UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Tese

MPC-SDP: um classificador genérico baseado em múltiplas perspectivas para a predição da evasão estudantil

Míriam Pizzatto Colpo

Pelotas, 2024

Míriam Pizzatto Colpo

MPC-SDP: um classificador genérico baseado em múltiplas perspectivas para a predição da evasão estudantil

Tese apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.

Orientador: Prof. Dr. Tiago Thompsen Primo
Coorientador: Prof. Dr. Marilton Sanchotene de Aguiar

Pelotas, 2024

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

C721m Colpo, Miriam Pizzatto

MPC-SDP [recurso eletrônico] : um classificador genérico baseado em múltiplas perspectivas para a predição da evasão estudantil / Miriam Pizzatto Colpo ; Tiago Thompsen Primo, orientador ; Marilton Sanchotene de Aguiar, coorientador. — Pelotas, 2024.
187 f. : il.

Tese (Doutorado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2024.

1. Evasão estudantil. 2. Modelos de predição. 3. Mineração de dados educacionais. 4. Aprendizado de máquina. I. Primo, Tiago Thompsen, orient. II. Aguiar, Marilton Sanchotene de, coorient. III. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

Míriam Pizzatto Colpo

MPC-SDP: um classificador genérico baseado em múltiplas perspectivas para a predição da evasão estudantil

Tese aprovada, como requisito parcial, para obtenção do grau de Doutor em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 07 de março de 2024

Banca Examinadora:

Prof. Dr. Tiago Thompsen Primo (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul

Prof. Dr. Marilton Sanchotene de Aguiar (coorientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul

Profa. Dra. Elaine Harada Teixeira de Oliveira
Doutora em Informática na Educação pela Universidade Federal do Rio Grande do Sul

Profa. Dra. Isabela Gasparini
Doutora em Computação pela Universidade Federal do Rio Grande do Sul

Prof. Dr. Júlio Carlos Balzano de Mattos
Doutor em Computação pela Universidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Agradeço aos meus pais, Cláudio e Eliane, e aos meus irmãos, Danieli e Miccael, por serem meu suporte e incentivo, sempre.

Aos meus afilhados, Benjamin, Martim, e Olívia, por serem luz e alegria em minha vida.

Aos meus amigos, em especial aos também compadres e vizinhos, Glivia e Diander, pela cumplicidade e pelos momentos de descontração, que tornaram essa caminhada mais leve.

Ao meu orientador, Tiago Thompsen Primo, e meu coorientador, Marilton Sanchotene de Aguiar, por se mostrarem sempre acessíveis e dispostos a compartilhar seus conhecimentos e experiências, pela compreensão, confiança e atenção dispensadas a mim nesse período.

Ao Programa de Pós-Graduação em Computação da Universidade Federal de Pelotas, por proporcionar formação pública, gratuita, e de qualidade. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e ao Programa Institucional de Incentivo à Qualificação Profissional em Programas Especiais, do Instituto Federal Farroupilha, pelo apoio financeiro.

Por fim, agradeço aos membros da banca examinadora por aceitarem o convite e dedicarem parte de seu tempo para contribuir com este trabalho.

RESUMO

COLPO, Míriam Pizzatto. **MPC-SDP: um classificador genérico baseado em múltiplas perspectivas para a predição da evasão estudantil**. Orientador: Tiago Thompsen Primo. 2024. 187 f. Tese (Doutorado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

Considerando o grave problema da evasão estudantil, diversos estudos têm aplicado técnicas de mineração de dados aos grandes volumes de dados educacionais, a fim de desenvolver modelos que permitam a identificação de estudantes e/ou fatores de risco. Essas pesquisas costumam representar o comportamento dos estudantes a partir de dados relacionados aos seus aspectos acadêmico, social e/ou econômico. Também é comum o desenvolvimento de modelos especializados em cursos e/ou momentos de predição, embora seus resultados sejam menos abrangentes. Ademais, os estudos aplicam geralmente as técnicas de forma tradicional, sem adaptá-las às especificidades do domínio da evasão. Com o intuito de explorar essas lacunas, nesta Tese foi proposto o MPC-SDP, um modelo de *ensemble* destinado a prever a evasão de estudantes em diferentes cursos e estágios da trajetória acadêmica (predição genérica), mas guiado por perspectivas especializadas e importantes para o domínio da evasão estudantil. Mais especificamente, o MPC-SDP visa (i) dar maior atenção à variabilidade dos padrões de evasão ao longo da trajetória acadêmica dos estudantes, a partir da construção de um comitê de classificação formado por subclassificadores especializados por períodos/semestres letivos; e (ii) uma participação mais abrangente de diferentes aspectos de dados dos estudantes (acadêmico, contextual, econômico, interacional e social) na representação do comportamento de evasão, por meio de um processo de seleção de atributos orientado por aspectos. Considerando o contexto dos cursos de graduação presenciais do Instituto Federal Farroupilha como estudo de caso, ao ser avaliado e comparado a modelos de predição genérica construídos de forma tradicional, o MPC-SDP demonstrou melhorar a precisão da predição de estudantes evadidos nos semestres iniciais. Na prática, isso oportunizaria um melhor direcionamento de medidas preventivas nas etapas mais críticas do acompanhamento estudantil, já que os semestres iniciais concentram os maiores quantitativos de matrículas e evasões. Além disso, a seleção de atributos orientada por aspectos proporcionou um aumento na variabilidade dos aspectos representados entre os principais padrões dos subclassificadores do MPC-SDP, contribuindo para um entendimento mais amplo e multifacetado dos padrões de evasão.

Palavras-chave: evasão estudantil; modelos de predição; mineração de dados educacionais; aprendizado de máquina.

ABSTRACT

COLPO, Míriam Pizzatto. **MPC-SDP: Multi-Perspective Classifier for Student Dropout Prediction**. Advisor: Tiago Thompsen Primo. 2024. 187 f. Thesis (Doctorate in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2024.

Given the severe issue of student dropout, many researchers have employed data mining techniques on large volumes of educational data to develop models that identify at-risk students and factors. The studies usually represent students' behavior based on their academic, social, and economic aspects. Developing models specialized in degrees and prediction moments was common, although their results needed to be more comprehensive. Additionally, studies generally apply the techniques in a traditional way without adapting them to the specificities of the dropout domain. In order to address these gaps, this thesis proposes the development of the MPC-SDP, an *ensemble* model designed to predict dropout in different degrees and stages of the academic trajectory (generic prediction) but which is guided by specialized perspectives that are important for the domain of student dropout. More specifically, the MPC-SDP aims to (i) give more attention to the variability of dropout patterns along the students' academic trajectory through the construction of an ensemble made up of sub-classifiers specialized by academic periods/semesters and (ii) enable broader participation of different aspects of student data (academic, contextual, economic, interactional, and social) in representing dropout behavior, via an aspect-oriented attribute selection process. When evaluated and compared to generic prediction models built traditionally, the MPC-SDP improved the precision of predicting students who drop out in the initial semesters, considering the context of face-to-face undergraduate courses at the Instituto Federal Farroupilha as a case study. This feature would better target preventive actions at the most critical stages of student follow-up since the initial semesters have the highest enrollment and dropout figures. Additionally, the aspect-oriented feature selection increased the variability of aspects represented among the main patterns of the MPC-SDP sub-classifiers, potentially contributing to a broader and multifaceted understanding of their dropout patterns.

Keywords: student dropout; prediction models; educational data mining; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACG	Ampla Concorrência Geral
AdaBoost	*Adaptive Boosting*
ANN	*Artificial Neural Network*
API	*Application Programming Interface*
AUC	*Area Under the Curve*
AVA	Ambiente Virtual de Aprendizagem
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
Consup	Conselho Superior do IFFar
COVID-19	*COronaVIrus Disease 2019*
CSV	Comma-Separated Values
DNN	*Deep Neural Network*
DT	*Decision Tree*
EAD	Educação a Distância
EDM	*Educational Data Mining*
ERE	Ensino Remoto Emergencial
FCNN	*Fully Connected Neural Network*
FIC	Formação Inicial e Continuada
Fies	Fundo de Financiamento ao Estudante do Ensino Superior
FN	Falsos Negativos
FP	Falsos Positivos
FSA	*Feature Selection by Aspect*
Fundeb	Fundo de Manutenção e Desenvolvimento da Educação Básica
GBT	*Gradient Boosting Trees*
G-mean	*Geometric Mean*
GPA	*Grade Point Average*
GRU	*Gate Recurrent Units*

IBGE	Instituto Brasileiro de Geografia e Estatistica
IDH	Índice de Desenvolvimento Humano
IDHM	Índice de Desenvolvimento Humano Municipal
IES	Instituição de Ensino Superior
IFFar	Instituto Federal de Educação, Ciência e Tecnologia Farroupilha
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbor*
LDA	*Linear Discriminant Analysis*
LR	*Logistic Regression*
LSTM	*Long-Short Term Memory*
LVQ	*Learning Vector Quantization*
MEC	Ministério da Educação
MLP	*Multilayer Perceptron*
MOOC	*Massive Open Online Course*
MPC-SDP	*Multi-Perspective Classifier for Student Dropout Prediction*
MSE	*Mean-Squared Error*
NB	*Naive Bayes*
PCA	*Principal Component Analysis*
PcD	Pessoa com Deficiência
PNAD	Pesquisa Nacional por Amostra de Domicílios
PNP	Plataforma Nilo Peçanha
PPE	Programa de Permanência e Êxito dos Estudantes do IFFar
PR	*Precision-Recall*
ProUni	Programa Universidade para Todos
RF	*Random Forest*
RFECV	*Recursive Feature Elimination with Cross-Validation*
Reuni	Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais
ROC	*Receiver Operating Characteristic*
RSL	Revisão Sistemática da Literatura
SHAP	*Shapley Additive Explanations*
SIG	Sistema Integrado de Gestão
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas

SMOTE	*Synthetic Minority Over-sampling Technique*
SP	*Select Percentile*
SQL	*Structured Query Language*
STEM	*Science, Technology, Engineering, and Mathematics*
SVM	*Support Vector Machine*
TFP	Taxa de Falsos Positivos
TVN	Taxa de Verdadeiros Positivos
TVP	Taxa de Verdadeiros Negativos
UAR	*Unweighted Average Recall*
VIF	Variation Inflation Factor
VN	Verdadeiros Negativos
VP	Verdadeitos Positivos
WoS	*Web of Science*
XAI	*eXplainable Artificial Intelligence*

SUMÁRIO

1 INTRODUÇÃO

Preconizadas pela Constituição Federal de 1988, na década de 1990, foram implementadas diversas políticas públicas voltadas à educação e à proteção social (Costa et al., 2021). A criação da Lei de Diretrizes e Bases da Educação Nacional (LDB), nº 9.394/1996 (Brasil, 1996), por exemplo, marcou a obrigatoriedade do Ensino Fundamental, contribuindo para que, já ao final do século passado, fosse observada uma grande elevação na frequência escolar (Costa et al., 2021). No início do século XXI, novas políticas públicas foram implantadas, como o Programa Bolsa Escola Federal, incorporado posteriormente pelo Programa Bolsa Família, e o Fundo de Manutenção e Desenvolvimento da Educação Básica (Fundeb) (Costa et al., 2021). A LDB também foi alterada, inicialmente prevendo progressiva extensão de obrigatoriedade ao ensino médio, na Lei nº 12.061/2009; e, após, estabelecendo a obrigatoriedade da educação básica, composta por pré-escola, ensino fundamental e ensino médio, dos 4 aos 17 anos, na Lei nº 12.796/2013 (Brasil, 1996). Como resultado, em 2015, a taxa de escolarização de brasileiros de 7 a 14 anos chegou a 98.8%, enquanto a de jovens de 15 a 17 anos atingiu 85.0% (Osório, 2017 apud Costa et al., 2021).

Naturalmente, a demanda pelo ensino superior também aumentou, motivando o desenvolvimento de novas políticas públicas, como o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni), o Programa Universidade para Todos (ProUni), e o Fundo de Financiamento ao Estudante do Ensino Superior (Fies), destinados, respectivamente, a ampliar a estrutura e a oferta de matrículas no ensino superior público, e a disponibilizar bolsas de estudo, e empréstimos estudantis em Instituições de Ensino Superior (IES) privadas (Costa et al., 2021). Como reflexo, o número de matrículas ao nível de graduação no ensino superior brasileiro mais que dobrou de 2003 a 2022, passando de 3.936.933 para 9.443.597 matrículas, conforme o Censo da Educação Superior, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (Inep, 2023).

Apesar do inegável avanço em direção à ampliação e à democratização do acesso ao ensino, o sistema educacional brasileiro continua a enfrentar o grave problema da evasão estudantil, caracterizada pela saída definitiva do estudante de seu curso de ori-

gem, sem a devida conclusão/diplomação (ANDIFES; ABRUEM; SESu/MEC, 1996). Apesar de ser um problema histórico, a preocupação em torno da evasão estudantil se tornou ainda maior durante a pandemia de COVID-19 (do inglês, *COronaVIrus Disease* 2019), devido ao contexto de exceção e aos desafios associados à transição da educação presencial para o Ensino Remoto Emergencial (ERE) (Santos; Zaboroski, 2020). Conforme o Instituto Brasileiro de Geografia e Estatística (IBGE), que traçou o cenário educacional de 2022, por meio da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), aproximadamente 18.3% dos brasileiros de 14 a 29 anos não completaram o ensino médio, sendo a necessidade de trabalhar a causa mais apontada (Gomes; Ferreira, 2023). No ensino superior brasileiro, o Instituto Semesp aponta que as taxas de evasão, em 2021, foram de 31.0% na rede privada e de 20.7% na rede pública, ambas considerando o ensino presencial. Na modalidade a distância, essas taxas foram de 36.6% e 27.1%, respectivamente (Semesp, 2023).

Além de prejuízos acadêmicos e sociais, a evasão reflete negativamente no desenvolvimento econômico do país, dada a relação entre o nível de escolarização e os ganhos salariais da população (Pontili; Staduto; Henrique, 2018). Adicionalmente, enquanto alunos que não concluíram seus cursos estão mais propensos ao desemprego e a necessitar de subsídios de assistência social, a escassez de mão de obra qualificada pode reduzir a capacidade produtiva da nação (Lee; Chung, 2019). Economicamente, a perda se torna ainda maior sob a perspectiva da rede pública de ensino, na qual a evasão significa recursos do Estado, não só financeiros, mas também de pessoal e infraestrutura, investidos sem o devido retorno (Silva Filho et al., 2007).

Considerando que a identificação de estudantes em risco, baseada no conhecimento do perfil de alunos previamente evadidos, pode fornecer subsídios às instituições de ensino para a tomada de decisões e a implementação de políticas preventivas (Nagy; Molontay, 2018), alguns estudos tentam traçar perfis de evasão a partir de análises manuais, baseadas em dados oriundos de entrevistas ou questionários. Porém, devido à dificuldade de contato com parcela significativa da população evadida, os resultados dessas pesquisas são suscetíveis a enviesamentos (Silva, 2013). Além disso, os grandes volumes de dados educacionais, providos por Ambientes Virtuais de Aprendizagem (AVAs) e sistemas de gestão acadêmica, embora armazenem informações abrangentes e de alto potencial estratégico, também inviabilizam explorações manuais (Romero; Ventura, 2020).

Nesse contexto, a fim de automatizar a análise de evasão, são adotadas técnicas de Mineração de Dados Educacionais (do inglês, *Educational Data Mining* – EDM), área de pesquisa que foca no desenvolvimento de métodos capazes de explorar grandes volumes de dados educacionais, a fim de compreender de forma mais eficaz o comportamento de alunos e outros fatores relacionados à aprendizagem (Baker; Isotani; Carvalho, 2011). Por ser uma área multidisciplinar, a mineração de dados (educa-

cionais ou não) incorpora técnicas de diversos domínios (Han; Kamber; Pei, 2012). O aprendizado de máquina supervisionado, por exemplo, é a base da classificação, que é a tarefa de EDM mais empregada no contexto da evasão, seja para a descoberta de padrões e atributos relacionados a esse fenômeno, seja para a previsão de alunos em risco (Colpo et al., 2020).

Além de diferentes modalidades de ensino, os estudos que aplicam EDM no contexto da evasão estudantil podem considerar escopos distintos de evasão, abordando o abandono de disciplinas ou, de forma mais abrangente, a evasão a nível de curso ou instituição (Colpo et al., 2020). Porém, em geral, grande parte dos estudos se dedica ao nível de graduação (Colpo et al., 2024), onde os índices de evasão costumam ser maiores, assim como a disponibilidade de dados, decorrente da maior adoção de AVAs e sistemas de gestão acadêmica nas IES.

Na geração dos modelos de EDM, em geral, as pesquisas valorizam o critério da interpretabilidade, uma vez que o conhecimento dos padrões de evasão pode ser decisivo para o desenvolvimento de medidas preventivas, junto à identificação dos alunos em risco. Além disso, o comportamento dos estudantes costuma ser representado a partir de dados relacionados aos seus aspectos acadêmico, social e/ou econômico. Note que, neste texto, considera-se como “aspectos” as diferentes dimensões de análise dos estudantes, que são representadas/caracterizadas a partir de determinados dados/atributos[1].

Por fim, é comum que a predição da evasão seja direcionada a um determinado estágio da trajetória acadêmica dos estudantes e/ou a cursos específicos. Embora possam obter vantagem preditiva a partir do uso de dados mais especializados, essas abordagens também restringem a abrangência de seus resultados e potenciais contribuições (Colpo et al., 2024). Por isso, alguns estudos têm se dedicado ao desenvolvimento de classificadores que visam compreender os estudantes e prever o risco de evasão em diferentes cursos e períodos letivos, referenciados, neste texto, como modelos “genéricos”. Porém, embora tenham um objetivo preditivo mais complexo, que requer uma maior capacidade de generalização do conhecimento, os modelos genéricos costumam ser construídos seguindo fluxos de EDM semelhantes aos utilizados por modelos especialistas.

1.1 Objetivos e Questões de Pesquisa

Visando obter vantagem de particularidades do domínio da evasão estudantil na identificação de estudantes em risco, esta Tese tem como objetivo geral a proposição de um modelo de predição genérico de evasão (destinado a prever evasões em di-

[1] Exemplo: a média de notas, a idade de ingresso e a renda familiar são atributos que ajudam a caracterizar os aspectos acadêmico, social/demográfico e econômico dos estudantes, respectivamente.

ferentes cursos e períodos/semestres da trajetória acadêmica dos estudantes), cuja estrutura conceitual seja guiada por perspectivas especializadas no domínio da evasão. Embora reprodutível, cabe antecipar que, para a construção e a validação do modelo proposto, esta Tese considera, como estudo de caso, o contexto dos cursos de graduação presenciais do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar).

Para isso, foram estabelecidos os seguintes objetivos específicos:

- Investigar o estado da arte de trabalhos que utilizam técnicas de EDM e aprendizado de máquina no contexto de predição da evasão estudantil;
- Identificar e extrair dados de estudantes de cursos de graduação presenciais do IFFar, para serem utilizados como estudo de caso;
- Estabelecer uma estrutura de aplicação de EDM focada no desenvolvimento de um modelo de predição de evasão genérico, mas que considere conhecimentos já revelados na literatura de predição de evasão;
- Avaliar o modelo produzido, a fim de verificar não só sua performance preditiva, mas também o conhecimento evidenciado por seus padrões preditivos.

O atendimento desses objetivos está diretamente relacionado a três questões de pesquisa:

- Como representar adequadamente dados de estudantes de diferentes tipos de cursos e períodos letivos, para torná-los comparáveis entre si e, assim, potencializar o reconhecimento de padrões comuns?
- Como estruturar a aplicação das técnicas de EDM, a fim de favorecer a generalização do aprendizado do modelo de predição genérico, em meio a diferentes períodos letivos?
- Como oportunizar, de forma mais efetiva, a participação de dados de diferentes aspectos dos estudantes nos padrões preditivos?

1.2 Contribuições

Ao alcançar os objetivos descritos na Seção 1.1, o trabalho desenvolvido nesta Tese apresenta a concepção/proposição do *Multi-Perspective Classifier for Student Dropout Prediction* (MPC-SDP), um modelo de *ensemble* ou comitê de classificação que, embora voltado à predição genérica, considera perspectivas especializadas e importantes para o domínio da evasão estudantil em sua estrutura/organização, visando: (i) dar maior atenção à variabilidade dos padrões de evasão no decorrer da trajetória

acadêmica, a partir da construção de subclassificadores especializados por períodos/-semestres letivos; e (ii) oportunizar uma participação mais abrangente de diferentes aspectos dos estudantes (acadêmico, contextual, econômico, interacional, e social/demográfico) na representação do comportamento de evasão, a partir de um processo de seleção de atributos orientado por aspecto, que recebeu o nome de *Feature Selection by Aspect* (FSA).

Como consequência dessa contribuição conceitual/técnica (MPC-SDP), esta Tese traz outras contribuições, que podem impactar diretamente nas práticas preditiva e educacional, relacionadas à evasão:

- Melhoria de precisão na predição da evasão em semestres iniciais, o que pode oportunizar um melhor direcionamento das medidas preventivas, justamente nas etapas mais críticas/importantes do acompanhamento dos estudantes.
- Maior variabilidade de aspectos nos principais padrões/atributos preditivos, o que pode contribuir para uma representação mais ampla/multifacetada do comportamento de evasão.
- Possibilidade de interpretação do modelo MPC-SDP, caso seja considerado um algoritmo interpretável na construção de cada subclassificador especializado por período/semestre. Para além de contribuir com o entendimento dos padrões de evasão que baseiam o modelo, isso pode auxiliar os gestores educacionais na proposição e/ou na adequação de estratégias preventivas, no contexto institucional de aplicação.

1.3 Organização do Texto

A seguir, no Capítulo 2, são apresentados conceitos de evasão, uma caracterização do IFFar e de seu enfrentamento institucional à evasão estudantil, além de uma fundamentação teórica sobre técnicas de EDM. No Capítulo 3, é descrito o desenvolvimento de uma RSL, com o intuito de caracterizar o estado da arte de pesquisas que aplicam EDM e aprendizado de máquina no contexto da evasão estudantil, com foco nos escopos institucional e de curso. Identificadas possíveis lacunas de pesquisa, no Capítulo 4, é proposto e detalhado o MPC-SDP. Após, no Capítulo 5, é descrita a implementação do MPC-SDP, incluindo a realização das atividades de extração e pré-processamento de dados do IFFar, necessárias para o desenvolvimento e a, posterior, avaliação do modelo, em um estudo de caso. No Capítulo 6, é descrita a avaliação experimental realizada para a validação do MPC-SDP; e, no Capítulo 7, são apresentados e discutidos os resultados dessa avaliação. Por fim, no Capítulo 8, este texto é concluído, junto a considerações finais e o apontamento de possíveis trabalhos futuros.

2 CONTEXTUALIZAÇÃO E FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são descritos conceitos e fundamentos teóricos associados a este trabalho. Na Seção 2.1, são apresentados conceitos da evasão estudantil, além de sua situação e enfrentamento no IFFar (estudo de caso). Na Seção 2.2, é apresentada uma fundamentação geral acerca de conceitos e técnicas associados à EDM. Embora seja dada maior atenção à tarefa de classificação, considerando que esta Tese trata a predição de evasão como um problema de classificação binária, também buscou-se apresentar um panorama da área de EDM, considerando sua vinculação aos trabalhos relacionados, a serem discutidos no Capítulo 3. Além disso, embora todos os conceitos relacionados à classificação sejam úteis e se apliquem a este trabalho, faz-se importante destacar que as técnicas escolhidas e utilizadas no desenvolvimento da Tese serão devidamente identificadas e justificadas nos Capítulos 5 e 6.

2.1 Evasão Estudantil

Diversos estudos apontam a variedade de definições e interpretações encontradas para o termo evasão escolar/estudantil na literatura, a exemplo de Coimbra; Silva; Costa (2021), e Rigo et al. (2014). Marco do campo científico e referência recorrente (Silva; Mariano, 2021), Tinto (1975) apontou que definições de evasão imprecisas ou desatentas fazem com que pesquisadores considerem da mesma forma estudantes que deixaram a instituição de ensino por razões muito distintas. A não distinção entre evasão/partida voluntária e o desligamento institucional por fracasso acadêmico, o abandono temporário, ou a transferência para outra instituição, por exemplo, pode levar pesquisadores a conclusões contraditórias, dificultando aos gestores acadêmicos o estabelecimento de políticas preventivas e a delimitação do público-alvo adequado (Tinto, 1975).

Tinto (1975) propôs um modelo teórico onde a evasão universitária (voluntária) é vista como um processo longitudinal das interações entre o estudante e os sistemas acadêmico e social da instituição de ensino. A integração e as vivências do estudante nesses sistemas modificam continuamente seus objetivos e compromissos com a ins-

tituição, levando-o à decisão de persistir ou evadir. Segundo Barroso et al. (2022), no modelo de Tinto (1975), é possível visualizar quatro conjuntos de variáveis que podem influenciar o processo de evasão:

- **(i) atributos prévios à entrada no ensino superior**, como os que envolvem o contexto familiar (situação socioeconômica, nível educacional dos pais, etc.), as características individuais/pessoais (gênero, raça, habilidades, etc.), e as experiências educacionais anteriores (resultados escolares, características das instituições frequentadas, etc.);
- **(ii) objetivos e compromissos prévios e posteriores à entrada no ensino superior**, incluindo os objetivos e compromissos pessoais para com a instituição e o plano de estudos (expectativas educacionais ou de carreira), as intenções acadêmicas, e os compromissos externos (ter que trabalhar para arcar com os custos do ensino superior, ter pessoas dependentes de si, etc.);
- **(iii) experiências envolvendo os sistemas acadêmico e social da instituição**, como experiências de desempenho acadêmico, interações pontuais com professores e demais colaboradores da instituição, atividades extracurriculares, interações com seus pares, etc.;
- **(iv) integração acadêmica e social**, resultante das experiências institucionais, formais e informais, do estudante em cada um dos sistemas, sendo que experiências positivas favorecem a integração do estudante.

Apesar da importância apontada por Tinto (1975) de se considerar definições precisas, a literatura frequentemente apresenta/considera definições mais simplistas, que descrevem a evasão como saída/desligamento/quebra de vínculo do estudante com o curso, instituição ou sistema, seja um ato voluntário ou não (Coimbra; Silva; Costa, 2021). O próprio Ministério da Educação do Governo Federal (MEC), apesar de citar Tinto (1975) no Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica (SETEC/MEC, 2014), considera definições amplas em seus documentos, estudos, e indicadores.

A Comissão Especial para o Estudo da Evasão nas Universidades Brasileiras, instituída pelo MEC no contexto das discussões do Programa de Avaliação Institucional das Universidades Brasileiras, representou um dos primeiros esforços nacionais na direção de esclarecer o conceito da evasão, identificar possíveis causas e soluções, e uniformizar a metodologia de análise a ser utilizada pelas instituições (SETEC/MEC, 2014). Como resultado do estudo, em seu relatório técnico, a comissão definiu a evasão de curso, objeto de seu estudo, como "***a saída definitiva do aluno de seu curso de origem, sem concluí-lo***" (ANDIFES; ABRUEM; SESu/MEC, 1996, p. 15), mas reconheceu/caracterizou a evasão em três dimensões distintas:

- **evasão de curso:** quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional;
- **evasão da instituição:** quando o estudante desliga-se da instituição na qual está matriculado;
- **evasão do sistema:** quanto o estudante abandona de forma definitiva ou temporária o ensino superior (ANDIFES; ABRUEM; SESu/MEC, 1996, p. 16).

Provavelmente pela maior facilidade de obtenção de dados e quantificação, percebe-se que os indicadores de evasão nacionais se concentram no nível de curso. A própria comissão ressaltou, em seu relatório (ANDIFES; ABRUEM; SESu/MEC, 1996), que a evasão de sistema só poderia ser traçada a partir do cruzamento de dados intra e inter-universidades, para cada estudante. No documento Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior, o Inep considera uma definição de evasão estabelecida em suas análises da educação básica, replicando-a para o ensino superior:

> **Evasão:** saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência (independentemente do motivo), representando, portanto, condição terminativa de insucesso em relação ao objetivo de promover o aluno a uma condição superior a de ingresso, no que diz respeito à ampliação do conhecimento, ao desenvolvimento cognitivo, de habilidades e de competências almejadas para o respectivo nível de ensino. Obviamente, a interrupção do programa em decorrência de falecimento do discente não pode ser atribuída como insucesso, dado que, de forma geral, se trata de caso fortuito e não se pode presumir uma intencionalidade do indivíduo em interromper o curso, cessá-lo ou uma incapacidade do indivíduo de manter-se no programa educacional (Inep, 2017, p. 10).

Silva; Mariano (2021) discutem e problematizam essa definição em vários aspectos. Entre eles, criticam a associação imediata da evasão ao insucesso, e apontam que a contabilização da evasão se dá por vaga e não por estudante. Assim, o Inep assume a condição terminativa na quebra de vínculo, sem, porém, verificar se o estudante não retornou, posteriormente, ao ensino superior ou se ele já não se encontra vinculado a outro curso. Outro aspecto apontado é que

> aos olhos do INEP e do MEC, as razões da perda de vínculo não são importantes para caracterizar o fenômeno, pois, quaisquer que sejam as causas, a interpretação se limitará a compreendê-lo como uma simples perda de vínculo (Silva; Mariano, 2021, p. 7).

Ao não distinguir as motivações para delimitar a evasão às perdas de vínculo que, de fato, representem um problema público, mas retirar os casos de falecimento do cômputo da evasão, o Inep também cria um paradoxo. Afinal, diversas outras causas de perda de vínculo poderiam ser consideradas fortuitas, não intencionais ou não preditivas da capacidade do indivíduo (Silva; Mariano, 2021).

Ainda considerando documentos do governo federal, o Guia de Referência Metodológica da Plataforma Nilo Peçanha (PNP) destaca que as taxas de evasão na Rede Federal de Educação Profissional, Científica e Tecnológica – Rede Federal – não são baseadas nos resultados produzidos pelo Inep, pois "*as complexas e diversificadas oferta e dinâmica escolar da Rede Federal não podem ser representadas a contento pela compatibilização entre o Censo Escolar e o Censo da Educação Superior*" (Moraes et al., 2018, p. 5). Apesar de definir como evadido o aluno "*que perdeu o vínculo com a Instituição antes da conclusão do curso*" (Moraes et al., 2018, p. 19), ao acessar a PNP (Brasil, 2023), onde são divulgados os dados estatísticos, de governança e transparência da Rede Federal, é possível verificar que são contabilizadas como evasões as seguintes situações de matrícula: abandono, cancelada, desligada, transferência interna, e transferência externa. Ou seja, as estatísticas de evasão da PNP também consideram a perda de vínculo com a vaga do curso, e não com a Instituição, já que são contabilizadas como evasão as transferências internas.

Como os dados oficiais de evasão da Rede Federal são divulgados na PNP, faz-se também importante verificar as definições apresentadas no Guia (Moraes et al., 2018) para os indicadores Taxa de Evasão e Evasão por Ciclo:

- **Taxa de Evasão**: percentual de matrículas que perderam o vínculo com a instituição no ano de referência, sem a conclusão do curso. Ou seja, é o percentual das evasões ocorridas em um determinado ano, perante o total de matrículas desse mesmo ano;
- **Evasão por Ciclo**: percentual de evasões em um ciclo de matrícula. Nesse caso, o cálculo deixa de considerar estatísticas de um ano base e analisa dados reunidos por ciclos de matrículas, onde

  > Um ciclo de matrícula envolve a oferta de um curso com uma carga horária definida, com mesma data de início e mesma previsão de término, visando englobar um conjunto de matrículas de alunos para obtenção de uma mesma certificação ou diploma. A análise dos indicadores "por ciclo" será realizada considerando a situação de matrícula dos alunos com fim de ciclo previsto para o ano anterior ao de referência (Moraes et al., 2018, p. 19).

A seguir, na Seção 2.1.1, serão apresentadas políticas de enfrentamento à evasão

no IFFar, além de estatísticas oficiais da instituição, extraídas da PNP e que consideram as definições e formas de cálculo recém descritas.

2.1.1 Evasão no Instituto Federal Farroupilha

Criado pela Lei nº 11.892, de 29 de dezembro de 2008, atualmente o IFFar possui, além de centros de referência e polos de Educação a Distância (EAD), 11 unidades acadêmicas (*Campi*) e seu órgão de administração central (Reitoria) (Jardim, 2018). Abrangendo as regiões Central, Noroeste e Oeste do Rio Grande do Sul, o IFFar visa orientar a oferta formativa de suas unidades em benefício do fortalecimento dos arranjos produtivos, sociais e culturais locais (IFFAR, 2022), e prioriza cursos técnicos de nível médio e de graduação, especialmente nas áreas tecnológicas e na formação de professores, embora oferte cursos em todas as modalidades e níveis de ensino (Jardim, 2018). Em 2022, o IFFar contabilizou 16326 matrículas, sendo: 2421 em cursos de Formação Inicial e Continuada (FIC); 6659 em cursos técnicos integrados, concomitantes, ou subsequentes ao ensino médio; 6847 em cursos de graduação, incluindo bacharelados, licenciaturas e tecnólogos; e 399 em cursos de pós-graduação *lato* ou *stricto sensu*. Enquanto a modalidade EAD têm alta predominância nas matrículas de FIC, nos níveis técnico e graduação, priorizados institucionalmente e com maior concentração de estudantes, cerca de 93% (12612) das matrículas correspondem ao ensino presencial (Brasil, 2023).

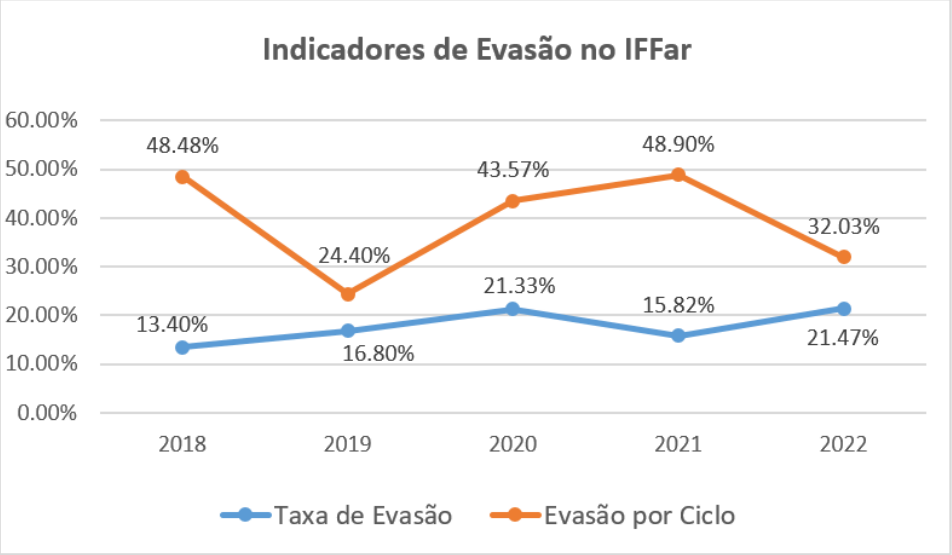
Inerente a qualquer instituição de ensino, o problema da evasão estudantil é uma preocupação constante no IFFar. Tratado como uma prioridade, no Plano de Desenvolvimento Institucional 2019-2026 (IFFAR, 2019), o combate à evasão também é foco do Programa de Permanência e Êxito dos Estudantes do IFFar (PPE), aprovado pela Resolução nº 178/2014, do Conselho Superior do IFFar (Consup), em 28 de novembro de 2014 (IFFAR, 2014). Considerando que a evasão pode ser causada por fatores individuais do estudante, e internos ou externos à instituição, o PPE estabelece um plano de ações de prevenção e acompanhamento, que visa mitigar potenciais desencadeantes do problema. As ações elencadas no PPE projetam desde a ampliação de auxílios da assistência estudantil e de bolsas de iniciação científica, para atender dificuldades sociais e financeiras dos estudantes; até a provisão de formação continuada a servidores de ensino do IFFar e a professores das Redes Públicas Municipal e Estadual, visando melhorar questões metodológicas, internamente, e a qualidade de ensino nas escolas públicas locais, externas ao IFFar.

Apesar de o problema ter recebido maior atenção institucional a partir de 2015, com o início da implantação do PPE, os dados de evasão disponíveis na PNP (Brasil, 2023), que consideram como referência os anos de 2018 a 2022, continuam a apontar um cenário desfavorável, conforme Figura 1. Após um atípico comportamento de

queda em 2021[1], a taxa de evasão anual do IFFar voltou a crescer em 2022, atingindo um valor similar a 2020, início do período pandêmico. Mais especificamente, os dados mais recentes revelam que uma parcela de **21.47%** dos alunos matriculados no IFFar, em 2022, evadiu. Em seu Relatório de Gestão 2022, o IFFar cita que, apesar da realização de busca ativa pelos estudantes em 2022, com o retorno integral das atividades letivas presenciais, muitos estudantes não conseguiram dar continuidade a suas atividades acadêmicas (IFFAR, 2023). Além disso, ao contextualizar o documento, a então reitora, Nídia Heringer, ressaltou que:

> O ano de 2022 teve inúmeras atipicidades, foi tão desafiador quanto os dois anos anteriores e apresentou complexidades de diferentes ordens: pedagógicas, econômicas e, sobretudo, interpessoais. Depois do distanciamento social, efetivar a readaptação da comunidade acadêmica ao presencial foi e continua sendo um dos maiores desafios de gestão, pois o sentir é único de cada indivíduo, que também é único quando se depara com novas realidades (IFFAR, 2023, p. 14).

Figura 1 – Evolução dos indicadores de evasão no IFFar, considerando dados de todos os níveis acadêmicos.



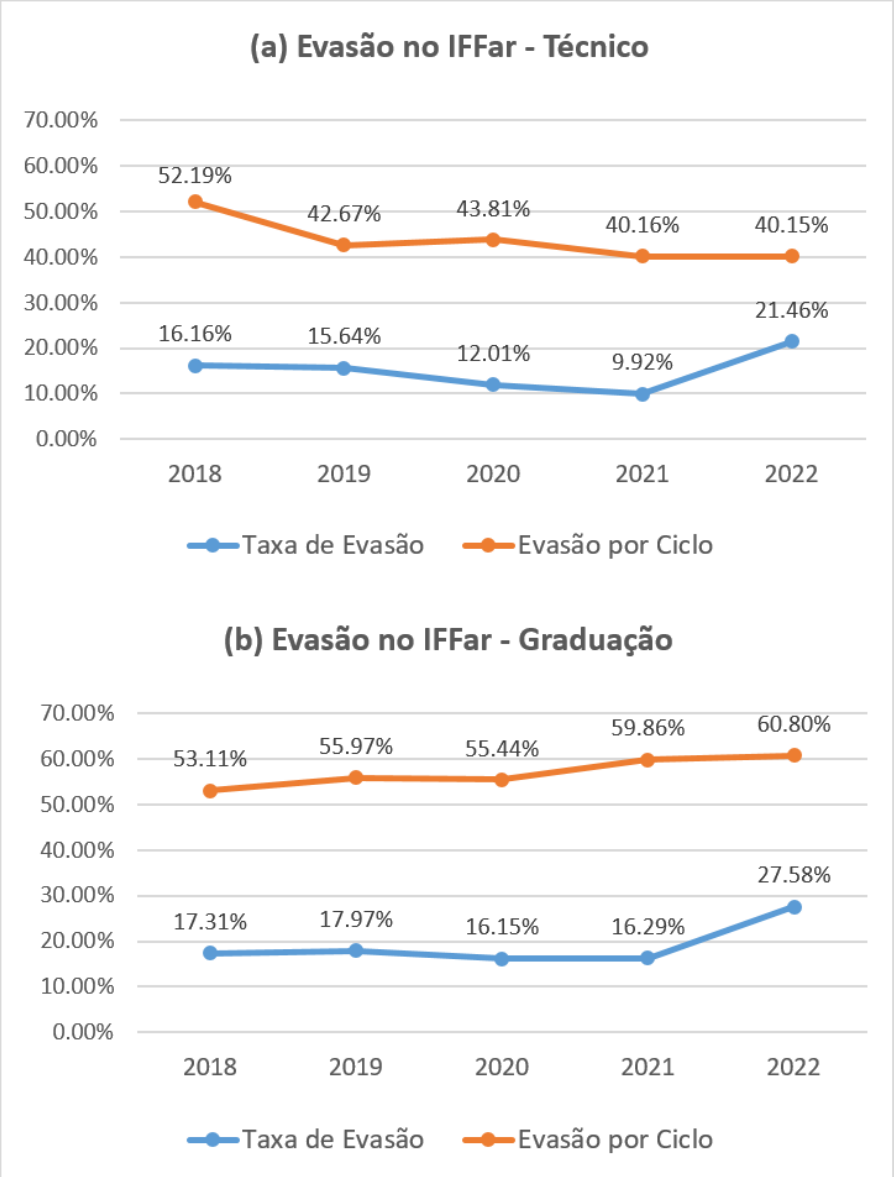
Fonte: Elaborada pela autora.

Nas Figuras 2(a) e 2(b) são considerados apenas os dados relacionados aos níveis técnico e graduação, priorizados institucionalmente. Pode-se observar que as taxas de evasão, que se mostravam em queda (técnico) ou estáveis (graduação) até

[1]Em seu Relatório de Gestão 2021, o IFFar atribui essa redução a ações iniciadas em 2020 e intensificadas em 2021, visando promover a permanência e êxito dos estudantes durante o contexto pandêmico e de ensino remoto, além de reverter as taxas preocupantes de 2020, que correspondem ao início do período pandêmico. Outro fator apontado pela instituição foi o retorno escalonado de atividades presenciais, no segundo semestre de 2021, que permitiu que muitos estudantes concluíssem o ano letivo (IFFAR, 2022).

2021, também passaram a apresentar uma tendência de crescimento, em 2022. O cenário se torna ainda mais preocupante ao se observar os altos valores dos últimos indicadores, principalmente no nível de graduação. Em 2022, dentre os estudantes matriculados, **21.46%** e **27.58%** evadiram de cursos de nível técnico e de graduação, respectivamente. E, no indicador acumulado por ciclo, os resultados recentes são ainda piores, alcançando **40.15%** de evasão nos cursos técnicos e **60.80%** nos de graduação.

Figura 2 – Evolução dos indicadores de evasão nos níveis técnico (a) e graduação (b) do IFFar.

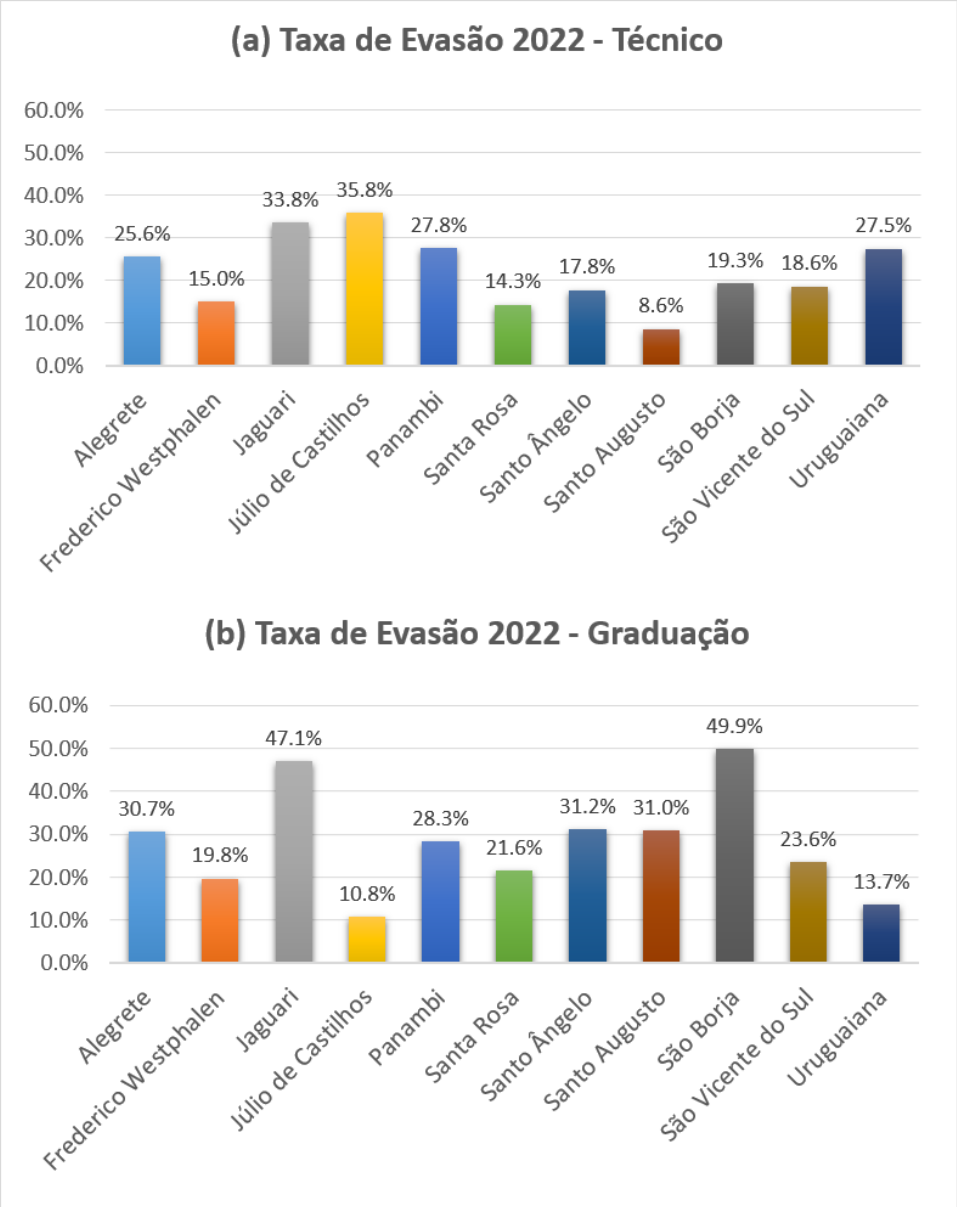


Fonte: Elaborada pela autora.

Ainda considerando dados da PNP (Brasil, 2023), na Figura 3 são apresentadas as taxas de evasão de 2022 por *campus*, considerando matrículas de cursos técnicos, na Figura 3(a), e de graduação, na Figura 3(b). No nível técnico, os *Campi* Santo

Augusto, Santa Rosa, e Frederico Westphalen apresentaram as menores taxas de evasão, enquanto os *Campi* Júlio de Castilhos, Jaguari, e Panambi apresentaram as maiores. Já na graduação, as menores taxas foram obtidas pelos *Campi* Júlio de Castilhos, Uruguaiana, e Frederico Westphalen; e as maiores pelos *Campi* São Borja, Jaguari, e Santo Ângelo. Pode-se observar que os *Campi* Frederico Westphalen e Jaguari estão, respectivamente, entre os de melhores e piores resultados, em ambos os níveis. Já o *Campus* Júlio de Castilhos apresentou a pior taxa de evasão do nível técnico e a melhor da graduação.

Figura 3 – Taxa de evasão por *campus*, no ano de referência 2022, para os níveis técnico (a) e graduação (b) do IFFar.



Fonte: Elaborada pela autora.

Esse cenário evidencia uma complexidade adicional na manifestação da evasão

em uma instituição de ensino tão plural quanto o IFFar. A oferta diferenciada de cursos, baseada nas potencialidades e características locais[2]; o cenário socioeconômico e cultural local; o perfil dos estudantes atendidos; a relação professor-aluno; as oportunidades de participação estudantil em projetos; e a forma com que as ações preventivas e de acompanhamento são planejadas e conduzidas; são exemplos de fatores que podem impactar os resultados de cada *campus*/unidade.

2.2 Mineração de Dados Educacionais

A mineração de dados é uma área de pesquisa cujo objetivo é analisar, de forma automática ou, mais frequentemente, semiautomática, grandes bases de dados, a fim de extrair padrões e conhecimentos úteis, que forneçam alguma vantagem estratégica a um determinado domínio de aplicação (Witten; Frank; Hall, 2011). Por seu caráter multidisciplinar, a mineração de dados se baseia em técnicas de diversas outras áreas, como: estatística, aprendizado de máquina, reconhecimento de padrões, inteligência artificial, e visualização de dados (Han; Kamber; Pei, 2012).

Com a informatização dos processos de gestão acadêmica/escolar e a expansão do ensino com suporte computacional, grandes volumes de dados educacionais tornaram-se disponíveis, despertando o interesse pela utilização de mineração de dados também nesse domínio. Assim, com o intuito de descobrir conhecimentos ocultos em dados educacionais e utilizá-los para compreender de forma mais eficaz o comportamento de alunos e outros fatores relacionados à aprendizagem, surgiu a área de pesquisa de EDM (Baker; Isotani; Carvalho, 2011).

As tarefas de mineração de dados, em geral, podem ser divididas entre **descritivas**, destinadas a encontrar padrões que caracterizem as propriedades dos dados de um determinado conjunto de dados; e **preditivas**, que visam realizar induções em dados atuais/correntes para prever algum fenômeno pré-estabelecido (Han; Kamber; Pei, 2012). Dentre as tarefas mais frequentemente empregadas na EDM, pode-se destacar, entre as descritivas, a mineração de associações e o agrupamento, e, entre as preditivas, a classificação e a regressão (Costa et al., 2013).

A **mineração de associações**, também conhecida como mineração de padrões frequentes, visa identificar padrões que ocorrem frequentemente em um conjunto de dados e, assim, descobrir associações e correlações entre os dados. Dentre os padrões de interesse, podem ser buscados conjuntos ou sequências (padrões sequenciais) de itens que aparecem associados com frequência (Han; Kamber; Pei, 2012).

[2]Embora os *campi* não sigam um mesmo catálogo de cursos, devido a diferentes características/necessidades locais, faz-se importante mencionar que a oferta entre os *campi* não é segmentada por área de pesquisa/atuação. Ou seja, diversos *campi* podem ofertar um mesmo curso ou cursos afins, como Bacharelado em Ciência da Computação, Licenciatura em Computação, e Tecnologia em Análise e Desenvolvimento de Sistemas.

Na EDM, um método de mineração de associação pode ser aplicado a uma base de notas de alunos, por exemplo, a fim de identificar regras do tipo “90% dos estudantes que apresentaram bom desempenho nas disciplinas de Lógica e Matemática também foram bem sucedidos em Programação” (Costa et al., 2013).

A tarefa de **agrupamento**, também conhecida como clusterização (*clustering*), procura dividir as instâncias de um conjunto de dados em grupos, de modo que as instâncias de um mesmo grupo apresentem alta similaridade entre si, mas pouca em relação aos componentes de outros grupos (Han; Kamber; Pei, 2012). A partir da caracterização/visualização de cada grupo, perante diferentes atributos, podem ser identificados seus padrões descritivos. Na EDM, um algoritmo de agrupamento pode ser aplicado a uma base de alunos, para mapear e caracterizar os perfis socioeconômicos dos estudantes, por exemplo.

Diferentemente das anteriores, as tarefas preditivas visam desenvolver modelos capazes de predizer um aspecto específico dos dados (atributo preditivo/alvo), a partir da análise dos demais aspectos do conjunto de dados (atributos preditores) (Costa et al., 2013). A construção desses modelos é feita a partir de aprendizado de máquina supervisionado, onde um conjunto de exemplos rotulados[3], denominado conjunto de treinamento, é utilizado para inferir uma função que mapeie o conjunto de atributos preditores, de maneira a predizer o atributo alvo/preditivo (Cechinel; Camargo, 2020). Enquanto, na tarefa de **regressão**, o atributo preditivo é contínuo; na **classificação**, essa variável é binária ou categórica (Costa et al., 2013), e também pode ser definida como “classe”. Como exemplo de aplicação, na EDM as tarefas de regressão e classificação podem ser utilizadas para predizer, respectivamente, notas e evasões, a partir de dados da interação dos estudantes com o AVA.

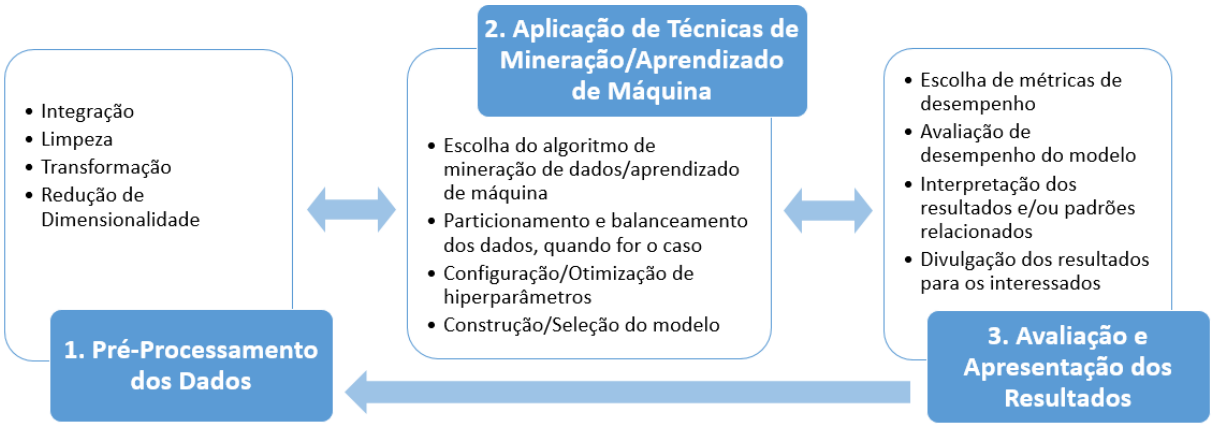
Após ser definida a tarefa de interesse, com base na questão educacional a ser analisada e nos dados disponíveis, a aplicação de EDM segue o processo geral de mineração de dados e de “Descoberta de Conhecimento em Bases de Dados” ou, do inglês, *Knowledge Discovery in Databases* (KDD) (Romero; Ventura, 2020). Embora muitos autores considerem KDD como um processo mais amplo, composto pelas macro-etapas de pré-processamento de dados, aplicação de métodos de mineração de dados, e avaliação e apresentação dos resultados, ambos os termos costumam ser tratados como sinônimos nos meios da indústria e da pesquisa, para se referir ao processo de descoberta de conhecimento como um todo (Han; Kamber; Pei, 2012).

Faz-se importante pontuar que, como representado na Figura 4, o processo de mineração de dados/KDD é, naturalmente, iterativo e cíclico. Por exemplo, se na segunda etapa for escolhido um algoritmo que não aceita um tipo de atributo presente no conjunto de dados pré-processado, será necessário retornar à etapa de pré-

[3]Em um conjunto de dados rotulado, cada instância, além dos dados de atributos preditores, apresenta o valor/rótulo da variável alvo (atributo preditivo).

processamento, para readequar os dados. Semelhantemente, se a etapa de avaliação demonstrar resultados insatisfatórios, é possível retornar às etapas de pré-processamento e de mineração de dados, a fim de testar a aplicação de técnicas diferentes.

Figura 4 – Representação das etapas do processo de mineração de dados ou KDD



Fonte: Elaborada pela autora.

A seguir, nas Seções 2.2.1, 2.2.2, e 2.2.3, são descritas, respectivamente, as etapas de pré-processamento, aplicação de métodos de mineração de dados, e avaliação e apresentação dos resultados, incluindo suas principais atividades (listadas na Figura 4).

2.2.1 Pré-processamento de dados

O pré-processamento de dados é a primeira etapa do processo de mineração de dados, e procura transformar dados brutos, previamente disponíveis ou coletados, em dados adequadamente formatados para a aplicação de algoritmos de extração de padrões/mineração de dados. Essencial para a descoberta de padrões úteis e de qualidade, o pré-processamento é uma fase complexa, que requer um significante esforço manual, e que pode consumir grande parte do tempo destinado ao processo de mineração de dados (Romero; Romero; Ventura, 2014).

Embora na EDM sejam adotadas técnicas tradicionais de pré-processamento de dados, é preciso observar que, no domínio educacional, existem diversas fontes de dados (como sistemas de gestão acadêmica, AVAs, repositórios de objetos de aprendizagem, jogos educacionais, e redes sociais), que fornecem diferentes tipos de dados (estruturados, textuais, multimídia), sob diferentes níveis de granularidade (cursos, turmas, estudantes, atividades). Dessa forma, a coleta de dados e as atividades de pré-processamento precisam ser devidamente planejadas e direcionadas para cada problema educacional específico (Romero; Ventura, 2020).

Dentre as atividades frequentemente adotadas no pré-processamento dos dados,

pode-se citar:

- **Integração** – consiste na tarefa de unir dados de múltiplas fontes, a fim de obter um conjunto de instâncias único e que contemple atributos de todas as fontes[4]. A integração precisa ser feita cuidadosamente, para evitar e/ou reduzir redundâncias e inconsistências no conjunto de dados resultante. A correta identificação das entidades entre as diferentes fontes de dados, por exemplo, é essencial para o mapeamento da integração e, consequentemente, para a obtenção de um conjunto de instâncias coerente (Han; Kamber; Pei, 2012).

- **Limpeza** – destinada a resolver problemas relacionados a ausências, duplicatas, e discrepâncias/inconsistências, também conhecidas como *outliers*, nos dados. O tratamento de dados faltantes é geralmente realizado a partir da remoção de atributos ou instâncias que possuam valores vazios; ou da imputação de dados, para o preenchimento desses valores (Raschka; Mirjalili, 2017). Nos demais casos, pode ser feita a remoção das instâncias inconsistentes ou duplicadas, além da suavização dos dados ruidosos, no caso de *outliers* (Han; Kamber; Pei, 2012).

- **Transformação** – consiste em transformar ou consolidar os dados, para representá-los de uma forma mais apropriada para a mineração (Han; Kamber; Pei, 2012) ou para adequá-los aos tipos de dados suportados por determinados algoritmos (Cechinel; Camargo, 2020). Exemplos de transformação são: (i) discretização, usada para transformar valores discretos em categóricos; (ii) binarização, em que cada valor de um atributo categórico é transformado em um novo atributo binário, cujos valores indicam a ocorrência ou não da categoria (Cechinel; Camargo, 2020); (iii) normalização ou padronização, usadas para reescalar os atributos numéricos, colocando-os na mesma ordem de grandeza (Raschka; Mirjalili, 2017); e (iv) derivação, onde novos atributos são criados/derivados a partir dos atributos originais (por meio de transformações matemáticas ou operações de agregação, por exemplo), com o intuito de potencializar/facilitar a identificação de padrões (Romero; Romero; Ventura, 2014).

- **Redução de dimensionalidade** – visa obter uma representação simplificada dos dados originais, por meio da redução de dimensionalidade (quantidade de atributos), sem causar perda significativa de informação (Cechinel; Camargo, 2020). A própria derivação de atributos pode proporcionar redução de dimensionalidade, se atributos derivados por agregações, por exemplo, forem utilizados

[4]Em mineração de dados, cada instância corresponde a um exemplo, individual e independente, do conceito a ser aprendido/analisado (Romero; Romero; Ventura, 2014); e as variáveis utilizadas para representar esses exemplos/instâncias são chamadas de atributos (Witten; Frank; Hall, 2011).

para substituir os atributos originais. Além disso, a redução de dimensionalidade também é realizada por técnicas de compressão, que transformam ou projetam os dados originais em um espaço menor, como a Análise de Componentes Principais (do inglês, *Principal Component Analysis* – PCA) (Han; Kamber; Pei, 2012), e por técnicas de seleção de atributos, que removem atributos redundantes ou de menor relevância (Colpo et al., 2021).

2.2.2 Aplicação de métodos de mineração de dados

Como mencionado anteriormente, a aplicação de métodos/algoritmos de mineração de dados corresponde à segunda macro-etapa do processo de mineração de dados. Existem diversas ferramentas e bibliotecas de mineração de dados/aprendizado de máquina disponíveis, fornecendo os mais variados tipos e implementações de algoritmos. Porém, a escolha do algoritmo a ser utilizado depende, essencialmente, da tarefa de mineração que se deseja executar, para atender/solucionar um problema (educacional, no caso da EDM) específico.

Além disso, na EDM, é muito importante avaliar o critério da **interpretabilidade** na escolha do algoritmo. Embora modelos mais complexos e não interpretáveis costumem apresentar melhores resultados, muitas vezes é preferível utilizar um algoritmo mais simples, mas que produza modelos interpretáveis, a fim de permitir a compreensão dos padrões descobertos e, assim, contribuir para processos de tomada de decisão (Romero; Ventura, 2020; Colpo et al., 2022). Apriori, K-means, árvore de decisão (do inglês, *Decision Tree* – DT), e regressão linear são exemplos de algoritmos clássicos, que produzem modelos interpretáveis, e que são popularmente empregados nas tarefas de mineração de associação, agrupamento, classificação e regressão, respectivamente (Witten; Frank; Hall, 2011).

Em geral, cada algoritmo de mineração de dados exige a configuração de um conjunto específico de hiperparâmetros, responsável por estabelecer características a serem observadas na construção dos modelos. Ou seja, cada possível combinação de hiperparâmetros produzirá um modelo diferente e, consequentemente, resultados diferentes (Cechinel; Camargo, 2020). Dessa forma, após escolhido o algoritmo, é importante otimizá-lo, por meio da **configuração de hiperparâmetros**.

Em tarefas preditivas, como os modelos são construídos a partir de aprendizado de máquina supervisionado, existem preocupações adicionais a serem consideradas na aplicação dos algoritmos. O problema de **sobreajuste** ou ***overfitting***, por exemplo, ocorre quando um modelo preditivo captura padrões adequados para os dados de treinamento, mas que demonstram pouca capacidade de generalização em dados desconhecidos (Raschka; Mirjalili, 2017). Ou seja, o sobreajuste ocorre quando o modelo se especializa demais aos exemplos de treinamento, aprendendo não apenas seus padrões gerais, mas também padrões específicos e ruídos presentes nos dados

(Cechinel; Camargo, 2020). Além da aplicação de técnicas de limpeza e seleção de atributos, no pré-processamento dos dados, a escolha e a devida parametrização do algoritmo de classificação/regressão pode reduzir o risco de *overfitting* (Han; Kamber; Pei, 2012). Algoritmos de comitê ou *ensemble*, por exemplo, costumam ser menos suscetíveis a esse problema, à medida que constroem e consideram um conjunto de diferentes modelos na predição.

Na tarefa de classificação, em que se deseja construir um modelo capaz de distinguir classes de dados (Han; Kamber; Pei, 2012), o processo de aprendizado pode ser prejudicado, caso exista uma grande disparidade entre as quantidades de exemplos de cada classe (Cechinel; Camargo, 2020). Ou seja, ao se basear em um conjunto de treinamento desbalanceado, o modelo tende a priorizar o aprendizado da classe majoritária e, consequentemente, prejudicar a identificação da classe sub-representada. A fim de evitar esse problema, antes da utilização do algoritmo de classificação, costuma ser aplicado algum método de amostragem para realizar o **balanceamento** dos dados de treinamento. Enquanto técnicas de subamostragem reduzem, aleatoriamente ou a partir de métodos baseados em *clusters*, a quantidade de exemplos/instâncias da classe majoritária; métodos de sobreamostragem incrementam, por reamostragem aleatória ou geração de dados sintéticos, a quantidade de amostras da classe minoritária (Cechinel; Camargo, 2020).

2.2.3 Avaliação e apresentação dos resultados

Embora algoritmos de mineração de dados possam identificar uma grande quantidade de padrões nos dados, nem todos esses padrões são interessantes, ou seja, nem todos esse padrões representam um novo e potencialmente útil conhecimento (Han; Kamber; Pei, 2012). Por isso, a última etapa do processo de mineração de dados corresponde à avaliação.

> Um padrão é **interessante** se for (1) *facilmente compreendido por humanos*, (2) *válido* em dados novos ou de teste com algum grau de *certeza*, (3) potencialmente *útil*, e (4) *novo*. Um padrão também é interessante se validar uma hipótese que o usuário *procurava confirmar*. Um padrão interessante representa **conhecimento**. (Han; Kamber; Pei, 2012, p. 21, tradução nossa).

Existem diversos **critérios objetivos** para avaliar a representatividade/credibilidade dos padrões descobertos. Em regras de associação, por exemplo, são observadas as medidas de suporte e confiança de cada regra, representando, respectivamente, o número de casos satisfeitos pela regra e o grau/porcentagem de certeza da associação (Witten; Frank; Hall, 2011). Porém, considerando que padrões podem ser bem avaliados por critérios objetivos, mas não representar um conhecimento novo ou de interesse, sempre que possível, medidas objetivas devem ser combinadas a

critérios subjetivos, que reflitam necessidades e interesses dos usuários, considerando cada contexto específico de aplicação (Han; Kamber; Pei, 2012). Em tarefas de agrupamento, onde não se tem um critério objetivo de sucesso, essa avaliação subjetiva é ainda mais importante e, em geral, reflete a percepção de utilidade dos grupos descobertos para o contexto de aplicação (Witten; Frank; Hall, 2011).

Nas tarefas preditivas, dependendo do algoritmo de regressão/classificação que for empregado, nem sempre é possível interpretar padrões/regras de predição diretamente. Porém, sempre é possível e necessária uma forte avaliação quantitativa dos modelos gerados, a fim de verificar o desempenho preditivo e, consequentemente, aproximar o grau de confiabilidade que esses modelos apresentarão em previsões reais, ou seja, quando os rótulos dos dados forem, de fato, desconhecidos. Considerando a possibilidade de sobreajuste, mencionada na Subseção 2.2.2, naturalmente, o desempenho preditivo de um modelo deve ser avaliado sobre um conjunto de dados diferente do utilizado no treinamento. Ou seja, a avaliação deve ser feita sobre instâncias rotuladas e independentes, que não foram utilizadas no processo de aprendizado do classificador/regressor (Cechinel; Camargo, 2020). Dada a função que exerce na avaliação, esse conjunto de dados adicional é denominado de **conjunto de teste**.

Quando não existem, de antemão, conjuntos de dados distintos[5], o **particionamento dos dados** entre os conjuntos de treinamento e teste é, em geral, realizado por uma das seguintes estratégias[6]:

- ***holdout*** – os dados são divididos aleatoriamente entre os conjuntos independentes de treino e teste, de modo que o conjunto de treinamento seja utilizado para derivar/treinar o modelo, e o de teste para estimar o desempenho preditivo (Han; Kamber; Pei, 2012). Essa estratégia pode ser adotada quando existe um grande volume de dados disponível para a mineração, sendo comum, nesses casos, considerar tamanhos de aproximadamente 75% e 25% para as partições de treinamento e teste, respectivamente (Cechinel; Camargo, 2020);

- **validação cruzada**/***cross-validation*** – os dados são divididos em k partições (k-*folds*), de aproximadamente mesmo tamanho, aleatoriamente. Inicialmente, uma das partições é reservada para teste e as outras k-1 são utilizadas para treinar o modelo preditivo. Após treinado, o modelo é avaliado sobre a partição de teste, reservada anteriormente. O procedimento é repetido k-1 vezes, sempre considerando uma partição ainda não utilizada como conjunto de teste e as outras k-1 partições como conjunto de treinamento. Dessa forma, ao término do

[5]Note que, embora as instâncias devam ser diferentes entre os conjuntos, a estrutura dos dados (atributos/variáveis) deve ser a mesma.

[6]Em tarefas de classificação, ambas as estratégias costumam realizar o particionamento estratificadamente, ou seja, garantindo que a amostragem aleatória mantenha as mesmas proporções de representação das classes, nos conjuntos de treinamento e teste (Witten; Frank; Hall, 2011).

processo, k modelos terão sido treinados e avaliados, e a média de seus resultados definirá o desempenho preditivo (Raschka; Mirjalili, 2017). Essa estratégia é utilizada, principalmente, quando a quantidade de dados é limitada. E, embora validações empíricas tenham indicado que o uso de dez partições (k=10) pode oferecer uma melhor estimativa de erro, tornando-o um padrão, o uso de 5 ou 20 partições pode proporcionar praticamente os mesmos resultados (Witten; Frank; Hall, 2011).

Em predições numéricas, ou seja, na tarefa de regressão, o **erro quadrático médio** (do inglês, *Mean-Squared Error* – MSE) é a métrica mais utilizada para avaliar o desempenho preditivo dos modelos[7] (Witten; Frank; Hall, 2011). Considerando como erro a diferença entre o valor predito e o valor real de uma instância, como o próprio nome sugere, o MSE calcula o quadrado do erro de predição de cada instância, e, após, calcula a média desses erros quadráticos. Ou seja, sendo $p_1, ..., p_n$ os valores preditos, e $r_1, ..., r_n$ os valores reais de um conjunto de teste de n instâncias:

$$MSE = \frac{(p_1 - r_1)^2 + ... + (p_n - r_n)^2}{n}$$

Já em tarefas de classificação (predições categóricas), cada predição de classe incorreta (ou seja, quando o modelo prediz, para uma instância de teste, uma classe diferente da real/correta) é considerada um erro de classificação. Dessa forma, a taxa de erro ou **taxa de classificação incorreta** é dada pela divisão da quantidade de erros de classificação pelo total de instâncias de teste; e, semelhantemente, a taxa de acerto/sucesso, mais comumente chamada de **acurácia**, corresponde à proporção de acertos de classificação perante o total de instâncias de teste (Raschka; Mirjalili, 2017). Embora avaliem o desempenho geral dos modelos, essas métricas não demonstram a capacidade preditiva do classificador perante as diferentes classes, e, por isso, não devem ser utilizadas isoladamente (Barros et al., 2019).

O desempenho preditivo de um classificador pode ser descrito, detalhadamente, por uma estrutura matricial quadrada e de ordem n, sendo n o número de classes envolvidas na classificação. Essa estrutura é denominada **matriz de confusão** e, a partir de suas informações, outras métricas de avaliação são calculadas (Han; Kamber; Pei, 2012). Considerando que, na EDM, os problemas de classificação geralmente são binários (só existem dois rótulos possíveis para o atributo classe/preditivo) (Cechinel; Camargo, 2020), a Tabela 1 apresenta um exemplo de matriz de confusão para as classes “SIM” e “NÃO”, no qual a classe de interesse é a classe “SIM” (positiva para algum fenômeno educacional de interesse[8]), e:

[7] Diversas outras métricas para avaliação de tarefas de regressão são descritas por Witten; Frank; Hall (2011), junto da análise de suas vantagens/desvantagens.

[8] No problema de predição de evasão, por exemplo, as classes “SIM” e “NÃO” podem indicar, respec-

- Verdadeiros Positivos (VP) são a quantidade de instâncias positivas (classe real = “SIM”) classificadas corretamente (classe predita = “SIM”);
- Falsos Positivos (FP) são a quantidade de instâncias negativas (classe real = “NÃO”) classificadas indevidamente como positivas (classe predita = “SIM”);
- Falsos Negativos (FN) são a quantidade de instâncias positivas (classe real = “SIM”) classificadas indevidamente como negativas (classe predita = “NÃO”);
- Verdadeiros Negativos (VN) são a quantidade de instâncias negativas (classe real = “NÃO”) classificadas corretamente (classe predita = “NÃO”).

Tabela 1 – Estrutura de uma matriz de confusão binária, com as classes “SIM” e “NÃO”

		Classe real	
		SIM	NÃO
Classe predita	SIM	VP	FP
	NÃO	FN	VN

Fonte: Elaborada pela autora.

Considerando a composição da matriz de confusão binária, na Tabela 2, além das fórmulas das métricas de acurácia e taxa de classificação incorreta, mencionadas anteriormente, são apresentadas outras métricas utilizadas/combinadas para avaliar o desempenho de um modelo de classificação/predição. Entre as mais adotadas, **precisão** corresponde à proporção de registros corretamente preditos entre as classificações positivas (exatidão), enquanto **revocação**, também conhecida como sensibilidade ou Taxa de Verdadeiros Positivos – TVP, indica a proporção de registros positivos que foram corretamente classificados/recuperados para essa classe (cobertura/abrangência). **F1**, por sua vez, integra essas duas métricas em uma média harmônica, considerando que precisão e revocação costumam apresentar relação inversamente proporcional (Han; Kamber; Pei, 2012). Apesar de menos utilizadas, de acordo com Barros et al. (2019), a média geométrica (do inglês, *Geometric Mean* – ***G-mean***) e a revocação média não ponderada (do inglês, *Unweighted Average Recall* – **UAR**) refletem de forma mais adequada o desempenho de classificação em um cenário de dados desbalanceados. Como o próprio nome sugere, *G-mean* é uma média geométrica entre a TVP e a Taxa de Verdadeiros Negativos (TVN). Também conhecida como especificidade, TVN corresponde à proporção de registros negativos que foram corretamente classificados, o que também pode ser interpretado como a revocação da classe negativa. Por isso, a UAR consiste na média aritmética entre TVP e TVN.

tivamente, instâncias que correspondem ou não a um estudante evadido, nos conjuntos de treinamento e teste.

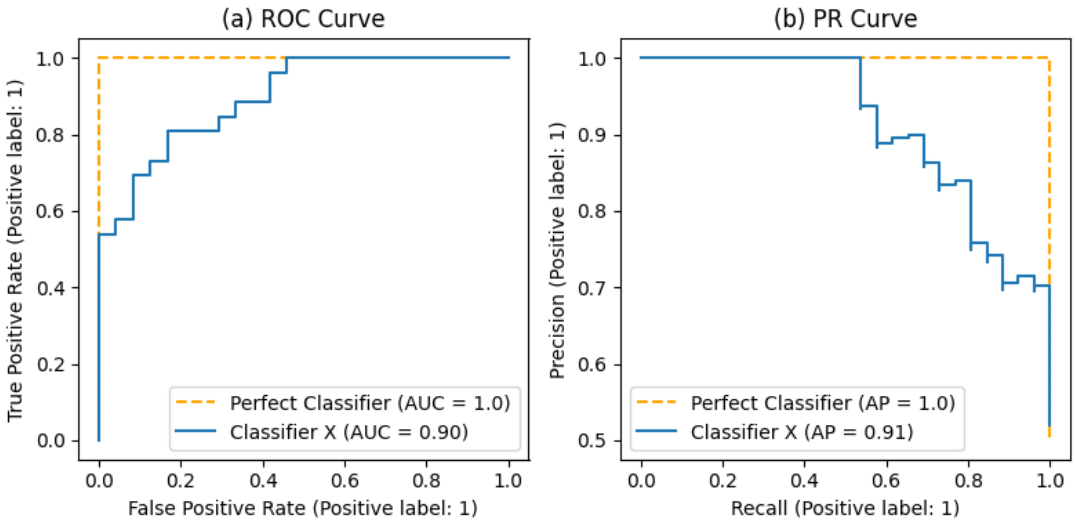
Tabela 2 – Métricas avaliativas para tarefas de classificação

Métrica	Fórmula
Acurácia, Taxa de acerto	$\frac{VP + VN}{VP + FP + FN + VN}$
Taxa de classificação incorreta, Taxa de erro	$\frac{FP + FN}{VP + FP + FN + VN}$
Revocação, Sensibilidade, Taxa de Verdadeiros Positivos (TVP)	$\frac{VP}{VP + FN}$
Especificidade, Taxa de Verdadeiros Negativos (TVN)	$\frac{VN}{VN + FP}$
Taxa de Falsos Positivos (TFP)	$\frac{FP}{FP + VN}$
Precisão	$\frac{VP}{VP + FP}$
F, F_1, *F-score*	$2\ x\ \frac{Precisão\ x\ Revocação}{Precisão + Revocação}$
G-mean	$\sqrt{TVP\ x\ TVN}$
UAR	$(TVP + TVN)/2$

Fonte: Adaptada de Han; Kamber; Pei (2012, p. 365) e Barros et al. (2019, p. 5).

Outras formas de avaliação importantes, no âmbito da classificação, são as **curvas ROC** (do inglês, *Receiver Operating Characteristic*) e **PR** (Precisão-Revocação). A ROC, exemplificada na Figura 5(a) é uma representação bidimensional do desempenho de um classificador, em que os eixos horizontal e vertical representam, respectivamente, as Taxas de Falsos Positivos (TFP) e de Verdadeiros Positivos (TVP), calculadas a partir do deslocamento do limiar de decisão do classificador (Raschka; Mirjalili, 2017). No espaço ROC, o ponto nas coordenadas (0,1) representa um modelo perfeito, capaz de identificar todas as instâncias positivas sem FP (Cechinel; Camargo, 2020). A curva PR, por sua vez, representa o *trade-off* precisão vs. revocação, nos eixos vertical e horizontal, respectivamente, e é mais apropriada quando há desbalanceamento de dados entre as classes (Scikit-learn, 2022). Como demonstrado na Figura 5(b), o ponto nas coordenadas (1,1) representa um modelo perfeito, em uma curva PR. As curvas ROC e PR costumam ser sintetizadas a partir da área abaixo da curva (do inglês, *Area Under the Curve* – AUC) e da precisão média (do inglês, *Average Precision* – AP), respectivamente. A AP corresponde à média ponderada das precisões alcançadas em cada limiar, considerando como peso o aumento de revocação em relação ao limiar anterior (Scikit-learn, 2022). Para ambas as métricas, quanto maior o valor alcançado, melhor o modelo.

Figura 5 – Exemplos de curvas ROC (a) e PR (b)



Fonte: Elaborada pela autora.

Por fim, como mencionado na Seção 2.2, o processo de EDM é tipicamente iterativo. Independentemente da tarefa de mineração e da metodologia de avaliação adotada, se os resultados produzidos não forem considerados satisfatórios, é possível retornar às etapas anteriores, a fim de tentar otimizar os modelos e, consequentemente, produzir conhecimentos/resultados mais úteis. Porém, ao término do processo, os conhecimentos resultantes devem ser sistematizados e apresentados aos demais interessados, para poderem contribuir, de fato, para a prática educacional. Essa apresentação pode se dar de diferentes formas, conforme o objetivo e o problema considerado. Estudos relacionados à previsão de evasão estudantil, por exemplo, podem prover desde um relatório com a descrição de fatores ou regras/padrões mais fortemente relacionados à evasão, até ferramentas/sistemas de predição, para monitoramento do risco de evasão de estudantes ativos/matriculados.

3 TRABALHOS RELACIONADOS: REVISÃO SISTEMÁTICA DA LITERATURA (RSL)

Este capítulo visa caracterizar, a partir do desenvolvimento de uma RSL, o estado da arte de pesquisas que aplicam EDM e aprendizado de máquina no contexto da evasão estudantil, com foco nos escopos institucional e de curso. Nas Seções 3.1 e 3.2 são apresentados, respectivamente, detalhes da metodologia adotada e da condução da RSL. Após, na Seção 3.3, são apresentados os resultados obtidos pela RSL, assim como os trabalhos relacionados. Na Seção 3.4, são apontadas tendências, oportunidades e desafios da área. Por fim, na Seção 3.5, são apresentadas as considerações finais sobre o capítulo, incluindo o posicionamento geral desta Tese frente aos trabalhos relacionados. A RSL associada a este capítulo encontra-se publicada, em Inglês, pela Revista Brasileira de Informática na Educação, em Colpo et al. (2024).

3.1 Metodologia da RSL

Segundo Kitchenham; Charters (2007), o desenvolvimento de uma RSL visa identificar e avaliar o maior número possível de pesquisas em uma área ou tema de interesse, por meio de uma metodologia justa e confiável. Neste trabalho, foi utilizada a metodologia de RSL proposta por esses autores, visando investigar o panorama internacional das pesquisas que adotam EDM e aprendizado de máquina no contexto da previsão de evasão estudantil. Mais especificamente, como principal questão de pesquisa, visamos identificar as características contextuais, técnicas e de dados relacionadas a esses estudos, considerando os escopos institucional e de curso. Para isso, inicialmente, as seguintes questões específicas de pesquisa foram estabelecidas:

- Q1. Quais os objetivos dessas pesquisas?
- Q2. Quais níveis, modalidades e redes de ensino são considerados?
- Q3. Qual a natureza, abrangência e volume dos dados utilizados?

- Q4. Quais tarefas, técnicas e ferramentas estão sendo aplicadas?

Nesta RSL foram considerados trabalhos publicados de 2016 a 2022, nas seguintes bases científicas: ACM Digital Library[1], IEEE Xplore[2], Web of Science[3] e Scopus[4]. Enquanto a delimitação temporal foi estabelecida contemplando os últimos sete anos que antecederam a execução da revisão, por ser uma amostragem bastante representativa dos avanços e do cenário atual de pesquisa, os repositórios de busca foram escolhidos por sua relevância internacional para a comunidade de Informática na Educação.

A expressão utilizada na busca foi refinada previamente, por meio da execução e da avaliação de consultas preliminares, e abrangeu termos relacionados ao Problema (“drop*out” OR “dropout”), à População (“student” OR “school” OR “college” OR “university”) e à Intervenção (“data mining” OR “machine learning”). No entanto, como os repositórios utilizam diferentes formas padrão de busca, padronizou-se que as consultas deveriam ser executadas sobre os metadados das publicações (títulos, resumos e palavras-chave). Para isso, foram criadas consultas avançadas, adaptadas às diferentes opções e sintaxes de cada repositório, conforme apresentado na Tabela 3.

Tabela 3 – Expressões de busca avançadas para cada base de dados científica.

Base	Expressões de busca avançada
ACM	Abstract:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning")) OR Title:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning")) OR Keyword:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))
IEEE	(("All Metadata":"drop*out" OR "All Metadata":"dropout") AND ("All Metadata":"student" OR "All Metadata":"school" OR "All Metadata":"college" OR "All Metadata":"university") AND ("All Metadata":"data mining" OR "All Metadata":"machine learning"))
Scopus	TITLE-ABS-KEY(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))
WoS	TS=(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))

Fonte: Traduzida de Colpo et al. (2024).

Para a seleção dos estudos, foram estabelecidos cinco critérios a serem atendidos por cada documento:

- ser escrito em Português ou Inglês, no formato de artigo científico;

[1] https://dl.acm.org
[2] https://ieeexplore.ieee.org
[3] https://www.webofscience.com
[4] https://www.scopus.com

- descrever, detalhadamente, a aplicação de uma solução específica, sendo excluídas revisões da literatura;
- adotar técnicas de EDM e aprendizado de máquina em soluções que tenham a intenção de contribuir para a previsão da evasão estudantil;
- considerar a evasão em um escopo institucional ou de curso, descartando trabalhos destinados ao abandono de disciplinas ou capacitações de curta duração;
- apresentar detalhes de seu desenvolvimento e resultados, transpondo o âmbito de proposta.

Por fim, para refinar o processo de seleção/exclusão dos estudos, também foram observados três critérios de qualidade:

- o artigo deve ter mais de cinco páginas (quer seja de coluna simples ou dupla), a fim de apresentar um detalhamento mínimo da solução;
- em suas validações, a pesquisa deve considerar um conjunto de dados de, no mínimo, 500 instâncias/alunos, a fim de garantir uma representatividade/qualidade mínima aos experimentos; e
- os dados utilizados no trabalho não devem ser coletados, majoritariamente, por questionários, devido à maior propensão de enviesamento desta abordagem.

3.2 Condução da RSL

Após planejada, a RSL foi executada em 2023. A condução das buscas e da seleção dos artigos foi realizada com o auxílio da ferramenta Parsifal[5] e compreendeu três fases. Inicialmente, na Fase 1, foram realizadas as buscas sobre as bases de dados científicas[6], e exportados os metadados dos resultados (incluindo os resumos) em formato BibTeX. Após, os arquivos BibTeX foram importados na ferramenta Parsifal, para agilizar a verificação de duplicatas e o processo de revisão dos artigos. Após a eliminação das duplicatas exatas, na Fase 2, os estudos resultantes da Fase 1 tiveram seus resumos e metadados (como tamanho, formato e idioma) analisados, tendo sido descartados os trabalhos que demonstraram não atender a algum dos critérios de seleção ou qualidade estabelecidos anteriormente[7]. Aqui, faz-se importante mencionar

[5] `https://parsif.al/`

[6] O acesso às bases científicas foi realizado a partir do Portal de Periódicos (`https://www.periodicos.capes.gov.br`) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que é uma fundação vinculada ao MEC.

[7] Note que, embora os resumos estivessem presentes, nem todos os metadados das publicações incluíam informações sobre páginas, idioma e formato. Consequentemente, alguns artigos que poderiam ter sido excluídos nesta fase passaram para a próxima

que, devido ao grande volume de publicações coletadas, não foi adotada a revisão por pares na análise dos artigos. Ou seja, cada estudo foi analisado por um único pesquisador. Finalmente, na Fase 3, foram extraídos os textos completos dos estudos resultantes da Fase 2, quando disponíveis[8], e analisados conforme os critérios de seleção e qualidade. Além disso, quando detectamos publicações semelhantes e de mesmos autores durante a leitura, mantivemos apenas os estudos mais recentes. Isso porque, embora não sejam duplicatas idênticas, estes artigos representarem o progresso de uma mesma pesquisa.

Na Figura 6 são apresentados, por base científica, o número de resultados retornados nas buscas e o número de exclusões para cada critério analisado e para cada fase. Observe que, na Fase 1, as buscas retornaram um total de 1275 resultados, sendo que grande parte dos estudos retornados pelas bases WoS (99) e Scopus (288) foram identificados como duplicatas e removidos. Isso se deve ao fato de essas bases terem sido as últimas a serem pesquisadas, de modo que muitos de seus resultados já haviam sido recuperados pelos outros repositórios. Também é possível observar que na Fase 2, durante a análise dos resumos e metadados, muitos estudos foram removidos por não utilizarem técnicas de EDM e de aprendizado de máquina ou por não abordarem o problema da evasão a nível institucional ou de curso. Mais especificamente, muitos estudos tratavam do abandono de turmas/disciplinas ou de cursos online abertos e massivos (do inglês, *Massive Open Online Course* – MOOC), além de outros problemas educativos, como a previsão de notas. Além disso, considerando especialmente a base de dados do IEEE, muitos artigos se mostraram completamente fora do domínio educacional, abordando a utilização da técnica de *dropout* para a regularização de modelos de redes neurais. Finalmente, na Fase 3, o número mais significativo de descartes foi o de publicações cujos arquivos/textos completos não estavam disponíveis/acessíveis. Como as bibliotecas digitais da ACM e do IEEE, em geral, detém suas próprias publicações, estas ocorrências concentraram-se nas bases WoS e Scopus.

Na Figura 7(a) são apresentados os quantitativos de artigos resultantes de cada fase de seleção, para cada base de pesquisa. Pode-se observar que, ao término, foram selecionados 72 trabalhos, listados no Apêndice A. Na Figura 7(b) é apresentada a distribuição desses estudos por ano de publicação, sendo possível observar que nenhum artigo publicado em 2016 foi selecionado e que houve um aumento significativo no quantitativo de trabalhos selecionadas em 2019. Esses resultados sugerem um crescimento recente no interesse de aplicação de técnicas de EDM e aprendizado de máquina no problema da evasão, confirmando o potencial de exploração dessa

[8]Note que a disponibilidade de texto integral não consta entre os critérios de seleção apresentados na Seção 3.1. Ela se caracteriza mais como um critério padrão de coleta, uma vez que a seleção depende da viabilidade de análise. O mesmo se aplica à verificação de duplicatas, uma vez que não há razão para se analisar um artigo igual/semelhante a outro.

Figura 6 – Progresso das etapas de análise, considerando os critérios de seleção e qualidade especificados no protocolo da RSL

	Searching for publications and removing duplicates	ACM	IEEE	WoS	Scopus	Total
Step 1	Search result	20	414	304	537	1275
	Duplicates	0	0	99	288	387
	Subtotal	**20**	**414**	**205**	**249**	**888**
	Analyzing abstracts and other metadata	**ACM**	**IEEE**	**WoS**	**Scopus**	**Total**
Step 2	Language or format not covered	0	0	0	31	31
	Literature review	1	11	20	23	55
	Not EDM or not institutional/degree dropout prediction	7	218	69	65	359
	Less than six pages in length	2	117	10	22	151
	Data set with less than 500 students/samples	1	4	7	9	21
	Data collected by questionnaires or surveys	0	0	2	1	3
	Subtotal	**9**	**64**	**97**	**98**	**268**
	Collecting and analyzing full texts	**ACM**	**IEEE**	**WoS**	**Scopus**	**Total**
Step 3	Duplicates/Similars	0	3	0	0	3
	Full text not available	0	1	46	66	113
	Language or format not covered	0	5	4	8	17
	Literature review	0	2	0	0	2
	Not EDM or not institutional/degree dropout prediction	0	18	7	7	32
	Less than six pages in length	0	1	0	0	1
	Data set with less than 500 students/samples	0	12	3	4	19
	Data collected by questionnaires or surveys	1	1	3	0	5
	Proposal stage	1	1	1	2	5
	Subtotal	**7**	**20**	**33**	**11**	**71**

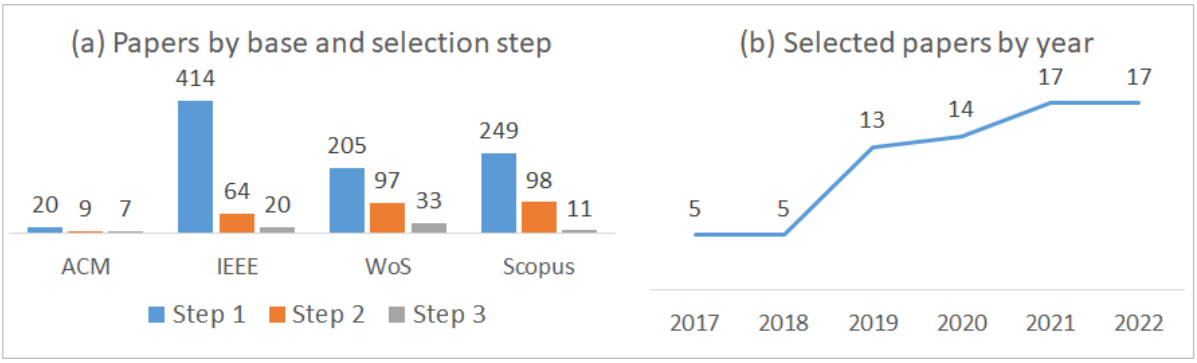
Fonte: Colpo et al. (2024).

temática de pesquisa.

Em relação à origem das publicações, é apresentado, na Figura 8, o mapeamento da quantidade de estudos selecionados por países envolvidos, considerando as afiliações dos autores de cada artigo. Para referência, o gráfico também sinaliza o continente, a Renda Nacional Bruta (RNB) per capita (linha cinza), e o Índice de Desenvolvimento Humano - IDH (linha azul) de cada país[9]. Os valores de RNB per capita e de IDH foram retirados do conjunto de dados de IDH do Programa das Nações Unidas para o Desenvolvimento, com ano de referência de 2021 (UNDP, 2021), último disponível em dezembro de 2023. De modo geral, embora o IDH considere múltiplas dimensões, é possível observar um alinhamento entre os dois indicadores, de modo que um país com menor RNB tende a ter também um menor IDH. Também é visível

[9]Note que, por não ser reconhecida pela ONU, não há informação sobre o IDH da República da China/Taiwan.

Figura 7 – Quantitativo de artigos por base e etapa de seleção (a); e distribuição dos artigos selecionados por ano (b)



Fonte: Colpo et al. (2024).

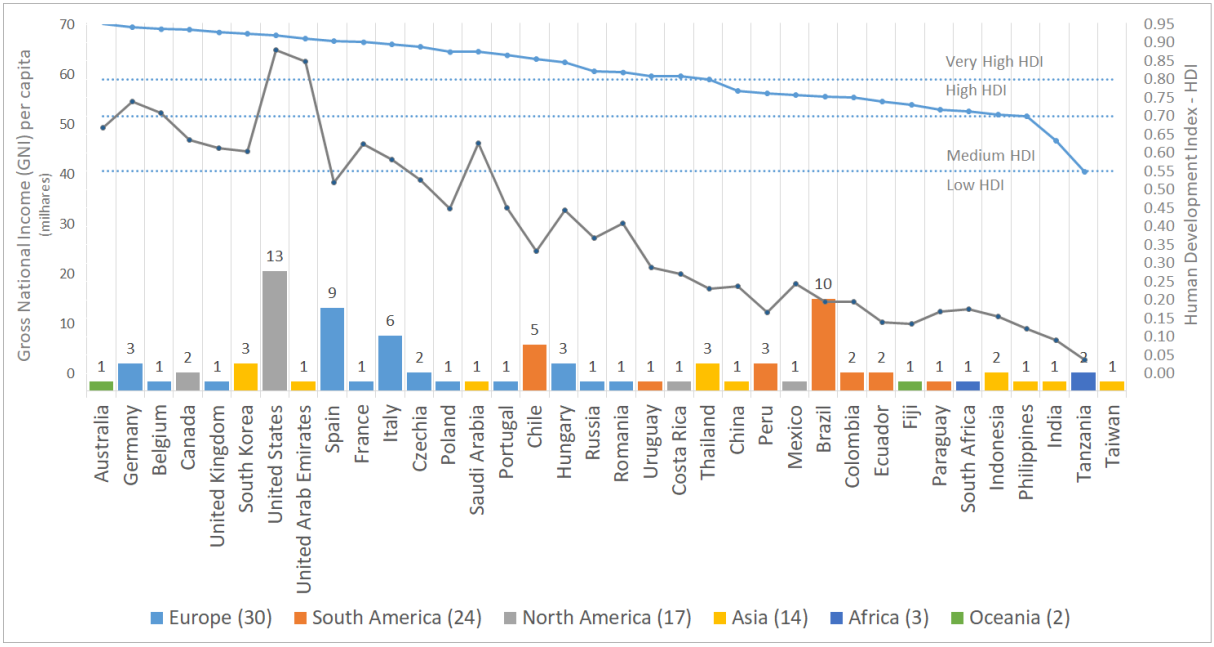
uma maior concentração[10] de trabalhos na Europa e nas Américas do Sul e Norte, com destaque aos EUA, ao Brasil, e à Espanha, que detém, respectivamente, os três maiores quantitativos de publicações. Embora a maioria dos estudos tenha sido realizada em países desenvolvidos, com maior RNB e IDH muito elevado (IDH igual ou superior a 0.8), é notório o interesse de países em desenvolvimento econômico e de alto IDH (entre 0.7 e 0.8) nessa área de pesquisa, principalmente da América do Sul. Porém, apenas dois e um dos artigos selecionados dizem respeito a países com IDH médio (IDH inferior a 0.7) e baixo (IDH inferior a 0.55), respectivamente, e com menores RNBs. Isso demonstra que o crescimento das pesquisas voltadas à aplicação de EDM ao problema da evasão estudantil não se confirma justamente nos países que mais necessitariam desse tipo de iniciativa, considerando que, como mencionado no Capítulo 1, a redução da evasão estudantil é essencial ao processo de desenvolvimento socioeconômico. Dessa forma, ainda se faz necessário difundir conhecimentos acerca dessa temática, a fim de estimular o desenvolvimento de novos estudos, principalmente em países de maior carência.

3.3 Resultados da RSL

Ao término da seleção, foi realizada a extração de informações dos artigos selecionados, a fim de responder às questões de pesquisa estabelecidas na Seção 3.1. Essas informações são sintetizadas e discutidas nesta Seção, considerando a seguinte organização: enquanto as Seções 3.3.1 e 3.3.2 apresentarão os resultados sob as perspectivas contextuais, relacionadas a Q1 e Q2; as Seções 3.3.3 e 3.3.4 descreverão características acerca dos dados e técnicas utilizados nos estudos, respondendo a Q3 e Q4, respectivamente.

[10]Note que o total de artigos de cada continente é apresentado junto a sua respectiva legenda, entre parênteses.

Figura 8 – Quantitativos de artigos selecionados por país e continente



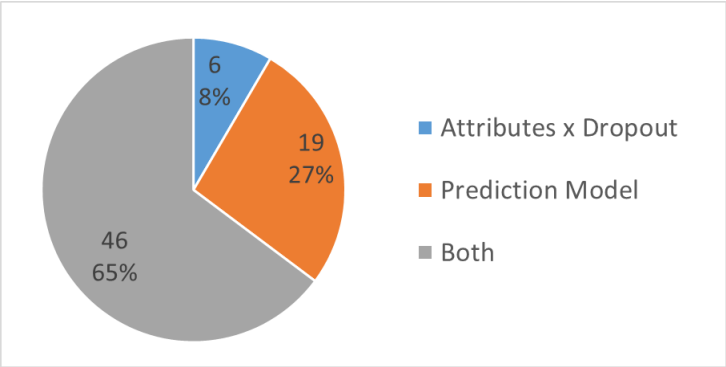
Fonte: Colpo et al. (2024).

3.3.1 Objetivos

Em relação à primeira questão de pesquisa, os trabalhos que utilizam técnicas de EDM e aprendizado de máquina no contexto da evasão estudantil, de forma geral, visam (i) a descoberta automática de padrões e atributos relacionados a esse fenômeno, a fim de melhor entender suas causas e, assim, auxiliar profissionais a prever, mediante acompanhamento pedagógico (previsão "manual"), alunos em risco; (ii) o desenvolvimento de modelos preditivos que possam identificar automaticamente alunos com padrões de evasão (previsão automática); ou (iii) ambos os objetivos anteriores, ou seja, buscam prover modelos preditivos de evasão, além de investigar padrões ou indícios dos atributos mais fortemente associadas ao abandono. Na Figura 9, é apresentada a distribuição dos artigos selecionados nesta RSL perante esses objetivos. Pode-se observar que apenas seis trabalhos (8%) se restringem à investigação de padrões e da relação entre as variáveis analisadas (atributos) e a evasão. A grande maioria dos estudos, 65 (92%), foca no desenvolvimento de modelos preditivos, dentre os quais grande parte, 46 (65% do total de artigos), também aborda a investigação dos padrões ou atributos relacionados à evasão.

Considerando a primeira categoria, em comum, os seis estudos demonstram uma forte relação entre a evasão e o baixo desempenho acadêmico dos alunos. Mais detalhadamente, Iam-On; Boongoen (2017a) utilizaram a abordagem de clusterização para construir modelos descritivos relacionados à evasão estudantil, considerando informações demográficas/sociais e acadêmicas (escolares e universitárias) de estudantes da

Figura 9 – Distribuição dos artigos selecionados perante seus objetivos



Fonte: Colpo et al. (2024).

Mae Fah Luang University (Tailândia). Foram gerados grupos considerando, separadamente, estudantes recém admitidos e que concluíram o primeiro ano de curso. A partir da análise do perfil representativo de cada grupo, os autores identificaram que alunos com alto desempenho escolar continuam tendo bom rendimento na universidade, e que alunos de desempenho escolar inferior tendem a evadir após o primeiro ano do curso.

Yoo; Woo; Park (2017) e Assis et al. (2022) utilizaram dados acadêmicos e demográficos, e apenas acadêmicos, na mineração de regras de associação para identificar padrões relacionados com a desistência de alunos de Ciências da Computação de uma universidade pública nos EUA e de Engenharia de Produção do Centro Federal de Educação Tecnológica do Rio de Janeiro (Brasil), respectivamente. Através da mineração de padrões sequenciais (que também envolve a temporalidade) sobre o histórico de disciplinas cursadas, Yoo; Woo; Park (2017) verificaram que os alunos tendem a evadir após cursarem - normalmente mais de uma vez, devido a reprovações - disciplinas introdutórias de programação introdutória e matemática. Assis et al. (2022) efetuaram a extração de regras de associação em dados de desempenho dos alunos, aumentados por métricas de redes sociais, em que o grau de propagação da média de notas (*Grade Point Average* – GPA) foi utilizado como *proxy* da ligação existente entre os alunos. Como resultado, os autores identificaram que o baixo desempenho escolar e a pouca participação na rede social levam ao atraso na graduação e à evasão.

A partir de dados acadêmicos e demográficos de estudantes de engenharia de seis universidades da União Europeia, Prada et al. (2020) desenvolveram uma ferramenta web para apoio à tutoria. Além de utilizar técnicas de redução de dimensionalidade, agrupamento, e visualização para permitir a análise exploratória dos dados, a ferramenta constrói/utiliza modelos de classificação do desempenho acadêmico e da evasão com um objetivo descritivo. Ou seja, fornecer pistas globais sobre o comportamento dos estudantes. Entre os padrões encontrados, os autores apontaram que a

idade de ingresso apresentou impacto negativo na graduação, enquanto a nota de admissão e o desempenho do aluno no primeiro semestre demonstrou impacto positivo.

Também considerando dados acadêmicos e demográficos dos estudantes, Yang; Olson; Puder (2021) empregaram técnicas de análise estatística exploratória, correlação e classificação, a fim de investigar como disciplinas individuais e sequências de disciplinas influenciam na evasão e na conclusão de estudantes de Ciência da Computação, na *San Francisco State University* (EUA). No que diz respeito ao emprego de classificação, foram desenvolvidos modelos preditivos de evasão com base nos classificadores *Random Forest* (RF), *Logistic Regression* (LR) e *Support Vector Machine* (SVM), utilizando dados de desempenho em disciplinas relacionadas, aproximadamente, a 2, 3, 4 e 5 períodos letivos. Entre os resultados, analisando os *rankings* de importância dos modelos, os autores identificaram que o desempenho nas disciplinas de matemática e física tem maior impacto preditivo na fase inicial do curso.

Por fim, Perchinunno; Bilancia; Vitale (2019) desenvolveram modelos LR e DT, considerando dados demográficos e acadêmicos de alunos que concluíram o primeiro ano de graduação ou de mestrado de ciclo único (combinação de graduação e mestrado), na *University of Bari Aldo Moro* (Itália). Os modelos foram utilizados descritivamente, na análise de atributos e padrões relacionados às evasões ocorridas entre o primeiro e o segundo ano de universidade. Dentre os resultados, foi identificado que o risco de evasão é maior para alunos do sexo masculino ou que tenham integralizado menos de doze créditos no primeiro ano.

Em geral, as pesquisas que desenvolvem modelos preditivos demonstram maior preocupação com o refinamento dos resultados, explorando e avaliando diferentes técnicas, tanto no pré-processamento dos dados, quanto na geração dos modelos de classificação. Como o processo de avaliação da qualidade dos modelos preditivos envolve, no mínimo, simular sua performance perante um conjunto de dados de teste, considerando diferentes métricas avaliativas, talvez haja uma maior clareza da necessidade de melhoria e, consequentemente, uma busca mais efetiva por técnicas que contribuam para o aumento do desempenho e da confiabilidade desses modelos.

Gamao; Gerardo (2019) e Queiroga et al. (2020), por exemplo, além de avaliarem diferentes algoritmos de classificação no desenvolvimento dos modelos preditivos, utilizaram algoritmos evolutivos para otimizar os resultados de classificação. Em Gamao; Gerardo (2019) foi proposto um algoritmo de otimização por enxame para selecionar um subconjunto de atributos ótimo, que maximizasse a acurácia de modelos preditivos de evasão. Para isso, foram considerados dados demográficos/sociais, acadêmicos e econômicos de calouros da *Davao del Norte State College* (Filipinas), além dos classificadores *Naive Bayes* (NB) e DT, tendo esse último demonstrado melhores resultados preditivos. Já Queiroga et al. (2020) propuseram a utilização de um algoritmo genético na otimização de hiperparâmetros de modelos preditivos de evasão, consi-

derando dados de contagem de interações dos alunos com um AVA, em um curso técnico a distância do Instituto Federal Sul Rio-grandense (Brasil). Na abordagem proposta, modelos baseados nos algoritmos DT, LR, RF, *Multilayer Perceptron* (MLP), e *Adaptive Boosting* (AdaBoost), com diferentes configurações de hiperparâmetros, competiram entre si. Assim, ao término do processo evolutivo, o melhor classificador e sua melhor combinação de hiperparâmetros foram utilizados. Comparando com o método de otimização *Grid Search* e os modelos configurados de forma padrão, os autores demonstraram o melhor desempenho preditivo de sua abordagem.

Em Barros et al. (2019) também foram considerados dados de estudantes de cursos técnicos de um Instituto Federal Brasileiro, o do Rio Grande do Norte, mas dos tipos acadêmicos, demográficos e econômicos, e de alunos de cursos presenciais integrados[11]. Além de uma otimização automática de hiperparâmetros e da avaliação dos algoritmos de classificação DT, MLP, e *Balanced Bagging*, foram testadas diferentes técnicas de balanceamento e métricas avaliativas. Como resultado, o classificador *Balanced Bagging*, sem técnica de balanceamento adicional, superou o desempenho dos demais modelos, associados a diferentes técnicas de balanceamento. Além disso, os autores identificaram que as métricas UAR e *G-mean* se mostraram mais adequadas para capturar o erro da classe minoritária, no contexto de dados desbalanceados.

Utilizando informações demográficas e acadêmicas de um conjunto de dados público de estudantes de escolas da Índia, Masood; Begum (2022) também avaliaram diferentes métricas e técnicas de reamostragem para tratar o desbalanceamento dos dados, no desenvolvimento de classificadores LR e SVM. Os autores identificaram que modelos SVM construídos após a aplicação das técnicas de subamostragem aleatória e SMOTE (*Synthetic Minority Over-sampling Technique*) produziram resultados superiores. Além disso, Masood; Begum (2022) apontaram maior eficácia da métrica AUC-ROC na avaliação do desempenho preditivo da classe minoritária.

Em um último exemplo, no trabalho de Solis et al. (2018) foram utilizados dados acadêmicos, demográficos e econômicos de alunos de graduação do Instituto Tecnológico de Costa Rica. Os autores avaliaram os classificadores RF, MLP, LR e SVM, considerando diferentes perspectivas de representação dos dados dos estudantes. Essas perspectivas tinham como objetivo avaliar se o aprendizado dos modelos preditivos seria beneficiado pela inclusão de alunos ativos (matriculados) dentre os exemplos da classe de não evadidos e se o comportamento de evasão seria melhor representado a partir de toda a trajetória (um registro de dados para cada semestre cursado pelo aluno) ou apenas do último semestre do aluno. Como resultado, além de apontarem o melhor desempenho do RF, os autores identificaram ser benéfico utilizar dados de todos os semestres cursados e considerar como exemplos negativos de evasão

[11] Em cursos integrados, o processo formativo dos níveis médio e técnico ocorrem de forma integrada e simultânea.

apenas registros de estudantes formados/concluídos.

Por fim, dentre os estudos que desenvolvem modelos preditivos, os que também investigam padrões ou atributos mais fortemente relacionados à evasão costumam fazer isso a partir da aplicação de técnicas que forneçam a importância de cada atributo ou da análise do próprio classificador, caso se trate de um modelo interpretável. Costa et al. (2021) e Naseem; Chaudhary; Sharma (2022), por exemplo, construíram modelos preditivos de evasão, a partir de dados de estudantes de Ciência da Computação, da Universidade Federal de Pelotas (Brasil) e da *University of the South Pacific* (Fiji), respectivamente. Enquanto Costa et al. (2021) utilizaram dados acadêmicos e socioeconômicos dos três primeiros semestres dos alunos no desenvolvimento de modelos DT, RF e LR; Naseem; Chaudhary; Sharma (2022) consideraram dados acadêmicos, econômicos e interacionais (da interação com um AVA) do primeiro ano dos estudantes para construir modelos DT, RF, NB, LR e KNN (K-Nearest Neighbor), considerando três momentos de predição distintos (matrícula e final do primeiro e do segundo semestre). Os *rankings* de importância estabelecidos pelos modelos e os resultados do algoritmo de seleção de atributos Boruta foram analisados pelo primeiro e segundo estudo, respectivamente, para identificar os atributos com maior influência na evasão. Em ambos os trabalhos, atributos associados ao desempenho acadêmico dos estudantes demonstraram maior importância preditiva. Enquanto Costa et al. (2021) apontaram melhor desempenho do modelo RF, Naseem; Chaudhary; Sharma (2022) identificaram que os modelos NB e LR se sobressaíram para predições no momento de matrícula e ao final dos semestres, respectivamente.

Berka; Marek (2021) construíram diferentes modelos preditivos de evasão, considerando dados acadêmicos e demográficos disponíveis na admissão e ao término de cada um dos quatro semestres iniciais de alunos de cursos de bacharelado, presenciais ou EAD, de uma universidade da República Tcheca. Foram gerados modelos para cada momento de predição e diferentes combinações de atributos, utilizando os algoritmos DT, LR e RF, e considerando dentre os exemplos de evasão (i) apenas estudantes que foram desligados pela universidade, por não cumprirem alguma regra/requisito acadêmico; ou (ii) incluindo também os alunos que abandonaram o curso por conta própria. Como resultado, os autores identificaram que a primeira forma melhora ligeiramente o aprendizado da classe evadida, e que os classificadores LR apresentaram melhor desempenho preditivo. Além disso, para verificar a influência dos atributos e padrões relacionados à evasão, os autores interpretaram os modelos DT e realizaram uma análise de dependência, utilizando diferentes algoritmos de regras de associação e apenas dados disponíveis na admissão, para identificar diferenças entre os estudantes que seguiam ou não no curso após o primeiro semestre. Dentre os resultados, os autores identificaram o percentual de créditos perdidos no último semestre como o atributo mais importante para os modelos preditivos, enquanto a análise de dependên-

cia apontou o tempo entre o término do ensino médio e o ingresso no ensino superior como o fator de maior risco para o abandono no primeiro semestre.

Embora menos comum por ser uma área de pesquisa mais recente, técnicas de inteligência artificial explicável (do inglês, *eXplainable Artificial Intelligence* – XAI), em especial a abordagem *Shapley Additive Explanations* (SHAP), também aparecem como um recurso para auxiliar na interpretação de modelos caixa preta. Em Baranyi; Nagy; Molontay (2020), por exemplo, foi proposta a aplicação de um modelo baseado na arquitetura *Fully Connected Deep Neural Network* (FCNN) na previsão de evasão. Esse, outros dois modelos de redes neurais profundas (TabNet, e BaggingFCNN), e dois *ensembles* de DTs (RF e XGBoost) foram treinados e avaliados a partir de dados acadêmicos e demográficos, disponíveis na admissão de estudantes de graduação da *Budapest University of Technology and Economics* (Hungria). Os autores observaram um desempenho preditivo ligeiramente superior da FCNN e, para entender melhor as decisões de seu modelo, avaliaram a importância dos atributos com base em permutações e na abordagem SHAP. Como resultado, identificaram que o tempo entre o término do ensino médio e o ingresso no ensino superior, a pontuação de admissão e a nota de matemática no exame de admissão demonstraram os maiores potenciais preditivos.

Semelhantemente, Bassetti et al. (2022) utilizaram valores SHAP para medir o impacto dos atributos nas predições de um modelo baseado no algoritmo *Gradient Boosting Trees* (GBT). Mais especificamente, os autores apresentaram o ISIDE, o protótipo de um sistema de alerta de evasão integrado ao portal *online* de discentes da *Sapienza University of Rome* (Itália). O ISIDE funciona de forma assíncrona, fazendo previsões com base em dados coletados semanalmente e retornando uma lista com a probabilidade de evasão dos alunos. Antes de optarem pelo melhor desempenho do GBT na tarefa preditiva, os autores também avaliaram modelos DT, RF, LR, SVM e MLP, todos construídos a partir de dados demográficos, econômicos e acadêmicos dos estudantes de graduação e pós-graduação da universidade. Analisando os valores SHAP, Bassetti et al. (2022) identificaram que os alunos com mais tempo desde a última matrícula, com menos créditos integralizados ou com médias de notas mais baixas têm maior probabilidade de evadir.

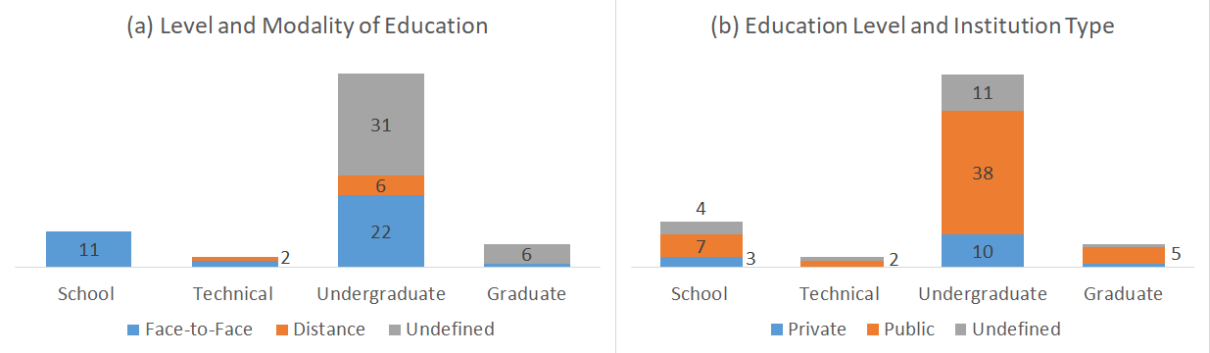
3.3.2 Níveis, modalidades e redes de ensino

Também relacionadas ao contexto, mas respondendo à segunda questão de pesquisa, as Figuras 10(a) e 10(b) apresentam o número de artigos selecionados[12] por níveis e, respectivamente, modalidades ou redes de ensino consideradas. Analisando

[12]Note que um mesmo artigo pode ser contabilizado em mais de uma categoria de cada gráfico, segundo os dados explorados. Essa observação é válida para todos os gráficos apresentados neste trabalho.

a Figura 10(a), percebe-se pouca exploração nos níveis de educação básica e técnica, sendo a grande maioria dos trabalhos vinculada à Graduação, o que é compreensível, à medida que essas pesquisas costumam ser desenvolvidas em universidades, explorando dados próprios. Além disso, sistemas de gestão acadêmica e AVAs são utilizados, de forma mais ampla, por instituições de nível superior, gerando maior disponibilidade de dados nesse nível de ensino. Em relação à modalidade, pode-se observar que os trabalhos selecionados tratam majoritariamente do ensino presencial[13]. Porém, esse resultado pode estar influenciado pelo interesse de pesquisa desta revisão, que restringe a evasão ao escopo institucional e de curso, desconsiderando estudos ao nível de disciplinas, que parecem ser uma tendência na modalidade EAD, por explorarem a disponibilidade e a atualização constante de dados interacionais dos alunos em turmas dos AVAs. Considerando a Figura 10(b), também fica evidente que a grande maioria dos trabalhos aborda a evasão em instituições públicas. Naturalmente, a maioria dos trabalhos apresentados na Seção 3.3.1 acompanham esses resultados, focando no ensino superior, público e presencial. Porém, também se faz importante descrever as poucas iniciativas que consideram os demais níveis, modalidades e redes de ensino, a fim de exemplificar e incentivar a expansão de pesquisas nesses contextos.

Figura 10 – Quantitativo de artigos por nível e modalidade (a), e nível e rede (b) de ensino



Fonte: Colpo et al. (2024).

Dentre os onze estudos voltados à educação básica, Masood; Begum (2022) já foi apresentado. Sorensen (2019) e Orooji; Chen (2019) utilizaram dados acadêmicos, socioeconômicos e comportamentais, obtidos a partir de bases de dados administrativas de escolas públicas dos estados norte-americanos da Carolina do Norte e da Luisiana, respectivamente. No primeiro trabalho, que utilizou dados do terceiro ao oitavo ano do ensino fundamental para prever a evasão de estudantes matriculados no ensino médio, modelos de SVM e *boosting* de DTs demonstraram, nessa ordem, me-

[13]Note que, embora muitos trabalhos tenham sido categorizados com modalidade "*Undefined*", por não mencionarem ou darem indícios dessa informação, provavelmente esses estudos tratam da modalidade presencial, dada sua predominância no ensino formal.

lhores acurácias, quando comparados a modelos preditivos mais simples. Porém, a partir da interpretação de modelos DT e LR, e da análise de importância dos atributos de um *bagging* de DTs, Sorensen (2019) identificou que atributos comportamentais e acadêmicos apresentaram maior poder preditivo. Por sua vez, ao avaliar diferentes algoritmos de classificação e de balanceamento de dados, Orooji; Chen (2019) apontou que as técnicas de balanceamento beneficiaram a revocação dos modelos.

Já em Chung; Lee (2019) e Lee; Chung (2019), foram utilizados dados comportamentais/disciplinares e acadêmicos de estudantes de ensino médio, oriundos do sistema nacional de administração educacional da Coreia do Sul. Enquanto no primeiro trabalho foi realizado o balanceamento dos dados e desenvolvido um modelo RF, no segundo foi apresentada uma análise comparativa acerca do uso de sobreamostragem, considerando o classificador RF e um *boosting* de DTs. Como resultados, além de validarem o desempenho preditivo do modelo, os autores verificaram, no primeiro estudo, que atributos relacionados a ausências e atrasos não autorizados apresentaram, respectivamente, as maiores importâncias preditivas. Na segunda pesquisa, identificaram que o *boosting* de DTs sem a aplicação da técnica de balanceamento obteve o melhor desempenho, demonstrando a necessidade de avaliar técnicas de interesse, antes de incluí-las em uma solução.

Mduma; Machuve (2021) e Mduma; Kalegele; Machuve (2019) utilizaram conjuntos de dados públicos com dados acadêmicos, demográficos e econômicos de alunos da Tanzânia, Quênia e Uganda, e apenas da Tanzânia, respectivamente, para desenvolver modelos de previsão do abandono escolar. Ambos os estudos utilizaram técnicas de balanceamento de dados, assim como permutação das importâncias dos atributos para identificar as características com maior influência na evasão. Como resultado, Mduma; Machuve (2021) apontou o melhor desempenho do modelo LR quando comparado com os modelos MLP e RF, bem como a maior importância de atributos relacionados a gênero, idade, rendimento, verificação de livros pelos pais, e refeições diárias. Por sua vez, embora não apresentem dados comparativos ou de desempenho preditivo, Mduma; Kalegele; Machuve (2019) reconheceram como melhor modelo um *ensemble* desenvolvido pela combinação suave de modelos LR e MLP otimizados. Além dos atributos identificados em Mduma; Machuve (2021), Mduma; Kalegele; Machuve (2019) também apontaram como importantes atributos que indicam se o aluno lê livros com os pais e se os pais discutem o progresso da criança com o professor. Mduma; Kalegele; Machuve (2019) selecionaram esses atributos importantes e os utilizaram como entrada em um protótipo de sistema web. Baseado no modelo *ensemble*, o sistema foi desenvolvido não só para prever se um aluno abandonará a escola com base nas informações introduzidas, mas também para apresentar uma visualização das escolas com mais evasões.

Considerando a ameaça à utilidade dos modelos preditivos em decorrência dos

desvios de conceito durante a pandemia de COVID-19, Xu; Wilson (2021) utilizaram simulações baseadas em imputação para analisar o impacto da qualidade e disponibilidade dos dados no desempenho do modelo de previsão de evasão implementado no *Early Warning System* (EWS) de *Rhode Island* (EUA). O EWS conta com um modelo RF treinado a partir de dados acadêmicos, demográficos e econômicos de alunos do ensino médio público de *Rhode Island*, considerando técnicas de balanceamento. Ao construir e avaliar modelos com diferentes estratégias de imputação e simulações de desvios de conceitos, os autores descobriram que, em determinadas circunstâncias, apesar do contexto pandêmico, alguns modelos preditivos ainda poderiam ser úteis na tomada de decisões.

Crespo (2020) simulou e comparou a eficácia de mecanismos baseados na previsão da evasão escolar e em indicadores de renda (*income-proxy means test* – PMT) para identificar pobres e alunos em risco de evasão, no contexto de direcionamento de políticas sociais e de transferência de renda. Para isso, o autor utilizou dados acadêmicos, demográficos e econômicos de bases de dados governamentais chilenas, abrangendo alunos a partir do sétimo ano, presentes no arquivo de proteção social do Chile. Os modelos GBT, SVM, RF, *elastic net*, *lasso*, e *generalized additive models* foram avaliados na predição da evasão escolar, tendo o *elastic net* demonstrado melhor desempenho. Além disso, com base nos *rankings* dos modelos, Crespo (2020) identificou como mais importantes os atributos de idade, notas, frequência, série/ano escolar, e taxa de evasão da escola. O autor também observou que o uso de resultados preditivos em conjunto com o PMT aumentou a eficácia do direcionamento, exceto quando a avaliação social dos pobres e futuros desistentes é muito diferente.

Queiroga et al. (2022) descrevem um projeto de *learning analytics* que visa mitigar a evasão e a retenção escolar, em âmbito nacional, no ensino secundário do Uruguai. Utilizando dados acadêmicos (do primeiro ano do ensino primário ao segundo ano do ensino secundário) e socioeconômicos dos alunos, os autores construíram oito modelos RF para prever a evasão em diferentes momentos (antes do início do ano letivo e após a primeira reunião de avaliação de cada ano) dos dois primeiros anos do ensino secundário básico regular e técnico. Além de considerarem técnicas de balanceamento, seleção de atributos e otimização na construção dos modelos, Queiroga et al. (2022) realizaram uma análise de viés, considerando três atributos protegidos, o que resultou na aprovação de sete dos oito modelos. O projeto ainda desenvolveu uma API web, para permitir o uso e a disponibilização dos resultados/previsões dos modelos aprovados, tanto de forma síncrona quanto assíncrona.

Assim como Queiroga et al. (2022), Barros et al. (2019) também considera a educação técnica e escolar, mas no contexto do ensino técnico integrado brasileiro. Dessa forma, além de estarem vinculados ao nível escolar, esses trabalhos também são contabilizados no nível técnico, juntamente com o de Queiroga et al. (2020). Barros et al.

(2019) e Queiroga et al. (2020) já foram apresentados na Seção 3.3.1, assim como Perchinunno; Bilancia; Vitale (2019) e Bassetti et al. (2022), que abordaram a evasão em cursos de graduação e pós-graduação, mesma situação de Oreshin et al. (2020), Del Bonifro et al. (2020) e Fernández-García et al. (2021).

Em Oreshin et al. (2020) foram utilizados dados acadêmicos, demográficos e econômicos, estáticos (disponíveis na admissão) e dinâmicos (adicionados semestralmente), de estudantes de uma universidade pública russa, a *University of Information Technologies, Mechanics and Optics*. Além disso, dentre as informações dinâmicas, foram considerados dados da interação e da avaliação de sentimentos das publicações dos alunos em uma rede social do país. Assim, tendo a evasão como variável alvo, além do modelo precoce, destinado a fornecer previsões imediatamente após o ingresso do aluno, foram desenvolvidos oito modelos preditivos, considerando dados extraídos de um até oito semestres consecutivos de curso e o algoritmo GBT. Como resultado, os autores destacaram bons resultados preditivos e apontaram o tipo de certificação do ensino médio, a cidade natal e o centro acadêmico/área de conhecimento do aluno como os três principais atributos do modelo precoce. Porém, nos modelos dinâmicos, atributos de desempenho acadêmico em semestres anteriores demonstraram maior potencial preditivo.

Del Bonifro et al. (2020) utilizaram dados sociais e acadêmicos (escolares e universitários) de estudantes de cursos de graduação e de ciclo único (combinação de graduação e mestrado) de uma universidade não identificada, para construir modelos de predição precoce de evasão, considerando apenas dados disponíveis na admissão ou incluindo dados disponíveis antes do término do primeiro ano de curso. Considerando a subamostragem aleatória para tratar o desbalanceamento dos dados, além de diferentes combinações de hiperparâmetros, no treinamento, os autores desenvolveram modelos baseados nos algoritmos SVM, RF, e *Linear Discriminant Analysis* (LDA), para os diferentes conjuntos de atributos. Como resultado, os autores optaram por modelos SVM e RF para os cenários com e sem atributos relacionados ao primeiro ano de curso, respectivamente. Além disso, apesar de ambos os modelos serem necessários e úteis para a antecipação de riscos, foi observada uma grande melhora no desempenho preditivo ao introduzir dados relacionados ao primeiro ano de curso.

Dada a dificuldade de acesso a dados pessoais e a questões de privacidade, Fernández-García et al. (2021) buscaram desenvolver os melhores modelos possíveis de previsão de evasão com base apenas em dados socioeconômicos e acadêmicos[14] dos estudantes, considerando o contexto de cursos de graduação e mestrado

[14]Embora os autores mencionem utilizar apenas dados acadêmicos, dentre os atributos utilizados constam informações relativas à data de nascimento e ao município do estudante, ao qual consideramos sociais/demográficos. Além disso, é considerado um indicador de recebimento de bolsa/auxílio do governo, que, segundo os próprios autores, está diretamente relacionado com a situação econômica da família do estudante.

de uma escola de engenharia de uma universidade pública espanhola. Para isso, os autores utilizaram diversas técnicas de engenharia de atributos e instâncias no pré-processamento, incluindo balanceamento/reamostragem. Considerando otimização de hiperparâmetros, Fernández-García et al. (2021) desenvolveram classificadores GBT, RF e SVM, juntamente com um *ensemble* heterogêneo desses modelos, para cinco etapas preditivas: na matrícula/admissão e ao final de cada um dos quatro primeiros semestres. Como diferencial, com exceção dos modelos destinados à primeira etapa, os demais utilizaram entre seus atributos o resultado preditivo do modelo da etapa imediatamente anterior. Como resultado, Fernández-García et al. (2021) apontou melhor desempenho do GBT no momento da matrícula e do *ensemble* heterogêneo nas predições ao final do primeiro e do segundo semestre.

Por outro lado, Alturki; Cohausz; Stuckenschmidt (2022) e Shilbayeh; Abonamah (2021) consideraram apenas dados de estudantes de pós-graduação, mais especificamente de mestrado. Alturki; Cohausz; Stuckenschmidt (2022) utilizaram dados demográficos e de desempenho acadêmico de estudantes de mestrado em Informática Empresarial, na *University of Mannheim* (Alemanha). Considerando a técnica de sobreamostragem SMOTE, os autores desenvolveram modelos LR, RF, KNN, NB, SVM e ANN (*Artificial Neural Networks*) para prever a situação (conclusão/não conclusão do curso) e o desempenho acadêmico final (média, acima da média e abaixo da média) dos estudantes, ao fim do primeiro e do segundo semestre. Entre os resultados, Alturki; Cohausz; Stuckenschmidt (2022) identificaram melhor desempenho nos modelos RF treinados após a sobreamostragem, além de melhores resultados após o segundo semestre. Além disso, usando de permutação para verificar a importância dos atributos no RF, os autores identificaram as notas do semestre e a distância da residência à universidade como as características mais importantes para prever os resultados dos alunos. Por sua vez, Shilbayeh; Abonamah (2021) utilizaram dados socioeconômicos e acadêmicos (de admissão e anteriores) de estudantes de mestrado da *Abu Dhabi School of Management* (Emirados Árabes Unidos). Os autores desenvolveram um *boosted regression tree ensemble* para prever o número de matrículas de estudantes nos próximos anos letivos e aplicaram o algoritmo Apriori para extrair regras de associação e, assim, identificar padrões relacionados aos estudantes matriculados e aos que tem maior probabilidade de evadir. Como resultado, o modelo de *ensemble* demostrou melhor performance, quando comparado a uma única *boosted regression tree*. Além disso, as características relacionadas à faixa etária de 30-40 anos e à faixa de GPA de 2.5-3 demostraram recorrência entre as regras associadas à evasão.

Queiroga et al. (2020), Berka; Marek (2021), Kang; Wang (2018), Yu; Lee; Kizilcec (2021), e Ortigosa et al. (2019) são exemplos de artigos que consideram o contexto da EAD. Kang; Wang (2018) e Yu; Lee; Kizilcec (2021) utilizaram dados acadêmicos e so-

ciais/demográficos de alunos de graduação de universidades públicas dos EUA para desenvolver modelos destinados a prever evasões durante as férias de inverno e o primeiro ano de curso, respectivamente. Kang; Wang (2018) desenvolveram modelos de predição logística usando ou não a subamostragem aleatória dos dados, e, como cada modelo favoreceu ou a precisão ou a revocação, ambos foram utilizados para prever o risco de evasão de estudantes ativos, gerando duas listagens de alunos em risco aos gestores educacionais. Em Yu; Lee; Kizilcec (2021) foram produzidos modelos separados para alunos de cursos EAD e presenciais, a partir de dados disponíveis ao término do primeiro período do curso. Os modelos foram gerados a partir dos algoritmos LR e GBT, considerando ou não a inclusão de atributos sensíveis, como gênero e cor/raça, a fim de avaliar se esses dados produziriam modelos discriminatórios. Como resultado, após avaliarem o desempenho geral e a justiça das predições (variação do desempenho preditivo em relação a grupos minoritários), os autores identificaram que os modelos GBT apresentaram melhores desempenhos e que atributos sensíveis tiveram pouco impacto na justiça algorítmica das predições, sendo preferível utilizá-los.

Além de descrever o desenvolvimento e a avaliação de modelos preditivos, Ortigosa et al. (2019) apresentaram os desafios e as mudanças técnicas impostas pela implantação de um sistema de previsão contínua do risco de evasão de estudantes de graduação da *Universidad a Distancia de Madrid*. Baseado em dados sociais, econômicos, acadêmicos e interacionais, o sistema fornece previsões (i) precoces, que consideram dados disponíveis nas matrículas dos alunos em seus respectivos anos/ciclos acadêmicos; e (ii) periódicas, que incluem dados da interação dos estudantes no AVA, coletados e atualizados periodicamente. Para isso, dividindo o ano acadêmico em 10 períodos mensais, os autores desenvolveram 22 modelos especializados (um estático e 10 periódicos para cada tipo de aluno – novato ou veterano), a partir de um algoritmo DT. Como resultados, sob a perspectiva dos modelos, além de desempenhos considerados satisfatórios institucionalmente, os autores identificaram que idade e forma de ingresso foram atributos considerados importantes em modelos de alunos novatos, mas ignorados por modelos de veteranos, que passaram a considerar atributos relacionados à performance acadêmica dos alunos em anos/ciclos anteriores. Acompanhando esse comportamento, quanto maior a quantidade de dados considerada pelos modelos, melhor foi o desempenho apresentado.

Assim como Ortigosa et al. (2019), outros 13 trabalhos se relacionam à rede privada. Perez; Castellanos; Correal (2018) e Hannaford; Cheng; Kunes-Connell (2021), por exemplo, dedicaram suas pesquisas a cursos de graduação específicos, desenvolvendo modelos para prever a evasão de estudantes de Engenharia de Sistemas e o risco de estudantes de Enfermagem não se graduarem, considerando universidades privadas da Colômbia e dos EUA, respectivamente. Perez; Castellanos; Correal (2018) avaliaram classificadores DT, LR, NB, e RF treinados a partir de atributos de-

mográficos/sociais e acadêmicos. Já Hannaford; Cheng; Kunes-Connell (2021) desenvolveram 126 modelos preditivos, considerando diferentes combinações entre dados demográficos e acadêmicos (do ensino médio da universidade); oito classificadores e um *ensemble* baseado na média ponderada dos resultados desses classificadores; e cinco momentos de predição (início de diferentes anos acadêmicos). Como resultado, Perez; Castellanos; Correal (2018) identificaram que o modelo RF demonstrou desempenho superior, enquanto o *ensemble* obteve melhor resultado em Hannaford; Cheng; Kunes-Connell (2021), que também apontou que quanto mais avançado o momento de predição maior o desempenho do modelo. Em comum, ambos os trabalhos identificaram que dados relacionadas ao desempenho acadêmico universitário possuem maior importância preditiva, incluindo, dentre os atributos mais importantes, a média cumulativa de notas e a média de notas em disciplinas específicas do curso.

Hutagaol; Suharjito (2019) também propuseram combinar os resultados de três modelos preditivos (*K-Nearest Neighbor* -– KNN, NB e DT) em um *ensemble*, utilizando, porém, um meta-classificador GBT para prever, a partir de dados demográficos, econômicos e acadêmicos, a evasão de estudantes da Faculdade de Ciências Sociais e Políticas de uma universidade privada da Indonésia. No pré-processamento, os autores realizaram o balanceamento dos dados e a seleção de atributos, considerando a técnica SMOTE e a quantização vetorial (em Inglês, *Learning Vector Quantization* – LVQ), respectivamente. Como resultado, além de o *ensemble* superar o desempenho dos classificadores individuais, a LVQ apontou maior importância preditiva de atributos relacionados à assiduidade e ao desempenho acadêmico dos alunos.

Utilizando dados demográficos, acadêmicos e econômicos de estudantes de cursos presenciais de universidades públicas e privadas, oriundos do Censo e dos Indicadores de Fluxo da Educação Superior Brasileira, da Silva et al. (2019) construíram e avaliaram três *baggings* de regressores homogêneos, que consideraram, individualmente, regressores do tipo linear, *robust* ou *ridge*, além de reproduzirem um *ensemble* heterogêneo proposto por outro trabalho, com sete algoritmos base combinados por um meta-regressor ridge. Esses quatro *ensembles* de regressão foram treinados a partir de dois subconjuntos distintos de atributos, selecionados, respectivamente, pelos métodos *Stepwise* e correlação de Pearson. Como resultado, os *ensembles* homogêneos superaram os heterogêneos, com destaque ao *bagging* de regressores lineares, treinado sobre os atributos selecionados pelo método *Stepwise*, que apresentou menores taxas de erros. Além disso, os autores identificaram como os atributos de maior importância preditiva as taxas de permanência e de conclusão do curso, a quantidade de estudantes remanescentes e o turno de estudos do aluno.

Em Palacios et al. (2021), a partir de dados acadêmicos (escolares e universitários), demográficos e econômicos de estudantes de graduação da *Universidad Católica del Maule* (Chile), foram desenvolvidos quatro tipos de modelos, destinados a

prever a evasão de forma (i) global (independente de período), ou especializada no (ii) primeiro, (iii) segundo ou (iv) terceiro ano de curso. Considerando a técnica de balanceamento SMOTE, foram treinados classificadores DT, KNN, LR, NB, RF, e SVM, para os quatro objetivos de predição. Porém, apenas os modelos destinados a prever abandonos no segundo e no terceiro ano consideraram dados de desempenho universitário. Como resultados, os autores demonstraram a importância do balanceamento, a boa performance preditiva e a superioridade dos modelos RF. Além disso, por meio da análise do ganho de informação, os autores identificaram que atributos relacionados ao desempenho escolar e ao contexto socioeconômico do local de origem do aluno apresentaram maior importância para os três primeiros cenários, embora atributos de desempenho universitário também tenham se mostrado relevantes no terceiro contexto. Já para a previsão de evasões no terceiro ano de curso (quarto cenário), atributos relacionados ao desempenho universitário dos estudantes mostraram-se mais importantes.

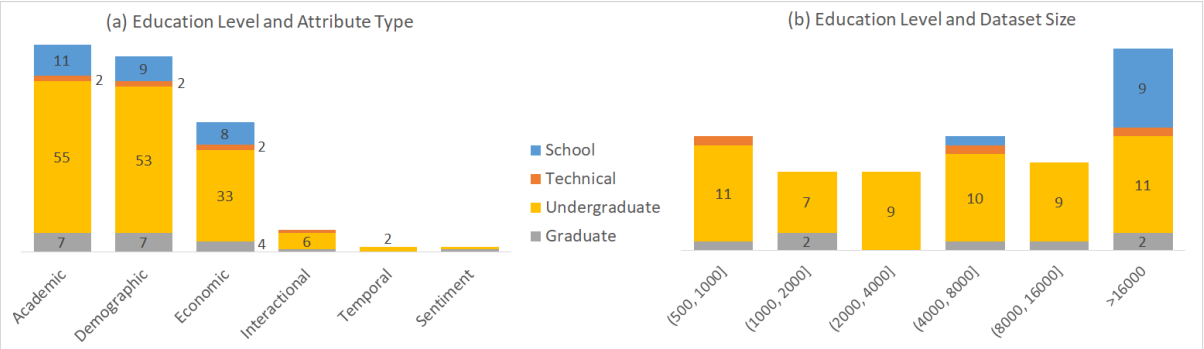
Como um último exemplo, Shiau (2020) abordou o problema da evasão de estudantes de graduação de tempo integral, em um campus de uma universidade privada e tecnológica do Taiwan. O autor utilizou uma tabela de decisão e uma DT para analisar, descritivamente, as regras determinantes de evasão, a fim de auxiliar no estabelecimento de estratégias de aconselhamento e na identificação de alunos em risco. Ambas as técnicas relacionaram as regras de evasão, majoritariamente, ao menor desempenho acadêmico. Considerando o modelo DT, Shiau (2020) identificou, por exemplo, que estudantes de primeiro e segundo ano, sem desconto na mensalidade e com média de desempenho acadêmico inferior a 45.6 no último semestre, tendem a evadir. Além disso, com o intuito de prover predições precoces, o autor desenvolveu um modelo baseado no método de análise de discriminantes, considerando dados relacionados a diferentes tipos de ausências dos estudantes. A função resultante, influenciada de forma mais expressiva por ausências relacionadas a licenças médicas, foi incorporada a um formulário, permitindo que orientadores acadêmicos informem os quantitativos atualizados de ausências de um aluno e obtenham, em tempo real, seu risco de evasão.

3.3.3 Dados

Sob a perspectiva dos dados utilizados, relacionada à terceira questão de pesquisa desta RSL, é apresentada, na Figura 11(a), a distribuição dos artigos selecionados em relação aos tipos (natureza) de atributos e aos níveis de ensino considerados. Faz-se importante mencionar que, nessa categorização, informações relacionadas às ausências dos alunos foram classificadas como acadêmicas, embora alguns trabalhos referenciem esses dados como comportamentais. Além disso, atributos utilizados complementarmente por alguns estudos e extraídos a partir de questionários/pesquisas,

como o tempo diário destinado a lazer e lições de casa, não foram considerados na categorização. Isso porque, além de não ser uma prática comum dentre os trabalhos, dados resultantes de questionários estão sujeitos a limitações e enviesamentos, como mencionado no Capítulo 1.

Figura 11 – Quantitativo de artigos por nível de ensino e tipo de atributo (a) ou tamanho de conjunto de dados (b)



Fonte: Colpo et al. (2024).

Considerando que foram ocultados os rótulos relacionados a um único artigo, para não poluir os gráficos, note que o nível de Graduação é o único a possuir trabalhos vinculados a todos os tipos de dados, o que é compreensível por ser o nível mais pesquisado. Além disso, pode-se observar que dados acadêmicos, sociais e econômicos são amplamente utilizados no contexto de predição de evasão institucional ou de curso, ao contrário de dados interacionais, temporais e de análise de sentimentos. Embora não conste na Figura 11(a), cabe especificar alguns atributos apontados pelos estudos como de maior importância preditiva. Dentre as informações acadêmicas, pode-se destacar as de desempenho acadêmico, como nota em exame de admissão, médias de notas e frequências (semestrais ou acumuladas); e de posicionamento do aluno no curso ou semestre, como o número de semestres cursados, e o total de créditos integralizados ou perdidos (semestrais ou acumulados). Já entre os dados demográficos, destacam-se os atributos de idade de ingresso, sexo, tempo entre o término do ensino médio e o ingresso no ensino superior, e os relacionados à origem dos estudantes, como cidade natal e indicadores de seu contexto socioeconômico. Por fim, dentre as informações econômicas, pode-se citar atributos relacionados à renda, e ao recebimento de auxílios ou benefícios estudantis.

Enquanto Naseem; Chaudhary; Sharma (2022), Oreshin et al. (2020), Ortigosa et al. (2019), Queiroga et al. (2020), Deho et al. (2022), Park; Yoo (2021) e Villegas-Ch; Palacios-Pacheco; Lujan-Mora (2020) correspondem aos artigos vinculados ao tipo interacional, o trabalho de Oreshin et al. (2020) foi o único a utilizar atributos preditivos resultantes de análise de sentimento, representando, porém, os níveis de graduação e pós-graduação. Já Yoo; Woo; Park (2017) e Kurniawati; Maulidevi (2022)

consideraram em suas análises a temporalidade associada ao histórico acadêmico dos estudantes[15].

Utilizando dados acadêmicos, demográficos e econômicos de estudantes de graduação de uma universidade indonésia, e considerando a temporalidade/sequencialidade semestral dos dados de percurso acadêmico, Kurniawati; Maulidevi (2022) desenvolveram modelos de redes neuronais recorrentes, considerando os algoritmos *Long-Short Term Memory* (LSTM) e *Gate Recurrent Units* (GRU), para prever a conclusão/evasão e o desempenho final (abaixo da média, média, bom, muito bom) dos estudantes. Como diferencial, os autores construíram três modelos: um para cada tarefa preditiva (evasão e desempenho), separadamente; e outro para predizer os dados com rótulos cruzados/combinados de evasão e desempenho. Avaliando o desempenho preditivo dos modelos ao longo de diferentes semestres, Kurniawati; Maulidevi (2022) verificaram que os modelos GRU e LSTM tiveram um desempenho superior na previsão de graduação/evasão e de desempenho, respectivamente. Além disso, os autores destacaram que os modelos construídos separadamente demonstraram melhores resultados, com desempenhos satisfatórios a partir do primeiro e do segundo semestre nas predições de graduação/evasão e desempenho, respectivamente.

Ainda considerando características dos dados, na Figura 11(b) é apresentada a relação entre os artigos selecionados e os tamanhos dos conjuntos de dados (quantidade de estudantes/registros evadidos e diplomados) utilizados, considerando cada nível de ensino separadamente. Embora, em geral, os trabalhos se baseiem em conjuntos de dados de tamanhos bastante variados, é notório que os estudos relacionados ao ensino escolar se concentram majoritariamente na faixa de maior abrangência dos dados. Isso é compreensível à medida que, como visto na Seção 3.3.2, essas pesquisas costumam utilizar dados de diversas escolas da rede pública de um estado ou país. Semelhantemente, os demais trabalhos dessa faixa, como da Silva et al. (2019), Yu; Lee; Kizilcec (2021), e Oreshin et al. (2020)[16], costumam utilizar dados de estudantes de graduação de todos os cursos ou de diversos departamentos de uma instituição.

Em Beaulac; Rosenthal (2019) e Viloria et al. (2019), porém, foram contemplados apenas cursos da Faculdade de Artes e Ciências da Universidade de Toronto, e dos Departamentos de Engenharia da Universidade de Mumbai, respectivamente. Mais detalhadamente, Beaulac; Rosenthal (2019) utilizaram dados de desempenho acadêmico no primeiro ano de curso para construir, a partir dos algoritmos RF e LR

[15]Embora outros estudos tenham utilizado dados de caráter temporal, como o semestre em que o aluno está matriculado ou o intervalo entre o término do ensino médio e o ingresso no ensino superior, eles não extraíram informações baseadas na sequencialidade desses dados e nem consideraram suas séries temporais.

[16]Note que, no gráfico, o estudo de (Oreshin et al., 2020) é contabilizado tanto no nível de graduação quanto no de pós-graduação.

(*baseline*), modelos destinados a prever se estudantes obteriam ou não seus diplomas. Como resultados, os autores confirmaram o melhor desempenho do modelo RF e apontaram a grande importância preditiva de atributos associados a créditos e notas de uma disciplina de seminário e de disciplinas relacionadas a departamentos específicos, especialmente o de matemática. Já Viloria et al. (2019) utilizaram dados acadêmicos, demográficos e econômicos na construção de modelos preditivos de evasão, baseados nos algoritmos *Bayesian Networks*, ANN e DT. Além de considerarem os resultados de todos os modelos satisfatórios, os autores confirmaram, por meio da interpretação dos modelos e de seus *rankings* de atributos, a hipótese de que condições acadêmicas, como a pontuação no exame de admissão, e socioeconômicas, como benefícios estudantis, são determinantes para a permanência ou a evasão no contexto analisado.

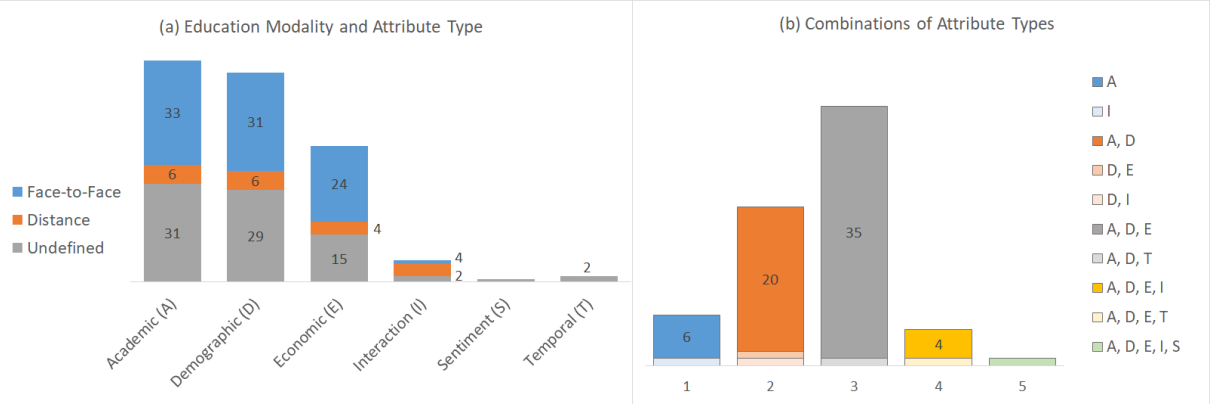
Por outro lado, os trabalhos que consideram conjuntos de dados de tamanho mais restrito (faixa de 500 a 1000 alunos/registros) abordam, em geral, a previsão de evasão em apenas um curso, como o caso de Costa et al. (2021), Hannaford; Cheng; Kunes-Connell (2021), Naseem; Chaudhary; Sharma (2022), Perez; Castellanos; Correal (2018), Queiroga et al. (2020), e Yoo; Woo; Park (2017). Porém, existem exceções, como os trabalhos de Iam-On; Boongoen (2017a,b), nos quais a base de dados, embora pequena, abrange dados acadêmicos e sociais de estudantes de 26 departamentos acadêmicos da *Mae Fah Luang University*. Assim como no primeiro trabalho, descrito na Seção 3.3.1, Iam-On; Boongoen (2017b) consideraram, separadamente, os contextos de estudantes recém admitidos e que concluíram o primeiro ano de curso. Porém, com o intuito de desenvolver modelos preditivos de evasão e, assim, validar uma abordagem de transformação de dados, baseada em *Link-Based Cluster Ensemble*. Como resultado, os autores demonstraram que sua abordagem, apesar de exigir maior esforço computacional, melhorou as predições e superou outras técnicas de redução de dimensionalidade, considerando modelos desenvolvidos para ambos os contextos, a partir de algoritmos DT, KNN, NB e ANN.

A relação entre os artigos e a natureza dos dados utilizados, considerando agora a perspectiva da modalidade de ensino, é apresentada na Figura 12(a)[17]. Observe que, para o ensino presencial, a adoção dos tipos de dados segue, em geral, a mesma tendência da totalidade dos estudos selecionados. Porém, na EAD, os dados interacionais demonstram a mesma representatividade que os econômicos, considerados por 57% (4 de 7) dos trabalhos dessa modalidade, confirmando a maior exploração da interação dos alunos com os AVAs nesse contexto. Além disso, na Figura 12(b), é possível visualizar o mapeamento acerca da quantidade de tipos de dados considerada nos artigos. Percebe-se que a maioria dos estudos utiliza três tipos de dados,

[17]Observe que, em ambos os gráficos da Figura 12, foram ocultados os rótulos das partições relacionadas a um único artigo.

seguida pelos que adotam dois e um tipo, respectivamente.

Figura 12 – Artigos por tipos de dados e modalidades (a); e por combinações de tipos (b)



Fonte: Colpo et al. (2024).

O uso de quatro ou cinco tipos foi identificado apenas nos estudos de Ortigosa et al. (2019), Naseem; Chaudhary; Sharma (2022), Park; Yoo (2021), Villegas-Ch; Palacios-Pacheco; Lujan-Mora (2020), Kurniawati; Maulidevi (2022), e Oreshin et al. (2020). A maioria destes estudos considera dados acadêmicos, demográficos/sociais, econômicos e interacionais (A, D, E, I), com exceção de Kurniawati; Maulidevi (2022), que considera dados temporais e não interacionais (A, D, E, T), e Oreshin et al. (2020), que acrescenta dados de análise de sentimento aos primeiros quatro tipos (A, D, E, I, S). Os trabalhos ligados a três tipos utilizam geralmente dados acadêmicos, demográficos/sociais e econômicos (A, D, E), como em Barros et al. (2019), Costa et al. (2021), Gamao; Gerardo (2019) e Solis et al. (2018). Como exceção, Yoo; Woo; Park (2017) utilizou dados acadêmicos, demográficos e temporais (A, D, T).

Para os trabalhos relacionados a dois tipos, em geral, a combinação de dados acadêmicos e demográficos (A, D) foi utilizada, como em Berka; Marek (2021), Iam-On; Boongoen (2017a,b), e Perchinunno; Bilancia; Vitale (2019). As exceções correspondem a Deho et al. (2022) e Freitas et al. (2020), que adotaram dados demográficos e interacionais (D, I) ou econômicos (D, E), respectivamente. Freitas et al. (2020) desenvolveram um sistema para a predição precoce de evasão, considerando modelos baseados em apenas sete atributos socioeconômicos de estudantes de engenharia do Instituto Federal do Ceará, Campus Fortaleza, e nos algoritmos LR, DT, SVM, KNN, MLP e *Deep Neural Network* (DNN). Como resultado, além de disponibilizarem uma interface para os usuários poderem, dentre outras funcionalidades, selecionar o modelo a ser utilizado na predição, submeter registros de alunos a serem previstos, ou consultar o risco de estudantes com dados já armazenados, os autores apontaram melhores desempenhos para os modelos DT e DNN.

Por fim, quando apenas um tipo foi utilizado, geralmente ele foi acadêmico (A), caso de Beaulac; Rosenthal (2019), Chung; Lee (2019), Lee; Chung (2019), Rovira;

Puertas; Igual (2017), Assis et al. (2022), e Kuzilek; Zdrahal; Fuglik (2021). A única exceção foi o estudo de Queiroga et al. (2020), que considerou apenas dados interacionais (I), como descrito na Subseção 3.3.1.

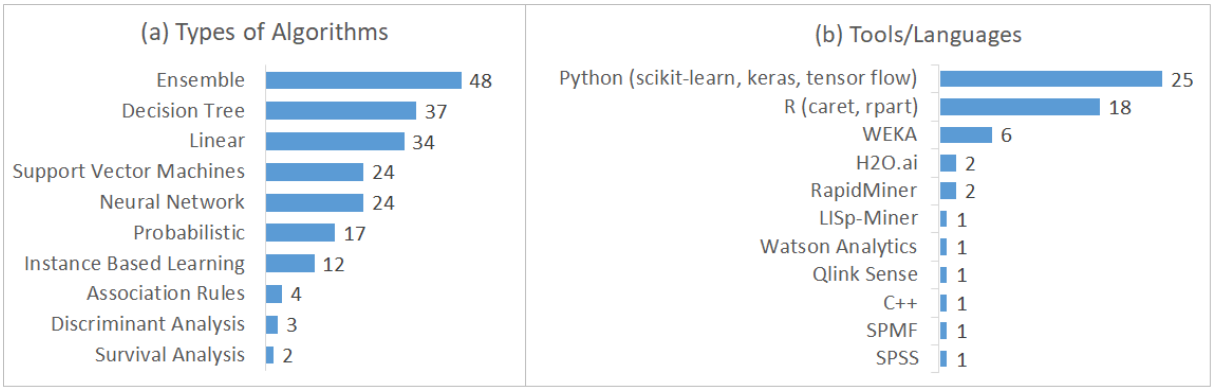
3.3.4 Ferramentas e Técnicas

Respondendo ao primeiro item da terceira questão de pesquisa, os artigos selecionados na RSL abordaram a previsão de evasão, majoritariamente, por meio da tarefa de Classificação, seja para identificar alunos em risco, seja para verificar atributos e padrões relacionados ao abandono. As excepções foram os estudos de Iam-On; Boongoen (2017a), Yoo; Woo; Park (2017), da Silva et al. (2019), Assis et al. (2022) e Shilbayeh; Abonamah (2021), que utilizaram apenas técnicas de Clusterização, Associação ou Regressão.

Os algoritmos de classificação utilizados pelos trabalhos foram categorizados em relação aos métodos/técnicas representadas. Na Figura 13(a) são apresentadas as frequências dessas categorias. Pode-se observar que métodos de *ensemble* foram adotados na maioria dos estudos, mais especificamente em 48. Isso é compreensível, à medida que algoritmos de *ensemble* têm ganhado atenção nos últimos anos, por combinarem diversos classificadores e aumentarem consideravelmente a qualidade das predições, devido à maior capacidade de generalização desse conjunto de modelos (Lee; Chung, 2019). Embora não seja apresentado no gráfico, cabe mencionar que a maioria dessas ocorrências (43) considerava DTs como classificadores base, com destaque à alta frequência do algoritmo RF, utilizado, por exemplo, em Chung; Lee (2019), Solis et al. (2018) e Queiroga et al. (2022). Também é possível visualizar que DTs não são aplicadas apenas como classificadores base. Suas aplicações na forma individual, a exemplo de Freitas et al. (2020), Perez; Castellanos; Correal (2018), e Ortigosa et al. (2019), correspondem à segunda categoria de algoritmo mais recorrente. A explicação pode estar associada não apenas a seus bons desempenhos preditivo e computacional, mas também à interpretabilidade de seus resultados, já que as DTs são constituídas por regras que descrevem os padrões descobertos (Han; Kamber; Pei, 2012). Motivação semelhante pode justificar o fato de modelos lineares, utilizados em Kang; Wang (2018), Mduma; Machuve (2021), Perchinunno; Bilancia; Vitale (2019) e Shiau (2020), por exemplo, aparecerem como a terceira categoria mais frequente, com destaque aos modelos de LR.

Ainda em relação à Figura 13(a), percebe-se que categorias relacionadas a algoritmos de SVM, redes neurais (simples ou profundas), e probabilísticos (baseados no teorema de Bayes) também apresentaram frequências significativas. Já a categoria de aprendizado baseado em instâncias, majoritariamente representada pelo algoritmo KNN nos estudos selecionados, apresentou recorrência mais discreta. Freitas et al. (2020), Hannaford; Cheng; Kunes-Connell (2021), e Palacios et al. (2021) são exem-

Figura 13 – Artigos por categorias de algoritmos (a) e ferramentas (b) utilizadas



Fonte: Colpo et al. (2024).

plos de trabalhos que utilizaram o KNN, dentre outros algoritmos, em suas análises comparativas. Embora também seja um algoritmo tradicional, essa menor adoção do KNN pode estar associada ao custo computacional que ele exige na classificação de novas instâncias, que pode aumentar linearmente, no pior caso, com o tamanho do conjunto de treino (Raschka; Mirjalili, 2017). Por fim, as demais categorias foram pouco utilizadas, tendo sido citadas em menos de cinco estudos.

Santos et al. (2020) consideraram dados acadêmicos e demográficos de alunos de cursos de graduação presenciais de uma instituição federal brasileira, no desenvolvimento de um modelo preditivo de evasão, denominado *EvolveDTree*. Com o intuito de melhorar o aprendizado do classificador DT em meio ao conjunto de dados desbalanceado, os autores empregaram o algoritmo de clusterização *K-Means* no pré-processamento, para estratificar os dados em 10 grupos de registros e utilizá-los no treinamento por validação cruzada. A fim de evitar o *overfitting* do modelo DT, também foi empregado um método de seleção de atributos baseado em um algoritmo evolucionário (genético), que selecionou oito atributos: indicador de participação em programas sociais, semestre de encerramento do curso, nota de redação no exame de admissão, ano e semestre de ingresso, GPA, raça/etnia/cor, e número de anos no curso. Como resultados, além de validar os resultados de cada etapa de sua abordagem, os autores demonstraram a superioridade de desempenho de seu modelo perante modelos de KNN, AdaBoost, MLP, RF, NB, SVM, e *Quadratic Discriminant Analysis*. Por fim, a visualização do modelo também apontou que o EvolveDTree relacionou à evasão, principalmente, estudantes com menor desempenho acadêmico ou que ainda não passaram do segundo ano de curso.

Em Chen; Johri; Rangwala (2018) foi proposto o uso de análise de sobrevivência, baseada nos modelos de *Aalen's Additive* e *Cox's Proportional Hazard*, para prever não apenas a ocorrência mas também o momento da evasão de estudantes de graduação das áreas STEM (do inglês, *Science, Technology, Engineering, and Mathema-*

tics), geralmente associadas a maiores taxas de abandono. Foram treinados modelos a partir de dados demográficos e acadêmicos (escolares e universitários) de estudantes da *George Mason University* (USA), considerando informações disponíveis em diferentes cortes temporais, incrementalmente (partindo de dados disponíveis no fim do segundo semestre até dados disponíveis ao término do quinto semestre). Os modelos de análise de sobrevivência foram comparados a modelos tradicionais de aprendizado de máquina (LR, DT, RF, NB e AdaBoost), e, como resultado, além de observarem que os resultados preditivos melhoravam à medida que novos dados semestrais eram considerados, os autores identificaram que os modelos de análise de sobrevivência apresentaram melhores desempenhos quando poucos dados semestrais estavam disponíveis. Além disso, os atributos de médias de notas semestrais (GPAs) e número de semestres cursados/matriculados demonstraram maior capacidade preditiva.

Na Figura 13(b) são apresentadas as ferramentas e linguagens de programação utilizadas pelos artigos selecionados, no processo de EDM. Embora alguns trabalhos não mencionem as ferramentas adotadas, analisando o gráfico, é possível verificar uma ampla utilização das linguagens Python e R. Faz-se importante mencionar que o uso de Python aparece, em geral, associado à biblioteca de aprendizado de máquina "scikit-learn", a exemplo de Barros et al. (2019), Queiroga et al. (2020), e Rovira; Puertas; Igual (2017). No trabalho de Rovira; Puertas; Igual (2017), ainda não apresentado, foram desenvolvidos modelos para prever, dentre outros alvos, o risco de estudantes de graduação da Universidade de Barcelona não voltarem a se matricular no segundo e no terceiro ano de curso, considerando dados do desempenho acadêmico dos alunos no primeiro ano dos cursos de Direito, Ciência da Computação, e Matemática, separadamente. Além de aplicarem a técnica de balanceamento SMOTE, os autores criaram cinco modelos preditivos para cada curso, por meio dos algoritmos LR, SVM, RF, NB e AdaBoost. Como resultado, os modelos RF e AdaBoost demonstraram os melhores desempenhos para o curso de Direito, enquanto os modelos LR e NB se destacaram nos outros dois cursos. Considerando os tamanhos dos conjuntos de dados utilizados para cada curso, os autores apontaram que classificadores não paramétricos, como LR e NB, podem ser mais adequados ao contexto de pouca disponibilidade de dados, enquanto modelos paramétricos, como RF e AdaBoost, podem se ajustar melhor à situação oposta.

Similarmente, o pacote "caret" é frequentemente utilizado nos trabalhos desenvolvidos em R, como em Perchinunno; Bilancia; Vitale (2019), Lee; Chung (2019), e Solis et al. (2018). Entre as outras ferramentas, apenas WEKA, RapidMiner e H2O.ai foram referenciadas mais do que uma vez pelos estudos selecionados. Palacios et al. (2021), Viloria et al. (2019), e Vega et al. (2022) são exemplos de pesquisas que utilizaram a ferramenta WEKA. As duas únicas referências ao RapidMiner ocorreram em Berka; Marek (2021) e Nagy; Molontay (2018). Embora Berka; Marek (2021) tenham utilizado

o RapidMiner na classificação, os autores também aplicaram a ferramenta LISp-Miner na mineração de associações. Semelhantemente, Nagy; Molontay (2018) também utilizaram implementações da plataforma H2O.ai, também adotada por Deho et al. (2022), em alguns dos seus modelos.

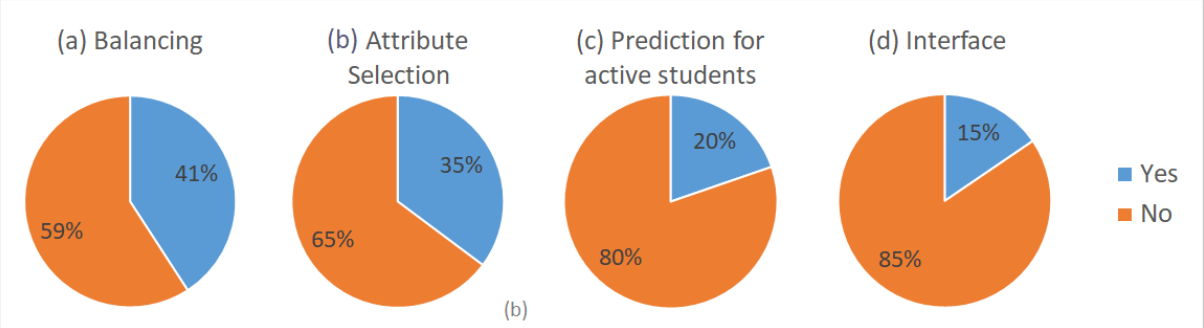
Com base em dados acadêmicos e socioeconômicos de estudantes de graduação da Faculdade de Engenharia de Sistemas e Informática da *Universidad Nacional Mayor de San Marcos* (Peru), Vega et al. (2022) desenvolveram um modelo de DT e um sistema (não web) para prever a evasão estudantil. Os autores utilizaram seis métodos disponíveis na ferramenta WEKA para selecionar os atributos frequentemente considerados importantes, e aplicaram a sobreamostragem SMOTE para lidar com o desbalanceamento dos dados. Como resultado, Vega et al. (2022) apresentaram um diagrama de casos de uso e protótipos de interface relacionados ao sistema desenvolvido, além de destacar a importância dos atributos de médias ponderadas de notas acumuladas e do último ciclo, e número de créditos integralizados.

Em Nagy; Molontay (2018) foram desenvolvidos modelos de predição precoce de evasão, baseados em dados acadêmicos e demográficos, disponíveis na admissão de estudantes de graduação da *Budapest University of Technology and Economics*. Além de utilizarem diferentes estratégias de seleção de atributos (evolucionária, por correlação, e baseada em métricas dependentes de modelos GBT e DNN), os autores consideraram diversos classificadores, mais especificamente os de regressão linear, DT, RF, GBT, LR, NB, KNN, DNN, e AdaBoost. Dentre os atributos selecionados como importantes, os autores identificaram que o desempenho em disciplinas de humanas no ensino médio foi considerado relevante, mesmo para o contexto de estudantes de graduação de engenharias e exatas, sugerindo a importância de uma boa e ampla formação para o sucesso universitário. Além disso, os modelos GBT e DNN demonstraram melhores desempenhos sobre os conjuntos selecionado e completo de atributos, respectivamente. Como resultado, foi desenvolvida e disponibilizada uma aplicação web, considerando o seleto conjunto de atributos, para os usuários poderem informar dados de um estudante e obter a previsão de seu risco de evasão, calculado pelo modelo GBT.

Além das características técnicas já apresentadas, faz-se importante posicionar os estudos em relação ao atendimento/satisfação de alguns aspectos adicionais, conforme apresentado na Figura 14. As proporções dos trabalhos perante a sinalização de adoção de técnicas de balanceamento e seleção de atributos são mostradas nas Figuras 14(a) e 14(b), respectivamente. Essas análises são necessárias, pois dados desbalanceados tendem a ser uma realidade no contexto da evasão, podendo prejudicar a aprendizagem da classe sub-representada. Além disso, técnicas de seleção de atributos visam remover dimensões irrelevantes ou redundantes dos dados, reduzindo o custo de processamento e evitando que essas variáveis influenciem negativamente

na geração dos modelos de classificação (Han; Kamber; Pei, 2012).

Figura 14 – Proporções de artigos em relação a diferentes características técnicas



Fonte: Colpo et al. (2024).

Pode-se observar que apenas 41% (29) dos trabalhos citou a aplicação ou a avaliação de técnicas de balanceamento. Embora nem todas as bases de dados sejam naturalmente desbalanceadas, esse baixo percentual revela que muitas pesquisas não atentam para essa importante questão. Porém, dentre os estudos que realizaram o balanceamento dos dados, destaca-se o uso da técnica SMOTE, a exemplo de Hutagaol; Suharjito (2019), Lee; Chung (2019), e Rovira; Puertas; Igual (2017). Já as técnicas de seleção de atributos foram consideradas em 35% (25) dos estudos. Esse resultado é compreensível, à medida que muitos trabalhos já partem de um número reduzido de atributos, por falta de acesso a mais informações ou por basearem a escolha de seus atributos em experiências, ou pesquisas anteriores. Dentre as abordagens de seleção utilizadas nos trabalhos já descritos, foram citados métodos evolutivos, como em Gamao; Gerardo (2019) e Santos et al. (2020); de correlação e *stepwise*, a exemplo de Fernández-García et al. (2021) e da Silva et al. (2019); LVQ, em Hutagaol; Suharjito (2019); métricas de importância dependentes de modelos, em Nagy; Molontay (2018); e o algoritmo Boruta, que opera como um *wrapper* de RF, em Naseem; Chaudhary; Sharma (2022).

Outro exemplo de estudo que realizou seleção de atributos é Tsai et al. (2020), no qual foram utilizados dados demográficos, econômicos e de desempenho acadêmico do primeiro ano de curso, de estudantes de uma universidade de Taiwan, no desenvolvimento de modelos destinados a prever evasões nos três anos seguintes da graduação. Porém, para selecionar os atributos a serem utilizados no treinamento desses modelos, os autores realizaram análise estatística por regressão logística, que apontou quatro atributos principais: performance acadêmica (*class ranking percentage*), indicador de pedido de empréstimo estudantil, número de ausências, e número de disciplinas em que o estudante recebeu alertas de desempenho acadêmico. Considerando esses quatro atributos para cada um dos dois primeiros semestres de curso, foram treinados classificadores MLP e LR, que demonstraram melhores especifici-

dade e sensibilidade, respectivamente. Como resultado, o modelo mais sensível foi utilizado na tarefa preditiva, provendo listagens de alunos em risco à universidade e, assim, possibilitando a execução de estratégias preventivas.

Na Figura 14(c) é apresentada a proporção dos trabalhos que preveem automaticamente o risco de evasão para alunos matriculados (ativos). Percebe-se que, embora muitas pesquisas (80%) apliquem e validem modelos preditivos em dados históricos de evasão, a fim de identificar padrões relacionados a esse fenômeno ou definir o melhor algoritmo de predição, apenas 20% (14) dos trabalhos utilizam os modelos construídos em dados do presente, para prever e detectar automaticamente evasões futuras. Especificamente, Kang; Wang (2018), Nagy; Molontay (2018), Ortigosa et al. (2019), Mduma; Kalegele; Machuve (2019), Freitas et al. (2020), Shiau (2020), Tsai et al. (2020), Shilbayeh; Abonamah (2021), Xu; Wilson (2021), Bassetti et al. (2022), Queiroga et al. (2022), Vega et al. (2022), Demeter et al. (2022) e Urbina-Najera; Mendez-Ortega (2022) correspondem ao último caso.

Interessados na utilidade preditiva dos registos de auxílio financeiro, Demeter et al. (2022) utilizaram dados acadêmicos, demográficos e econômicos de estudantes universitários que se candidataram a auxílio federal gratuito nos dois primeiros anos consecutivos de matrícula em uma universidade pública do sudeste dos EUA. Além de selecionar atributos a partir de um processo orientado teoricamente, os autores utilizaram técnicas orientadas por dados, como a regressão logística *stepwise* e a importância dos atributos do modelo. Como diferencial, Demeter et al. (2022) conduziram um processo de classificação hierárquico de dois níveis para fazer predições a partir de registos disponíveis ao final do segundo ao sexto semestre. Para isso, foram utilizados dois modelos RF, um para prever inicialmente se os alunos iriam evadir ou concluir a graduação, e outro para ser aplicado aos potenciais concluintes, para prever se a conclusão seria tardia ou em tempo regular (4 anos). Para além da validação de desempenho preditivo, os autores aplicaram seus modelos a dados de estudantes atuais, obtendo uma lista com os potenciais desistentes, compartilhada para análise de orientadores acadêmicos. Por fim, Demeter et al. (2022) também destacaram a importância preditiva dos atributos relacionados a créditos integralizados, GPA da universidade e do ensino médio, contribuição financeira da família, e inscrições e notas em cursos obrigatórios da área de especialização do aluno.

Urbina-Najera; Mendez-Ortega (2022) utilizaram dados acadêmicos, demográficos e econômicos de estudantes universitários de 25 cursos de engenharia, ciências sociais e ciências administrativas de uma universidade não identificada. Mediante uma sobreamostragem da classe minoritária (evadidos) e de uma subamostragem da classe maioritária (não evadidos), os autores avaliaram diferentes métodos de seleção de atributos (*wrapper* de RF e baseados em correlação e em consistência) no desenvolvimento de modelos RF e ANN para a previsão de evasão. Entre os resulta-

dos, destacaram que a reamostragem beneficiou a capacidade preditiva dos modelos e que o RF com atributos selecionados pelo *wrapper* de RF apresentou melhor desempenho. Urbina-Najera; Mendez-Ortega (2022) também relataram que o (melhor) modelo de RF foi implementado no sistema institucional, permitindo a previsão de um número aproximado de estudantes em risco por período e contribuindo para as ações de prevenção.

Outro fator a ser observado é que nem todos os estudos que previram evasões futuras disponibilizaram suas previsões aos gestores acadêmicos por meio de interfaces, como indicado na Figura 14(d). Ou seja, embora 92% das pesquisas tenham desenvolvido modelos preditivos de evasão (Figura 9), apenas 15% garantiram o acesso intuitivo e continuo às previsões reais, facilitando o acompanhamento e o desenvolvimento de estratégias preventivas junto aos alunos em risco. Além disso, faz-se importante mencionar que, dentre os onze trabalhos que disponibilizaram sistemas/interfaces, apenas os de Ortigosa et al. (2019), Xu; Wilson (2021) e Bassetti et al. (2022) se relacionam a projetos que integram os sistemas de previsão com as bases de dados de sistemas administrativos/institucionais. Os demais exigem que o usuário informe, por meio do preenchimento de um formulário ou da submissão de um arquivo .csv, os dados/atributos dos estudantes a terem sua situação predita.

3.4 Tendências, desafios e oportunidades

A análise dos estudos abrangidos por esta RSL evidenciou algumas tendências, oportunidades e desafios relacionados à aplicação de EDM e aprendizado de máquina no contexto da previsão da evasão institucional e de curso.

Do ponto de vista contextual, pode-se destacar a carência e consequente oportunidade de investigação na educação básica/escolar e técnica, e nos países com menor IDH, precisamente os contextos com maior necessidade. Isso porque os níveis de ensino básico (fundamental e médio) abrangem o maior volume de alunos e são extremamente importantes para o desenvolvimento socioeconômico e educacional da população. Também é preciso considerar os desafios que podem estar por trás dessas carências, que precisam ser enfrentados para impulsionar avanços nesse contexto de pesquisa. Embora concentrem a maioria dos alunos, as escolas públicas possuem recursos limitados e podem não ter equipes dedicadas e especializadas na área de tecnologia da informação, dificultando a realização desses estudos localmente (nas escolas). Portanto, expandir o uso de EDM e aprendizado de máquina no contexto da evasão escolar depende do investimento de órgãos governamentais no desenvolvimento dessas pesquisas internamente ou em facilitá-las externamente, por meio da disponibilização de conjuntos de dados públicos, que, preferencialmente,

abranjam diferentes aspectos dos alunos[18]. Mduma; Machuve (2021) também apontou a dificuldade de se encontrar conjuntos de dados que abordem o problema da evasão estudantil e que sejam disponibilizados publicamente, para o desenvolvimento de estudos envolvendo a aplicação de técnicas de aprendizado de máquina. Ainda em consonância com as observações sobre o contexto escolar, nesta RSL dois dos três estudos relacionados ao ensino técnico foram realizados no Brasil, utilizando dados de instituições da Rede Federal de Educação Profissional, Científica e Tecnológica, que, por também ofertarem cursos de graduação e pós-graduação, contam com um corpo técnico especializado, gerando um ambiente mais propício à pesquisa.

Em relação aos dados utilizados, muitos estudos utilizam combinações de dados demográficos, econômicos e acadêmicos, apesar da ampla adoção de AVAs nas universidades. Isso revela uma oportunidade de pesquisa, já que a análise de dados da interação dos estudantes em AVAs pode auxiliar na identificação de padrões comportamentais. Além disso, dados interacionais podem ser especialmente valiosos em modelos de previsão precoce, por estarem acessíveis ao longo de todo o semestre, mesmo antes de o estudante ter concluído qualquer período acadêmico ou obtido informações sobre notas (Park; Yoo, 2021). Outra oportunidade relacionada a dados interacionais é a utilização de textos de postagens e de participações de alunos em fóruns do AVA em um processo de mineração de sentimentos, cujo resultado também pode contribuir para a previsão da evasão. Semelhantemente, pode-se explorar melhor a temporalidade e a sequencialidade dos dados relativos à trajetória acadêmica. Estudos comparando os resultados preditivos de modelos de classificação tradicionais com os de séries temporais seriam de grande utilidade prática e esclareceriam se a temporalidade dos dados realmente oferece vantagem preditiva no contexto da evasão.

Em geral, os estudos relacionados à previsão de evasão também demonstram pouca adaptação das técnicas ao domínio específico. Ou seja, apesar de considerarem questões de pesquisa focadas no problema da evasão estudantil, os trabalhos costumam adotar técnicas tradicionais de EDM e aprendizado de máquina, sem adaptá-las ou propor diferentes formas de aplicação às particularidades do domínio. Fernández-García et al. (2021) e Demeter et al. (2022) são exemplos de estudos que propõem estruturas de classificação diferentes e que podem servir de inspiração para outras investigações nessa direção. Também é comum o desenvolvimento de modelos preditivos especializados em cursos ou períodos letivos. Embora a exploração de fatores mais específicos, como o desempenho dos alunos em determinadas disciplinas, possa beneficiar tais modelos, é importante notar que esta abordagem fornece

[18] Isso porque muitos dados públicos, como os de censos educacionais, restringem-se a informações socioeconômicas, limitando o potencial de pesquisa e, consequentemente, o interesse dos pesquisadores

resultados menos abrangentes institucionalmente. Assim, sempre que possível, o desenvolvimento de modelos focados na previsão genérica da evasão (ou seja, modelos destinados a prever a evasão em diferentes cursos e períodos da trajetória acadêmica) pode prover maior retorno à prática educacional.

Ainda considerando a perspectiva das técnicas utilizadas, embora se tenha observado entre os trabalhos a preocupação em avaliar diferentes algoritmos de aprendizado de máquina e isso realmente seja importante, é preciso destacar, para o desenvolvimento de novos estudos, a necessidade de se validar as decisões técnicas ao longo de todo o processo de desenvolvimento dos modelos preditivos, considerando a realidade institucional e os objetivos de cada pesquisa. No desenvolvimento de um modelo de predição genérico, destinado a prever evasões em diferentes cursos e estágios da trajetória acadêmica, por exemplo, considerar atributos absolutos (como o número de créditos integralizados) é a melhor opção? O uso de atributos relativos (como a porcentagem de créditos integralizados) não forneceria dados mais informativos e generalizáveis, perante estruturas curriculares distintas? Além disso, estudar a escolha e a priorização das métricas avaliativas, para cada realidade, é extremamente necessário. Caso contrário, todas as decisões técnicas podem se basear em resultados tendenciosos, gerando modelos falsamente viáveis. Embora diversos estudos apontem o bom desempenho de seus modelos, muitas vezes essas conclusões não apresentam o devido embasamento, já que não se deve avaliar modelos de classificação apenas por suas taxas gerais de acertos (acurácias), por exemplo, principalmente em um contexto de dados desbalanceados. Além disso, embora os modelos baseados em DT tenham sido amplamente utilizados por combinarem bons desempenhos de predição e execução com interpretabilidade, é preciso salientar que existem diversas possibilidades para a aplicação de modelos mais complexos/caixa preta. Isso porque, se se esses modelos proporcionarem uma maior vantagem preditiva e tempos de execução aceitáveis para a realidade do projeto/pesquisa, podem ser aplicadas técnicas de XAI na identificação de fatores associados à previsão de evasão (Rondado de Sousa et al., 2021).

Embora haja um interesse crescente na utilização da EDM para prever a evasão estudantil, a maioria dos estudos analisa apenas dados históricos, sem aplicar os modelos desenvolvidos a dados de alunos ativos/matriculados, para prever evasões futuras. Além disso, nem todos os poucos trabalhos que efetuam previsões fornecem aos gestores acadêmicos interfaces/ferramentas para acesso intuitivo e contínuo aos resultados das previsões. Acreditamos que esse subaproveitamento dos esforços e do potencial de muitos estudos na prática educacional é o maior desafio a ser enfrentado. Isso porque, como já mencionado, a maioria das pesquisas é realizada em ambiente acadêmico, muitas vezes associada a trabalhos de conclusão de curso, dissertações ou teses. Nesse tipo de pesquisa, devido a questões de privacidade, o acesso aos

dados costuma ser completamente anônimo e parcial. Como resultado, estes estudos não permitem avançar efetivamente na identificação de alunos em risco. A solução para este problema pode estar na realização de projetos que envolvam também pesquisadores e *stakeholders* ligados às equipes de tecnologias de informação e gestores das instituições detentoras dos dados. Desta forma, os modelos desenvolvidos no âmbito acadêmico podem ser integrados aos sistemas institucionais, em um esforço da instituição parceira.

Por fim, faz-se necessário apontar a existência de limitações e ameaças à validade desta RSL. Primeiramente, devido ao grande número de trabalhos considerados, não foi realizada revisão por pares na seleção e na análise dos artigos. Ou seja, cada publicação foi analisada por apenas um pesquisador. No contexto interpretativo inerente a qualquer estudo secundário, isso pode gerar decisões subjetivas. Outra ameaça pode ser a falta de bases de dados científicas importantes para os contextos específicos dos diferentes países. No entanto, como esta RSL foca no cenário internacional e não se tem conhecimento sobre as bases relevantes para cada país, a inclusão de algumas bases de dados específicas poderia introduzir enviesamento nos resultados. Por esta razão, foram consideradas apenas quatro grandes bases de dados científicas internacionais, especificamente ACM Digital Library, IEEE Xplore, Web of Science e Scopus.

3.5 Considerações sobre o Capítulo

Neste Capítulo foi apresentada uma RSL acerca da aplicação de EDM no contexto da previsão de evasão estudantil. Essa revisão abrangeu os escopos de evasão institucional e de curso e resultou na identificação de características contextuais, técnicas e de dados que têm sido exploradas nesse tópico de pesquisa. As Tabelas 4 e 5 posicionam os estudos analisados e o trabalho desenvolvido nesta Tese perante as principais características relacionadas às questões de pesquisa da RSL.

Na RSL, observou-se um maior foco de pesquisa no nível de graduação, na rede pública de ensino, e na modalidade presencial, podendo esta última tendência estar relacionada ao escopo analisado, que desconsiderou publicações relacionadas à predição de evasão em disciplinas, recorrentes na modalidade EAD. Além disso, embora, em geral, os trabalhos tenham buscado o desenvolvimento de modelos preditivos, a maioria também demonstrou interesse em obter indícios a respeito dos principais fatores relacionados à predição e, consequentemente, à evasão estudantil.

Tabela 4 – Posicionamento do trabalho proposto nesta Tese, perante os estudos e as principais características analisadas na RSL (Parte 1)

	Contexto								Objetivo		Tipo de Dados						Tarefa EDM				Técnica			
	Educação Básica/Escolar	Técnico	Graduação	Pós-Graduação	EAD	Presencial	Pública	Privada	Atributos x Evasão	Modelo Preditivo	Acadêmico	Demográfico	Econômico	Interacional	Sentimento	Temporal	Associação	Classificação	Clusterização	Regressão	Balanceamento	Seleção de Atributos	Previsão (matriculados)	Interface/Sistema
Agrusti et al. (2020)			✓				✓			✓	✓	✓	✓					✓						
Aguirre; Pérez (2020)			✓						✓	✓	✓	✓	✓							✓	✓	✓		
Alturki et al. (2022)				✓		✓	✓		✓	✓	✓	✓						✓			✓			
Assis et al. (2022)			✓				✓		✓		✓						✓							
Baranyi et al. (2020)			✓				✓		✓	✓	✓	✓						✓						
Barros et al. (2019)	✓	✓				✓	✓			✓	✓	✓	✓					✓			✓			
Bassetti et al. (2022)			✓	✓			✓		✓	✓	✓	✓	✓					✓					✓	✓
Beaulac; Rosenthal (2019)			✓				✓		✓	✓	✓							✓						
Berka; Marek (2021)			✓		✓	✓			✓	✓	✓	✓					✓	✓						
Bitencourt et al. (2022)			✓			✓	✓		✓	✓	✓	✓	✓					✓			✓			
Böttcher et al. (2021)			✓						✓	✓	✓	✓						✓						
Chen et al. (2018)			✓				✓		✓	✓	✓	✓						✓						
Chung; Lee (2019)	✓					✓	✓	✓	✓	✓	✓							✓			✓			
Costa et al. (2021)			✓			✓	✓		✓	✓	✓	✓	✓					✓						
Crespo (2020)	✓					✓	✓	✓	✓	✓	✓	✓	✓					✓						
da Silva et al. (2019)			✓			✓	✓	✓	✓	✓	✓	✓	✓							✓		✓		
Deho et al. (2022)			✓				✓			✓		✓		✓				✓						
Del Bonifro et al. (2020)			✓	✓					✓	✓	✓	✓						✓			✓			
Demeter et al. (2022)			✓			✓	✓		✓	✓	✓	✓	✓					✓				✓	✓	
Fernández–García et al. (2021)			✓	✓			✓			✓	✓	✓	✓					✓			✓	✓		
Flores et al. (2022)			✓			✓	✓		✓	✓	✓	✓						✓			✓	✓		
Fontana et al. (2021)			✓				✓			✓	✓	✓						✓						
Freitas et al. (2020)			✓			✓	✓			✓		✓	✓					✓					✓	✓
Gamao; Gerardo (2019)			✓				✓			✓	✓	✓	✓					✓				✓		
Hannaford et al. (2021)			✓			✓		✓	✓	✓	✓	✓						✓						
Hoffait; Schyns (2017)			✓			✓	✓		✓	✓	✓	✓	✓					✓						
Hutagaol; Suharjito (2019)			✓			✓		✓	✓	✓	✓	✓	✓					✓			✓	✓		
Iam-On; Boongoen (2017a)			✓				✓			✓	✓	✓						✓	✓			✓		
Iam-On; Boongoen (2017b)			✓				✓		✓		✓	✓							✓					
Kang; Wang (2018)			✓		✓		✓		✓	✓	✓	✓						✓			✓		✓	
Karimi-Haghighi et al. (2022)			✓				✓		✓	✓	✓	✓						✓						
Kiss et al. (2019)			✓			✓	✓		✓	✓	✓	✓						✓						
Kurniawati; Maulidevi (2022)			✓							✓	✓	✓	✓			✓		✓						
Kuzilek et al. (2021)			✓			✓	✓		✓	✓	✓							✓						
Lee; Chung (2019)	✓					✓	✓	✓		✓	✓							✓			✓			
Lottering et al (2020)			✓			✓	✓			✓	✓	✓	✓					✓			✓	✓		
Masood; Begum (2022)	✓					✓				✓	✓	✓						✓			✓			
Mduma et al. (2019)	✓					✓			✓	✓	✓	✓	✓					✓			✓	✓	✓	✓
Mduma; Machuve (2021)	✓					✓			✓	✓	✓	✓	✓					✓			✓	✓		
Nagy; Molontay (2018)			✓				✓		✓	✓	✓	✓	✓					✓				✓	✓	✓
Naseem et al. (2022)			✓			✓			✓	✓	✓	✓	✓	✓				✓			✓	✓		
Nuanmeesri et al. (2022)			✓			✓	✓		✓	✓	✓	✓	✓					✓				✓		
Opazo et al. (2021)			✓				✓	✓	✓	✓	✓	✓	✓					✓			✓	✓		
Oreshin et al. (2020)			✓	✓			✓		✓	✓	✓	✓	✓	✓	✓			✓						
Orooji; Chen (2019)	✓					✓	✓			✓	✓	✓	✓					✓			✓			
Ortigosa et al. (2019)			✓		✓			✓	✓	✓	✓	✓	✓	✓				✓			✓		✓	✓
Este trabalho			✓			✓	✓		✓	✓	✓	✓	✓	✓				✓			✓	✓		

Fonte: Elaborada pela autora.

Tabela 5 – Posicionamento do trabalho proposto nesta Tese, perante os estudos e as principais características analisadas na RSL (Parte 2)

	Contexto								Objetivo		Tipo de Dados						Tarefa EDM				Técnica			
	Educação Básica/Escolar	Técnico	Graduação	Pós-Graduação	EAD	Presencial	Pública	Privada	Atributos x Evasão	Modelo Preditivo	Acadêmico	Demográfico	Econômico	Interacional	Sentimento	Temporal	Associação	Classificação	Clusterização	Regressão	Balanceamento	Seleção de Atributos	Previsão (matriculados)	Interface/Sistema
Pachas et al. (2021)			✓			✓		✓	✓	✓	✓	✓						✓						
Palacios et al. (2021)			✓					✓	✓	✓	✓	✓	✓					✓			✓			
Park; Yoo (2021)			✓		✓					✓	✓	✓	✓	✓				✓			✓	✓		
Perchinunno et al. (2021)			✓	✓			✓		✓		✓	✓						✓						
Perez et al. (2018)			✓					✓	✓	✓	✓	✓						✓						
Prada et al. (2020)			✓						✓		✓	✓						✓	✓			✓		✓
Queiroga et al. (2022)	✓	✓				✓			✓	✓	✓	✓	✓					✓			✓	✓	✓	✓
Queiroga et al. (2020)		✓			✓		✓			✓				✓				✓						
Rovira et al. (2017)			✓				✓			✓	✓							✓			✓			
Santos et al. (2020)			✓			✓	✓		✓	✓	✓	✓						✓				✓		
Segura et al. (2022)			✓			✓	✓		✓	✓	✓	✓	✓					✓				✓		
Shiau (2020)			✓			✓		✓	✓	✓	✓	✓	✓					✓					✓	✓
Shilbayeh; Abonamah (2021)				✓				✓	✓	✓	✓	✓	✓				✓			✓			✓	
Solis et al. (2018)			✓				✓			✓	✓	✓	✓					✓						
Sorensen (2019)	✓					✓	✓		✓	✓	✓	✓	✓					✓						
Tsai et al. (2020)			✓			✓		✓	✓	✓	✓	✓	✓					✓				✓	✓	
Urbina–Najera; Mendez-Ortega (2022)			✓			✓			✓	✓	✓	✓	✓					✓			✓	✓	✓	✓
Verdugo et al. (2022)			✓							✓	✓	✓	✓					✓			✓	✓		
Vega et al. (2022)			✓			✓	✓		✓	✓	✓	✓	✓					✓			✓	✓	✓	✓
Villegas–Ch et al. (2020)			✓		✓				✓	✓	✓	✓	✓	✓				✓						
Viloria et al. (2019)			✓				✓		✓	✓	✓	✓	✓					✓						
Xu; Wilson (2021)	✓					✓	✓			✓	✓	✓	✓					✓			✓		✓	✓
Yang et al. (2021)			✓				✓		✓		✓	✓						✓				✓		
Yoo et al. (2017)			✓				✓		✓		✓	✓				✓	✓							
Yu et al. (2021)			✓		✓	✓	✓		✓	✓	✓	✓	✓					✓			✓			
Este trabalho			✓			✓	✓		✓	✓	✓	✓	✓	✓				✓			✓	✓		

Fonte: Elaborada pela autora.

Em relação às informações utilizadas, ficou evidenciado que, além de uma notável variação na quantidade de registros considerados, praticamente a totalidade dos estudos adota de um a três tipos de dados na representação dos estudantes, sendo utilizadas, em geral, combinações entre atributos acadêmicos, sociais e/ou econômicos. Dentre os fatores de maior potencial preditivo, os de desempenho acadêmico (relacionados a notas e assiduidade, por exemplo) se mostraram mais recorrentes. Consequentemente, trabalhos que construíram modelos relacionados a diferentes estágios da trajetória acadêmica dos alunos observaram que, quanto mais avançado o momento de predição e maior o volume de dados acadêmicos disponível, melhores os resultados preditivos. Porém, como também foi observada uma maior tendência de evasão em semestres iniciais, predições precoces, geralmente apoiadas em dados socioeconômicos, mantêm-se importantes.

Sob a perspectiva das técnicas e ferramentas adotadas, a maioria dos trabalhos

aborda o problema da evasão por meio da tarefa de classificação, podendo-se destacar o uso de modelos de DT, considerados na forma individual ou combinada (*ensembles*), e das linguagens Python e R. Além disso, embora já fosse esperada a adoção contida de técnicas de seleção de atributos, considerando que muitos trabalhos já partem de uma quantidade reduzida de variáveis, a baixa aplicação de técnicas de balanceamento causou surpresa, à medida que os conjuntos de dados utilizados no contexto da evasão costumam apresentar diferenças proporcionais consideráveis entre as classes, o que pode prejudicar a qualidade dos modelos.

Ainda considerando características técnicas, embora o interesse na previsão de evasão apoiada por EDM esteja crescendo, a grande maioria dos estudos realiza análises apenas em dados do passado, sem aplicar os modelos preditivos a dados atuais, de alunos ativos/matriculados, para realmente prever estudantes em risco de evasão. Além disso, nem todos os trabalhos que realizam previsões disponibilizam aos gestores acadêmicos interfaces/ferramentas para a consulta, de forma intuitiva e contínua, de seus resultados, a fim de auxiliar na tomada de decisões e no planejamento de estratégias preventivas. Esses resultados indicam um subaproveitamento dos esforços e potenciais de grande parte das pesquisas na prática educacional.

Para além das características envolvidas nas questões de pesquisa desta RSL, também foram apontadas algumas tendências gerais e, em alguns casos, preocupantes. Uma delas é a carência de pesquisas associadas a países de médio e baixo IDH, precisamente os que mais precisariam desses estudos. Também se observou pouca adaptação das técnicas ao domínio da predição de evasão. Ou seja, embora considerem questões/objetivos de pesquisa direcionados ao problema da evasão, em geral, os trabalhos adotam técnicas de mineração de dados e aprendizado de máquina em suas formas tradicionais, sem adaptá-las às particularidades do domínio. Além disso, mostrou-se comum o desenvolvimento de modelos especializados em cursos e/ou momentos de predição. Embora tais modelos possam ter sua capacidade preditiva beneficiada pela exploração de fatores mais específicos, como o desempenho de estudantes em determinadas disciplinas, essa abordagem provê resultados menos abrangentes, institucionalmente.

Acompanhando tendências identificadas na RSL, o trabalho proposto e desenvolvido nesta Tese considera, como estudo de caso, o contexto de estudantes de cursos de graduação presenciais do IFFar, uma instituição pública, no desenvolvimento de modelos preditivos de evasão e na identificação dos principais fatores envolvidos. Em relação aos dados, buscando ofertar uma maior variabilidade de recursos aos padrões preditivos, esta Tese se junta aos poucos estudos que utilizam mais de três aspectos de dados em seus modelos, e considera dados acadêmicos, demográficos/sociais, econômicos, e interacionais. Sob a perspectiva das técnicas utilizadas, assim como a grande maioria dos estudos, esta Tese trata o problema de predição de evasão

como uma tarefa de classificação binária. Porém, ao contrário de muitas pesquisas, esta Tese extrai uma grande quantidade de características/variáveis dos estudantes, e considera técnicas de seleção de atributos na construção dos modelos preditivos, além do balanceamento/reamostragem dos dados, com o intuito de oportunizar uma maior capacidade preditiva aos modelos. Por fim, por se tratar de um trabalho acadêmico em que o acesso aos dados foi atrelado à anonimização e cujo objetivo é propor uma estrutura de classificação voltada especificamente à predição genérica de evasão, a proposta desta Tese, como a maioria das pesquisas, é validada apenas sobre dados do passado (de estudantes com situação final – de evasão ou conclusão – já definida/conhecida), e não abrange o desenvolvimento de um sistema de previsão da situação de estudantes ativos/matriculados.

Embora as Tabelas 4 e 5 demonstrem semelhanças e diferenças contextuais, técnicas e de dados entre todas as pesquisas, também cabe posicionar conceitualmente o trabalho desenvolvido nesta Tese frente ao estudo de Fernández-García et al. (2021), que mais se assemelha a este trabalho na forma como utiliza conhecimentos do domínio educacional para adaptar/estruturar a aplicação das técnicas de EDM e aprendizado de máquina, na tarefa preditiva de evasão. Mais especificamente, Fernández-García et al. (2021) desenvolvem modelos preditivos especializados em diferentes estágios de predição (na matrícula/admissão e ao final de cada um dos quatro primeiros semestres), sendo que cada modelo, a partir do primeiro semestre, considera entre seus atributos o resultado gerado pelo modelo especializado no momento de predição anterior, a fim de explorar e aproveitar o conhecimento gerado anteriormente. Essa metodologia, assim como o trabalho desenvolvido nesta Tese, usa de conhecimento do domínio para representar a evasão como o resultado de um processo contínuo, que considera a predição da situação dos estudantes em semestres anteriores e, consequentemente, a variabilidade dos padrões preditivos ao longo dos semestres. Porém, nesta Tese, a predição é feita a partir de um comitê de classificação (*ensemble*) formado por subclassificadores especializados em diferentes períodos letivos, considerando, inclusive, semestres posteriores ao do estudante, a fim de possibilitar a identificação antecipada de padrões que caracterizem a evasão em semestres futuros. Além disso, enquanto Fernández-García et al. (2021) constroem modelos especialistas para serem utilizados nas predições de evasão de seus respectivos semestres, o comitê de classificação desenvolvido nesta Tese (MPC-SDP) visa a predição genérica de evasão, sendo, dessa forma, destinado a predições em qualquer semestre letivo. Adicionalmente, como será detalhado a seguir, no Capítulo 4, o MPC-SDP visa ampliar a representatividade de diferentes aspectos dos estudantes (acadêmico, social, econômico e interacional) entre seus padrões preditivos, além de possibilitar a interpretação de seus subclassificadores.

4 PROPOSTA: MULTI-PERSPECTIVE CLASSIFIER FOR STUDENT DROPOUT PREDICTION (MPC-SDP)

Considerando a importância de prever o risco de evasão estudantil para o fomento de estratégias preventivas, como Tese de doutorado, propõe-se o desenvolvimento de um modelo de predição de evasão genérico, destinado a prever a evasão em diferentes cursos e estágios da trajetória acadêmica. O interesse por um modelo genérico decorre da maior abrangência de sua aplicação e, consequentemente, do maior alcance de seus resultados e contribuições.

Diferentemente da maioria dos estudos apresentados no Capítulo 3, propõe-se uma estrutura de aplicação de EDM especializada ao problema de evasão estudantil, com o intuito de prover um modelo de predição único/genérico, mas que dê maior atenção a particularidades como:

- **Variabilidade dos padrões de evasão perante diferentes períodos letivos** – Conforme evidenciado em estudos apresentados no Capítulo 3, a importância preditiva dos atributos e dos padrões relacionados à evasão pode variar conforme o estágio no qual o estudante se encontra no curso. A pesquisa de Palacios et al. (2021), por exemplo, identificou maior importância de atributos relacionados ao desempenho escolar e ao contexto socioeconômico, para modelos especializados em prever a evasão nos dois primeiros semestres de curso. Já para a predição de evasões no terceiro semestre, os autores apontaram a preconização de atributos relacionados ao desempenho acadêmico universitário.

- **Importância da representatividade de diferentes aspectos entre os padrões de evasão** – Consoante com o item anterior, atributos de determinados aspectos do estudante (como sociais e econômicos) podem ser ignorados pelos modelos, se fatores relacionados a outro aspecto (como acadêmico) tiverem uma capacidade preditiva maior. Embora possa ser favorável ao desempenho preditivo, isso limita o conhecimento representado pelo modelo e, consequentemente, a descoberta de fatores que, apesar de menos importantes, ainda poderiam beneficiar o estabelecimento de medidas preventivas. Ou seja, embora modelos de DT

costumem ser adotados na predição de evasão em razão da interpretabilidade, os conhecimentos resultantes podem não ser úteis (novos), caso relacionem a evasão apenas ao baixo desempenho acadêmico, por exemplo.

Com o intuito de considerar a variabilidade dos padrões de evasão perante diferentes períodos letivos, propõe-se a construção de um modelo de *ensemble* genérico, destinado a prever evasões em diferentes períodos/semestres da trajetória acadêmica dos estudantes, mas que considere modelos de classificação (subclassificadores) especializados por períodos em seu comitê de decisão. Com isso, pretende-se que os dados atuais de um estudante, independentemente de seu estágio no curso, sejam avaliados por modelos especializados em diferentes períodos. Dessa forma, mesmo que um aluno apresente padrão de permanência para o modelo especializado em prever evasões em seu semestre atual, caso sejam satisfeitos padrões de evasão dos modelos especializados nos demais períodos, esses padrões também serão considerados na decisão do modelo de *ensemble*. Essa abordagem pode contribuir para a antecipação da identificação de estudantes em risco, por não considerar apenas os padrões de evasão mais abrangentes (modelos genéricos tradicionais) ou os padrões relacionados a um único momento de predição (modelos especializados).

Além disso, para oportunizar de forma mais contundente a participação de atributos relacionados a diferentes aspectos dos estudantes nos padrões preditivos, propõe-se que, antes do treinamento de cada um dos subclassificadores (especializados por semestre), seja realizado um processo de seleção de atributos orientado por aspecto. Ou seja, propõe-se que os dados de treino de cada um dos modelos especializados sejam divididos por aspecto, em dados acadêmicos, contextuais, econômicos, interacionais, e sociais/demográficos, e seja realizado um processo de seleção de atributos para cada um desses subconjuntos de dados. Assim, ao invés de considerar apenas os atributos de maior capacidade preditiva geral, o conjunto final de atributos selecionados contemplará os atributos mais importantes de cada aspecto. Espera-se, com isso, aumentar a representatividade de diferentes aspectos dos estudantes entre os padrões de evasão, a fim de prover uma percepção mais abrangente acerca dos fatores potencialmente relacionados ao abandono estudantil.

Por visar o desenvolvimento de um modelo de classificação genérico, mas guiado por perspectivas especializadas e importantes para o domínio de predição da evasão estudantil, esta proposta recebeu o nome de **MPC-SDP**, acrônimo, em Inglês, para ***Multi-Perspective Classifier for Student Dropout Prediction***. Mais especificamente, essas perspectivas resultam da exploração de padrões de evasão relacionados a múltiplos períodos, separadamente; e da priorização de múltiplos aspectos (acadêmicos, contextuais, econômicos, interacionais, e sociais/demográficos) na representação do comportamento de evasão dos estudantes.

Na Figura 15, é apresentada uma visão geral da estrutura de EDM proposta para

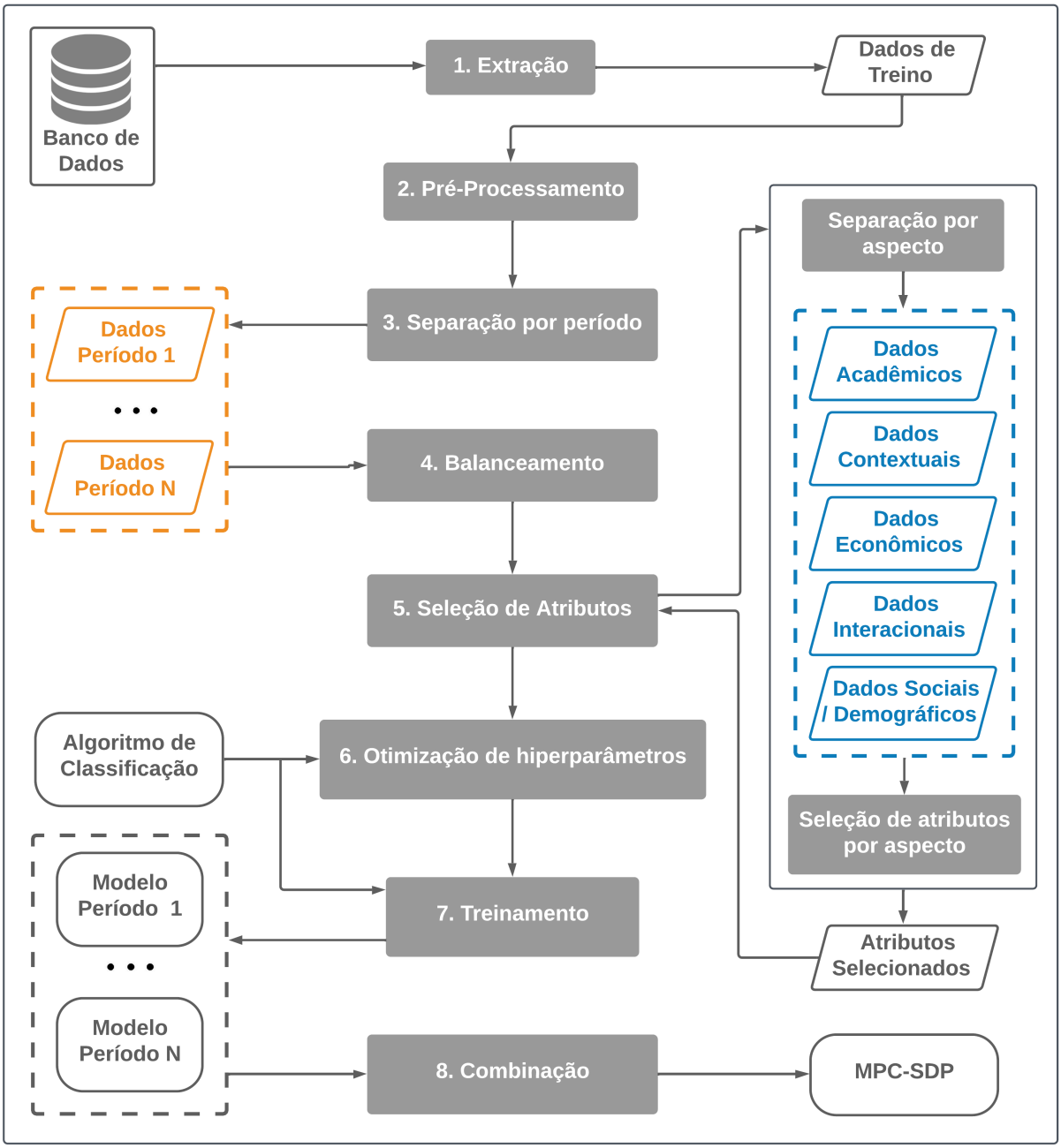
o desenvolvimento do MPC-SDP. Como em todo processo de EDM, inicialmente, faz-se necessário obter e preparar os dados a serem utilizados. Por isso, a **primeira** e a **segunda etapa** correspondem, respectivamente, às atividades de extração e pré-processamento dos dados[1]. Porém, como o MPC-SDP visa a predição de evasão em diferentes cursos e estágios da trajetória acadêmica, faz-se importante que as atividades de pré-processamento sejam cuidadosamente planejadas para oportunizar a equiparação e a generalização de dados que, em suas formas brutas/absolutas, podem não ser diretamente comparáveis entre si. Mais especificamente, o valor de um atributo absoluto, como a carga horária integralizada pelo estudante, pode carregar significados distintos perante cursos com cargas horárias e estruturas curriculares muito diferentes (como os de tecnologia e bacharelado, por exemplo). Nesse caso, a inclusão de atributos relativos, como o percentual de carga horária integralizada, pode oportunizar uma maior capacidade de generalização ao modelo.

Como um dos principais diferenciais da construção do MPC-SDP, após o pré-processamento dos dados de treino, propõe-se que, na **terceira etapa**, os dados dos estudantes sejam divididos segundo os períodos/semestres letivos que eles representam. Essa etapa visa preparar os subconjuntos de treinamento que serão considerados na construção dos subclassificadores (especializados por períodos). Porém, antes de serem utilizados, esses subconjuntos de dados devem ser submetidos, na **quarta etapa**, a um processo de balanceamento, uma vez que cada subconjunto pode apresentar proporções de representação bastante distintas entre as classes de estudantes evadidos e formados. Além disso, propõe-se que cada subconjunto de treino especializado por período seja, na **quinta etapa**, submetido a um processo de seleção de atributos orientado por aspecto. Nesse processo, conforme descrito anteriormente, cada subconjunto de dados deve ser dividido em novos subconjuntos, a fim de contemplar, separadamente, atributos relacionados a aspectos acadêmicos, contextuais, econômicos, interacionais e sociais/demográficos dos estudantes. Então, para cada novo subconjunto/aspecto de um determinado período, selecionam-se seus principais atributos, de modo que o resultado da seleção contemple, obrigatoriamente, atributos de todos os aspectos.

Definidas as representações finais dos subconjuntos de treino de cada período acadêmico, a **sexta** e a **sétima etapa** da Figura 15 correspondem, respectivamente, aos processos de otimização de hiperparâmetros e de treinamento dos subclassificadores (especializados por período). Nessa etapa, cada subconjunto de treino, relacionado a um período acadêmico, deve ser utilizado na configuração e na construção de seu respectivo modelo especialista, de modo que sejam disponibilizados “N” modelos

[1]Note que, por uma questão de simplificação, a Figura 15 apresenta apenas o fluxo de treinamento/-construção do MPC-SDP. Dessa forma, os processos de preparação dos dados de teste e de avaliação do modelo foram omitidos.

Figura 15 – Visão geral da estrutura de EDM proposta para a construção do MPC-SDP



Fonte: Elaborada pela autora.

preditivos de evasão, cada um especializado em um determinado período/momento de predição. Ao serem combinados, na **oitava etapa**, esses "N" modelos especializados por período darão origem a um comitê de classificação/predição, que constitui o MPC-SDP.

Além de oportunizar uma avaliação mais atenta a perspectivas de diferentes períodos e aspectos de predição, o MPC-SDP pode permitir a interpretação de seus padrões preditivos, caso seja considerado um algoritmo de classificação interpretável, como DT ou LR, na construção de seus subclassificadores (especializados por período). Isso porque, embora se caracterize como um modelo de *ensemble*/comitê, o número de subclassificadores a ser considerado no MPC-SDP tende a ser bem menor que o utilizado em modelos de *ensemble* padrão, como os baseados no algoritmo RF, geralmente compostos por 100 ou mais DTs. Considerando que os subclassificadores do MPC-SDP são organizados por períodos/semestres acadêmicos e que as estruturas curriculares dos cursos de graduação costumam se limitar a 8 ou 10 semestres, o MPC-SDP tenderia a utilizar um número cerca de dez vezes menor de subclassificadores, se comparado a um modelo RF padrão, por exemplo. Dessa forma, ao contrário de outros *ensembles*, ainda seria viável analisar os padrões preditivos de cada um dos modelos que compõem o MPC-SDP.

5 DESENVOLVIMENTO

Como mencionado no Capítulo 1, embora possua uma estrutura conceitual reprodutível, nesta Tese, o MPC-SDP foi desenvolvido e avaliado considerando, como estudo de caso, o contexto de estudantes de cursos de graduação presenciais do IFFar. Enquanto a instituição foi escolhida em razão do vínculo funcional da autora, as delimitações de nível e modalidade de ensino fundamentaram-se, respectivamente, na maior concentração de evasão entre estudantes de graduação (entre os níveis de ensino priorizados institucionalmente) e na grande predominância da modalidade presencial, no IFFar (Brasil, 2023).

Neste Capítulo são descritas, nas Seções 5.1 e 5.2, as atividades de extração e pré-processamento dos dados de estudantes do IFFar, necessárias, posteriormente, para o treinamento e a avaliação do MPC-SDP. Após, na Seção 5.3, são descritas as implementações do método de seleção de atributos orientado por aspecto e do comitê de classificação com subclassificadores especializados por semestre, que compõem o MPC-SDP.

5.1 Extração de dados

Como primeira etapa do desenvolvimento do MPC-SDP, foi realizada a extração de dados de estudantes de graduação do IFFar, da modalidade presencial, a fim de representar as classes positiva (estudantes evadidos) e negativa (estudantes concluídos/formados) de evasão, perante diferentes aspectos e períodos/semestres acadêmicos. A seguir, nas Seções 5.1.1 e 5.1.2, são apresentados detalhes da metodologia de extração e dos dados coletados, respectivamente.

5.1.1 Metodologia

Inicialmente, a anuência institucional para acesso aos dados dos estudantes foi solicitada para a então reitora do IFFar, Profa. Dra. Carla Comerlato Jardim, por meio de uma carta de apresentação da pesquisa, enviada por e-mail. Com o retorno positivo, em setembro de 2020, o estudo dos dados foi iniciado, a partir do banco

de dados do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) do IFFar. Parte de um conjunto de Sistemas Integrados de Gestão, desenvolvido pela Universidade Federal do Rio Grande do Norte e adotado por mais de 50 instituições de ensino brasileiras (STI/UFRN, 2022), o SIGAA inclui um AVA e é utilizado no IFFar desde 2014 (Colpo; Primo; Aguiar, 2021).

Após o estudo dos dados disponíveis na base do SIGAA, foram construídas consultas SQL (do inglês, *Structured Query Language*) para viabilizar a coleta dos dados de interesse. Além das já citadas restrições de nível (graduação) e modalidade (presencial) de ensino, a extração teve como foco estudantes que evadiram ou concluíram seus cursos, após terem se matriculado, pela última vez, entre os anos de 2017 e 2022[1]. Faz-se importante destacar que, dentre as matrículas finais de cada ano de referência, foram considerados evadidos e concluídos os estudantes com situação "cancelado" e "concluído", respectivamente.

Considerando que o SIGAA começou a ser utilizado institucionalmente em 2014, a extração também foi condicionada a abranger apenas estudantes cujo ingresso no IFFar tenha ocorrido a partir do início desse ano. Isso porque, embora também tenham sido inseridos no SIGGA, para acesso às turmas virtuais e demais funcionalidades, estudantes com ingresso anterior, mesmo que ativos, não teriam registros prévios armazenados no sistema. Seus históricos de interação com o sistema, por exemplo, se limitariam a parte de suas trajetórias acadêmicas. Por razão semelhante, também foram desconsiderados estudantes que nunca fizeram o auto-cadastro de seus usuários no SIGAA. Ou seja, estudantes que nunca tiveram interação com o sistema.

Como última restrição para a extração dos dados, foram desconsiderados estudantes que tenham sido desligados do IFFar por falecimento, cancelamento judicial, exclusão, cancelamento de cadastro, não regularização da transferência de outra IES, decisão administrativa, cancelamento por novo vestibular, cancelamento por reopção (transferência interna), e transferência para outra IES[2]. Em outras palavras, na implementação desta Tese, foi considerado evadido o estudante que, sem formalizar solicitação de trancamento, tenha tido seu vínculo com o curso cancelado, sem a devida conclusão/diplomação, por razões que não sejam as citadas inicialmente. Isso porque, a partir de um entendimento semelhante ao de Tinto (1975), embora possam apresentar situação "cancelado", os estudantes desligados por essas razões não representam, essencialmente, casos de evasão. Dessa forma, considerá-los entre os exemplos de treinamento poderia introduzir ruído ao processo de aprendizado do modelo preditivo.

[1] Note que, naturalmente, nem todos esses dados estavam disponíveis no início do trabalho de extração, tendo sido coletados de forma progressiva.

[2] Observe que essas razões/motivações de desligamento são cadastradas no sistema institucional do IFFar pelas Coordenações de Registros Acadêmicos, quando ocorre a formalização de um pedido de cancelamento do curso.

Em relação aos dados de interesse, como mencionado no Capítulo 4, buscou-se extrair atributos de diferentes aspectos, com o intuito de enriquecer a caracterização do comportamento dos estudantes. Mais especificamente, foram considerados atributos:

- **acadêmicos** – relacionados ao desempenho dos estudantes, como médias de notas e frequências em disciplinas já cursadas;
- **contextuais** – que situam/contextualizam os estudantes na instituição e em seus cursos, como período atual, unidade/*campus*, tipo e área de conhecimento do curso;
- **econômicos** – que refletem o contexto econômico dos estudantes e de sua família, como renda e quantidade de membros do grupo familiar;
- **interacionais** – relacionados à interação dos estudantes com as turmas virtuais/AVA do SIGAA, como participação em questionários, tarefas e fóruns disponibilizados nas disciplinas;
- **sociais/demográficos** – que indicam características sociais/demográficas dos estudantes, como sexo, raça/cor, e idade ao ingressar no IFFar.

Cada estudante, evadido ou formado/concluído, teve suas informações coletadas por período/semestre[3] cursado, de forma progressiva e acumulada, a fim de prover diferentes exemplos/registros ao longo de sua trajetória no curso. Ou seja, mesmo que um estudante tenha evadido em seu terceiro semestre acadêmico, seus registros de primeiro e segundo período também foram utilizados como exemplos de abandono, partindo do entendimento de que a evasão tende a decorrer de um processo contínuo, e não de uma causa isolada ou momentânea. Essa forma de representação foi escolhida a partir de uma avaliação empírica preliminar, inspirada no estudo de Solis et al. (2018) e publicada em Colpo; Primo; Aguiar (2021). Porém, considerando a importância de oportunizar predições precoces de evasão, nesta extração, além dos registros semestrais, foi incluído um registro contendo apenas os dados disponíveis no momento de admissão/ingresso (período 0) de cada estudante.

O processo de extração foi viabilizado a partir do desenvolvimento de um *script* em linguagem Python, cujo funcionamento é apresentado, sintetizadamente, no Algoritmo 1. Conforme representado na função *DataExtraction*, para cada ano de referência (de 2017 a 2022), foi obtida uma lista com a identificação dos estudantes de interesse e disparado, para cada estudante dessa lista, o procedimento de extração. Enquanto a

[3]No IFFar, os períodos letivos e os componentes curriculares (disciplinas) dos cursos de graduação possuem duração semestral.

seleção dos estudantes foi executada no banco de dados do SIGAA por meio da função *GetStudentIDs*, cada aluno teve seus dados extraídos a partir da função *Extract*. A extração foi executada simultaneamente entre diferentes estudantes, explorando o paralelismo dos dados. Embora representado pela função *Parallelize* no Algoritmo 1, esse paralelismo foi introduzido na implementação por meio da classe *Pool*, do pacote *multiprocessing* (Python, 2022).

A função *Extract* apresenta como foram coletados os dados de um determinado estudante, perante diferentes aspectos e períodos acadêmicos. Inicialmente, foram realizadas as extrações de dados contextuais e sociais/demográficos do estudante, representadas pelos procedimentos *GetContextualData* e *GetSocialData* (linhas 12 e 13). Isso porque, por serem de natureza permanente[4], esses dados independem do estágio no qual o estudante se encontra no curso e, desso forma, só precisam ser consultados uma única vez no banco de dados. Porém, para viabilizar a extração dos demais dados, de forma progressiva e acumulada por período acadêmico, foram consultadas todas as combinações de ano-semestre vinculadas ao histórico do estudante, conforme representado pela função *GetYearAndSemesterTuples* (linha 14). Assim, para cada período (combinação ano-semestre) cursado pelo estudante, foram extraídos os dados acadêmicos, econômicos e interacionais – Funções *GetAcademicData*, *GetEconomicData* e *GetInteractionalData* (linhas 17-19) – que representavam a situação do aluno nesses diferentes momentos. Dessa forma, o registro que representa/caracteriza o estudante em um determinado período letivo foi composto pela união de seus dados permanentes (contextuais e sociais) e de seus respectivos dados semestrais (econômicos, acadêmicos e interacionais) (linhas 20-26). Adicionalmente, após ter sido coletado o registro do primeiro período do estudante, seus dados foram utilizados na geração do registro de admissão (período 0), conforme representado no bloco condicional do Algoritmo 1 (linhas 27-30). Ou seja, dentre as informações que compunham o registro do período 1, foram selecionados os dados que já se encontravam disponíveis no momento do ingresso do estudante, para compor seu registro de admissão. Por fim, o conjunto de todos os registros do estudante (semestrais e de admissão), representado pela variável *student_data* no código, foi retornado pela função *Extract* (linha 33).

Assim, para cada ano de referência, após terem sido extraídos os registros dos estudantes de interesse, os dados foram anonimizados e salvos em um arquivo CSV (do inglês, *Comma-Separated Values*), conforme representado ao final do laço da função *DataExtraction* (linhas 5 e 6). A anonimização se fez necessária porque alguns identificadores dos estudantes (ID do usuário e matrícula) foram extraídos e utilizados, durante a execução do *script*, para viabilizar a extração dos dados de interesse. Porém,

[4]Apenas o atributo contextual que indica o período em que o estudante se encontra no curso é variável. Por isso, esse atributo foi extraído junto aos atributos acadêmicos.

Algoritmo 1: Extração dos conjuntos de dados dos estudantes, por ano.

```
1  DataExtraction ()
2      for year ← 2017 to 2022 by 1 do
3          student_ids ← GetStudentIDs(year)
4          results ← Parallelize(Extract, student_ids)
5          results ← AnonymiseData(results)
6          WriteCSV(results)
7      end
8      return
9
10 Extract (student_id)
11     student_data ← ∅
12     contextual_data ← GetContextualData(student_id)
13     social_data ← GetSocialData(student_id)
14     year_semester_tuples ← GetYearAndSemesterTuples(student_id)
15     foreach (year, semester) ∈ year_semester_tuples do
16         args ← [student_id, year, semester]
17         academic_data ← GetAcademicData(args)
18         economic_data ← GetEconomicData(args)
19         interactional_data ← GetInteractionalData(args)
20         period_data ← ConcatData(
21             contextual_data,
22             social_data,
23             economic_data,
24             academic_data,
25             interactional_data
26         )
27         if contextual_data["period"] = 1 then
28             admission_data ← GetAdmissionData(period_data)
29             student_data.append(admission_data)
30         end
31         student_data.append(period_data)
32     end
33     return student_data
```

para a proteção dos dados dos estudantes, não era desejável que esses identificadores fizessem parte dos resultados da extração. Dessa forma, eles foram removidos e substituídos, sem função de reversão ou mapeamento, por um único identificador falso/*fake*, criado exclusivamente com o intuito de indicar que diversos registros semestrais pertenciam a um mesmo estudante.

Por fim, também foram obtidos dados externos ao SIGAA, a fim de incluir informações a respeito do contexto de desenvolvimento humano de origem dos estudantes. Para isso, foram considerados dados do Atlas do Desenvolvimento Humano no Brasil, mais especificamente do *ranking* do Índice de Desenvolvimento Humano Municipal (IDHM), disponibilizado em formato de planilha, em ATLAS BRASIL (2022).

5.1.2 Dados extraídos

As Tabelas 6 e 7 apresentam a relação dos 75 atributos contextuais, sociais, acadêmicos, econômicos, e interacionais, que foram extraídos para cada estudante, por período/semestre cursado.

Na Tabela 6, foram sintetizados sete atributos acadêmicos (36-42) a partir da descrição “índices acadêmicos do SIGAA”. A seguir, são elencados esses índices, junto das descrições apresentadas pelo próprio SIGAA, em seu relatório de índices de rendimento acadêmico (SINFO/UFRN, 2022a):

- **Média de Conclusão (MC)** – média do rendimento acadêmico final obtido pelo estudante nos componentes curriculares em que obteve êxito, ponderada pela carga horária discente dos componentes;
- **Média de Conclusão Normalizada (MCN)** – padronização da MC do aluno, considerando a média e o desvio padrão das MCs de todos os estudantes que concluíram o mesmo curso/modalidade nos últimos cinco anos;
- **Índice de Rendimento Acadêmico (IRA)** – média do rendimento acadêmico final obtido pelo estudante nos componentes curriculares que concluiu, ponderadas pela carga horária discente dos componentes;
- **Índice de Eficiência em Carga Horária (IECH)** – divisão da carga horária com aprovação pela carga horária utilizada;
- **Índice de Eficiência em Períodos Letivos (IEPL)** – divisão da carga horária acumulada/integralizada pela carga horária esperada;
- **Índice de Eficiência Acadêmica (IEA)** – produto da MC pelo IECH e pelo IEPL;
- **Índice de Eficiência Acadêmica Normalizado (IEAN)** – produto da MCN pelo IECH e pelo IEPL.

Tabela 6 – Lista dos atributos contextuais, sociais e acadêmicos, extraídos para cada estudante

#	Atributo	Aspecto
1	identificador *fake*	–
2	ano de ingresso	Contextual
3	semestre de ingresso	Contextual
4	forma de ingresso	Contextual
5	unidade/campus	Contextual
6	curso	Contextual
7	tipo de curso	Contextual
8	turno(s) do curso	Contextual
9	área de conhecimento do curso	Contextual
10	período no curso	Contextual
11	idade ao ingressar no curso	Social
12	sexo	Social
13	cor/raça	Social
14	estado civil	Social
15	cursou ensino médio em escola pública	Social
16	anos entre o fim do ensino médio e o ingresso	Social
17-18	município e estado de residência	Social
19-20	município e estado de naturalidade	Social
21	tipo de reserva de vaga	Social
22	total de matrículas em disciplinas	Acadêmico
23	total de aprovações	Acadêmico
24	total de reprovações	Acadêmico
25	total de disciplinas trancadas	Acadêmico
26	total de exames	Acadêmico
27	média total de notas	Acadêmico
28	desvio padrão entre as notas	Acadêmico
29	média total de frequências	Acadêmico
30	média de notas no período	Acadêmico
31	desvio padrão entre as notas do período	Acadêmico
32	média de frequências no período	Acadêmico
33	carga horária do curso	Acadêmico
34	total de carga horária integralizada	Acadêmico
35	total de carga horária integralizada no período	Acadêmico
36-42	índices acadêmicos do SIGAA	Acadêmico
43	total de participação em projetos	Acadêmico
44	total de dias desde o ingresso	Acadêmico
45	total de dias em falta justificada/estudo domiciliar	Acadêmico
46	total de trancamento(s) do curso	Acadêmico
47	total de dias em situação de trancamento	Acadêmico
48	total de dias em situação ativa no período	Acadêmico
49	situação final no curso	Acadêmico

Fonte: Elaborada pela autora.

Tabela 7 – Lista dos atributos econômicos e interacionais, extraídos para cada estudante (Continuação da Tabela 6)

#	Atributo	Aspecto
50	renda bruta mensal do núcleo familiar	Econômico
51	quantidade de membros do núcleo familiar	Econômico
52	pontuação no Cadastro Único	Econômico
53	total de auxílios recebidos	Econômico
54	total de enquetes lançadas nas turmas	Interacional
55	total de enquetes respondidas	Interacional
56	total de questionários lançados nas turmas	Interacional
57	total de questionários respondidos	Interacional
58	total de fóruns lançados nas turmas	Interacional
59	total de fóruns respondidos	Interacional
60	total de tarefas lançadas nas turmas	Interacional
61	total de tarefas respondidas	Interacional
62	número de colegas com quem dividiu grupos de tarefas	Interacional
63	total de ações em turmas virtuais	Interacional
64	total de ações em turmas virtuais no período	Interacional
65	total de acessos ao SIGAA	Interacional
66	total de acessos ao SIGAA no período	Interacional
67	total de mensagens enviadas	Interacional
68	total de mensagens enviadas a docentes	Interacional
69	total de mensagens enviadas a estudantes	Interacional
70	total de mensagens recebidas	Interacional
71	total de mensagens recebidas de docentes	Interacional
72	total de mensagens recebidas de estudantes	Interacional
73	total de pessoas com quem trocou mensagens	Interacional
74	total de docentes com quem trocou mensagens	Interacional
75	total de estudantes com quem trocou mensagens	Interacional

Fonte: Elaborada pela autora.

Em relação aos atributos econômicos, faz-se importante mencionar que os dados relacionados à renda e ao núcleo familiar (atributos 50 e 51) são de preenchimento obrigatório ao estudante, para a alteração de seus dados pessoais de endereço e contato (SINFO/UFRN, 2022b). A atualização desses dados pessoais, por sua vez, costuma ser exigida no processo de solicitação de matrícula *on-line* (DTI/IFFAR, 2022), executado pelos estudantes antes de cada período letivo, com exceção do primeiro[5]. Além disso, o preenchimento de um questionário socioeconômico, denominado de Cadastro Único (referenciado no atributo 52) no SIGAA, é pré-requisito para a solicitação de bolsas e auxílios estudantis. Como algumas perguntas do Cadastro Único

[5]No IFFar, a primeira matrícula dos alunos ingressantes é realizada pelas Coordenações de Registros Acadêmicos dos *campi*, junto ao processo de confirmação de vaga

possuem respostas valoradas, as pontuações obtidas pelos estudantes refletem suas situações de vulnerabilidade socioeconômica.

Na Tabela 8 são apresentados, por ano de referência, os quantitativos de estudantes que tiveram seus dados extraídos (concluídos e evadidos)[6], além dos totais de registros coletados. Como cada aluno teve múltiplos registros extraídos, conforme a quantidade de períodos cursados, os totais de registros são, naturalmente, maiores que os de alunos concluídos ou evadidos. Além disso, como as evasões costumam ocorrer nos semestres iniciais, os estudantes evadidos possuem um número bem menor de registros semestrais. Dessa forma, apesar do desbalanceamento entre os quantitativos de estudantes evadidos e concluídos, os totais de registros/exemplos alcançam maior equilíbrio. Por fim, também ficam evidenciadas variações entre os percentuais anuais de evasão dos dados coletados e os da plataforma PNP, apresentados na Seção 2.1.1, devido às restrições consideradas na extração e descritas na Seção 5.1.1.

Tabela 8 – Quantitativos de estudantes e registros extraídos, por ano de referência

Ano	Situação	Estudantes	Registros
2017	Matriculados	3934	-
	Concluídos	35 (0.89%)	231
	Evadidos	620 (15.76%)	2054
2018	Matriculados	4609	-
	Concluídos	284 (6.16%)	2318
	Evadidos	737 (15.99%)	2749
2019	Matriculados	5220	-
	Concluídos	443 (8.49%)	3915
	Evadidos	1046 (20.04%)	4165
2020	Matriculados	5216	-
	Concluídos	399 (7.65%)	3752
	Evadidos	1053 (20.19%)	3918
2021	Matriculados	5368	-
	Concluídos	446 (8.31%)	4430
	Evadidos	1102 (20.53%)	4534
2022	Matriculados	5189	-
	Concluídos	548 (10.56%)	5379
	Evadidos	1263 (24.34%)	5655
Total	Concluídos	2155	20025
	Evadidos	5821	23075

Fonte: Elaborada pela autora.

[6]Note que, na Tabela 8, também são apresentados (entre parênteses) os percentuais representados pelos estudantes evadidos e concluídos, perante o total de matrículas de cada ano. Além disso, embora não conste na tabela, cabe mencionar que estudantes ingressantes representam, em média, cerca de 40% das matrículas de cada ano.

Na Tabela 9, são apresentados os dados disponíveis na planilha do *ranking* do IDHM brasileiro (ATLAS BRASIL, 2022). Observe que o IDHM brasileiro considera indicadores de renda, educação, e longevidade, mesmas dimensões consideradas no IDH Global. Porém, a metodologia de cálculo do IDHM brasileiro se adéqua ao contexto e à disponibilidade de indicadores nacionais (PNUD Brasil, 2022).

Tabela 9 – Lista de atributos da planilha do *ranking* de IDHM brasileiro

#	Atributo
1	Territorialidade (Município e Estado)
2	Posição IDHM
3	IDHM
4	Posição IDHM Renda
5	IDHM Renda
6	Posição IDHM Educação
7	IDHM Educação
8	Posição IDHM Longevidade
9	IDHM Longevidade

Fonte: Elaborada pela autora.

5.2 Pré-processamento

Após extraídos, os dados foram submetidos à etapa de pré-processamento, que buscou adequá-los ao problema da predição genérica de evasão e às restrições dos algoritmos de aprendizado de máquina de interesse. Realizado com o suporte das bibliotecas *numpy*, *pandas*, e *scikit-learn*, o pré-processamento resultou em 122 atributos, apresentados nas Tabelas 10 e 11. As atividades de pré-processamento realizadas para integração, limpeza, adição/derivação, e transformação dos dados brutos são descritas, respectivamente, nas Seções de 5.2.1, 5.2.2, 5.2.3, e 5.2.4. Após, na Subseção 5.2.5, é apresentada uma descrição estatística dos principais atributos resultantes da etapa de pré-processamento.

5.2.1 Integração

Inicialmente, foi realizada a integração entre os dados extraídos do SIGAA e os dados de IDHM. Ou seja, aos registros de cada estudante foram adicionados os dados de IDHM do seu respectivo município de origem/naturalidade. Para isso, foram mapeadas as informações de município e estado de naturalidade de cada estudante (atributos 19 e 20, da Tabela 6) com a territorialidade do IDHM (atributo 1, da Tabela 9). Embora envolvam diferentes dimensões, como os dados de IDHM ajudam a caracterizar o contexto de naturalidade dos estudantes, decidiu-se incluí-los entre os atributos sociais/demográficos. Mais especificamente, na Tabela 10, os atributos de 48 a 51

Tabela 10 – Lista dos atributos contextuais, sociais, e acadêmicos resultantes do pré-processamento

#	Atributo	Aspecto
1	identificador *fake*	–
2-8	indicadores de forma de ingresso	Contextual
9-18	indicadores de unidade/*campus*	Contextual
19-21	indicadores de tipo de curso	Contextual
22-24	indicadores de turno(s) do curso	Contextual
25-31	indicadores de área de conhecimento do curso	Contextual
32	período no curso	Contextual
33	idade ao ingressar no curso	Social
34	desvio da média de idade de ingresso	Social
35	indicador de sexo masculino	Social
36-40	indicadores de cor/raça	Social
41-43	indicadores de estado civil	Social
44	cursou ensino médio em escola pública	Social
45	anos entre o fim do ensino médio e o ingresso	Social
46	desvio da média de anos entre ensino médio e ingresso	Social
47	reside no município do *campus*	Social
48-51	IDHMs do município de naturalidade	Social
52	desvio da média de IDHM	Social
53-56	posições dos IDHMs do município de naturalidade	Social
57-60	indicadores de tipo de reserva de vaga	Social
61	minoria social	Social
62-63	médias de notas acumulada e no período	Acadêmico
64-65	desvios padrão de todas as notas, e as notas do período	Acadêmico
66	desvio da média de notas acumuladas	Acadêmico
67-68	médias de frequências acumulada e no período	Acadêmico
69	desvio da média de frequências acumuladas	Acadêmico
70-73	% de aprovações, reprovações, exames, e cancelamentos	Acadêmico
74	desvio da média de % de aprovações	Acadêmico
75-76	% de carga horária integralizada acumulada e no período	Acadêmico
77	desvio da média de % de integralização acumulada	Acadêmico
78	desvio da média de % de integralização no período	Acadêmico
79-85	índices acadêmicos do SIGAA	Acadêmico
86	média semestral de participações em projetos	Acadêmico
87	% de dias em falta justificada/estudo domiciliar	Acadêmico
88	% de dias em situação de trancamento	Acadêmico
89	diferença entre as médias de frequência 68 e 67	Acadêmico
90	diferença entre as médias de notas 63 e 62	Acadêmico
91	indicador de evasão (atributo classe)	Acadêmico

Fonte: Elaborada pela autora.

Tabela 11 – Lista dos atributos econômicos e interacionais, resultantes do pré-processamento (Continuação da Tabela 10)

#	Atributo	Aspecto
92	renda bruta mensal do núcleo familiar	Econômico
93	quantidade de membros do núcleo familiar	Econômico
94	renda bruta mensal per capita	Econômico
95	desvio da média de renda bruta per capita	Econômico
96	pontuação no Cadastro Único	Econômico
97	média semestral de auxílios recebidos	Econômico
98-101	% de enquetes, questionários, fóruns, tarefas respondidas	Interacional
102	% geral de resposta a todas as atividades da linha anterior	Interacional
103	desvio da média de % geral de resposta	Interacional
104	média semestral de colegas em grupos de atividades	Interacional
105	desvio da média semestral de colegas em grupos	Interacional
106-107	médias diárias de ações acumulada e no período	Interacional
108	desvio da média diária de ações acumulada	Interacional
109-110	médias diárias de acessos acumulada e no período	Interacional
111	desvio da média diária de acessos acumulada	Interacional
112	diferença entre as médias de ações 107 e 106	Interacional
113	diferença entre as médias de acessos 110 e 109	Interacional
114-115	médias semestrais de mensagens enviadas, e recebidas	Interacional
116-117	desvios das médias de mensagens enviadas, e recebidas	Interacional
118-119	% de mensagens enviadas, e recebidas de docentes	Interacional
120	média semestral de contatos com quem trocou mensagens	Interacional
121	% de docentes entre os contatos	Interacional
122	desvio da média semestral de contatos	Interacional

Fonte: Elaborada pela autora.

correspondem aos dados de IDHM, IDHM Renda, IDHM Educação, e IDHM Longevidade do município de naturalidade do estudante, enquanto os atributos de 53 a 56 indicam as posições desses IDHMs no *ranking* nacional. Durante a integração, alguns estudantes sem informação de naturalidade ou sem territorialidade correspondente tiveram o valor "-1" atribuído aos seus dados de IDHM, a fim de indicar a inexistência de informação.

5.2.2 Limpeza

Em relação à limpeza, inicialmente os dados foram analisados para a identificação de problemas cadastrais. Após, foi realizado o **tratamento das inconsistências**, identificadas nos seguintes atributos:

- Área de conhecimento (atributo 9, na Tabela 6) – cursos semelhantes foram cadastrados com áreas de conhecimento distintas, por diferentes *campi*. Para padronização/correção, foram analisadas as diferenças e mantidos os mapeamentos considerados mais adequados. A relação dos cursos, junto aos seus respectivos *campi* e áreas de conhecimento (já padronizadas), pode ser visualizada no Apêndice B;

- Idade de ingresso, e anos entre o fim do ensino médio e o ingresso (atributos 11 e 16, na Tabela 6) – 1481 registros relacionados a 260 estudantes[7] apresentavam valor nulo, maior que 100, ou menor que 16 ou 0, nos atributos que indicam a idade do estudante ao ingressar no curso e o intervalo de anos entre o fim do ensino médio e o ingresso, respectivamente. Essas inconsistências decorrem provavelmente de erros de digitação na data de nascimento e no ano de conclusão do ensino médio desses estudantes. Para correção, cada valor inconsistente na idade de ingresso de um aluno (nulo, menor que 16 ou maior que 100) foi substituído pela mediana dos valores válidos de seus pares. Como pares, foram considerados estudantes de mesmo período, unidade/*campus*, área de conhecimento e tipo de curso. Processo semelhante foi aplicado na correção dos valores inconsistentes (nulos, menores que 0 ou maiores que 100) do atributo de anos entre o fim do ensino médio e o ingresso. Porém, neste caso, a mediana foi obtida entre os pares que também compartilhavam da mesma idade de admissão/ingresso do estudante com registro inconsistente[8];

[7] Observe que os quantitativos citados nesta Seção consideram, unificadamente, todos os conjuntos de dados (com registros de estudantes concluídos e evadidos de 2017 a 2022). Porém, na prática e nos experimentos, o pré-processamento é realizado sobre cada combinação de treino e teste, separadamente.

[8] Em caso de falta de registro com valor válido e mesma idade de ingresso entre os pares, o cálculo considerou estudantes com idades de ingresso mais próximas (imediatamente inferior e superior à do registro inconsistente).

- Cursou ensino médio em escola pública (atributo 15, na Tabela 6) – 69 estudantes não apresentavam informação neste atributo, mesmo tendo reserva de vaga relacionada à escola pública. Foi incluído o valor “1” nos 465 registros relacionados a esses estudantes, a fim de resolver as inconsistências;

- Renda bruta mensal, e quantidade de membros do grupo familiar (atributos 50 e 51, na Tabela 7) – por serem de preenchimento obrigatório para acesso a algumas funcionalidades do SIGAA, ficou evidenciada a informação de dados fictícios nesses atributos, por alguns estudantes. Dessa forma, decidiu-se desconsiderar/apagar os valores de renda e de membros do grupo familiar que ultrapassassem, respectivamente, 30 mil reais e 15 pessoas, o que envolveu 113 registros semestrais de 29 estudantes;

- Atributos acadêmicos de frequência e notas (atributos de 27 a 32, na Tabela 5 6) – alguns registros apresentavam valores que ultrapassavam os limites lógicos de notas e frequências. Considerando que o IFFar trabalha com uma escala de notas de 0 a 10, e que as frequências consistem em porcentagens (de 0 a 100), foram removidos os registros de notas e frequências menores que 0 ou maiores que 10 e 100, respectivamente. Essa remoção abrangeu 46 registros relacionados a 28 estudantes.

Após corrigidas as inconsistências, foi realizado o **tratamento de dados faltantes** nos atributos numéricos[9]. Para sinalizar a inexistência das informações, dados numéricos faltantes em registros de admissão (período 0), de aspecto acadêmico, econômico e interacional, tiveram seus valores vazios preenchidos com “-1”. Considerando os registros relacionados aos demais períodos acadêmicos, atributos econômicos também tiveram seus dados faltantes preenchidos pelo valor “-1”, a fim de diferenciar a inexistência da informação da inexistência de renda/auxílio (valor “0”), por exemplo. Já atributos numéricos e de aspecto acadêmico tiveram seus valores nulos preenchidos com “0”, pois se entendeu que, nesse aspecto e a partir do primeiro período, a inexistência de informação já é um indicativo de desempenho. Se um estudante não tiver registro de carga horária integralizada, por exemplo, significa que ele não teve êxito em nenhuma disciplina. Por fim, para dados faltantes em atributos sociais numéricos, foi aplicado o método de imputação pela média dos vizinhos mais próximos (em inglês, *k-Nearest Neighbors Imputer*), disponível na biblioteca *scikit-learn*. O atributo mais afetado na imputação foi o que indica se o estudante é oriundo de escola pública (atributo 15 da Tabela 6), no qual cerca de 11% dos registros não possuía dados.

[9]Atributos categóricos não precisaram ter seus valores faltantes tratados, pois, como será descrito na Subseção 5.2.4, eles foram binarizados.

5.2.3 Adição/Derivação de Atributos

Como mencionado no Capítulo 4, a construção de modelos de predição de evasão genéricos exige um cuidado maior em relação à representação dos atributos. Por considerarem registros de diferentes períodos e cursos, é preciso oportunizar a esses modelos a equiparação e a generalização de dados que, em suas formas brutas/absolutas, podem não ser diretamente comparáveis entre si. Atributos absolutos, como o total de aprovações e o total de carga horária integralizada, por exemplo, podem ter significados variáveis perante diferentes períodos acadêmicos ou cursos com estruturas curriculares muito distintas (como os de tecnologia e bacharelado). Por isso, com o intuito de oportunizar uma maior capacidade de generalização ao modelo, **atributos absolutos foram utilizados na derivação/adição de atributos relativos**. Os já citados atributos de totais de aprovações e carga horária integralizada (atributos 23 e 34, da Tabela 6), por exemplo, foram substituídos pelas porcentagens de aprovações e carga horária integralizada (atributos 70 e 75, da Tabela 10).

Como apresentado nas Tabelas 10 e 11, além do cálculo de porcentagens, foram consideradas médias, diferenças entre médias do período e acumuladas, e desvios das médias, na derivação dos atributos relativos. Faz-se importante mencionar que os atributos de "desvio da média" se referem à diferença de um atributo de média do estudante perante o valor médio de seus pares nesse mesmo atributo. Por exemplo, o desvio da média de notas acumuladas (atributo 66, da Tabela 10) se refere à diferença entre a média de notas acumulada pelo estudante (atributo 62, da Tabela 10) e a média das médias de notas acumuladas por todos os registros de mesmo período, unidade/campus, área de conhecimento e tipo de curso. Além disso, para atributos relativos que podem assumir valores positivos e negativos, como os de desvios da média e diferenças, em casos de inexistência de base de cálculo no registro[10], o atributo de desvio recebeu o valor 0, a fim de manter uma informação de neutralidade. Já para os atributos relativos que só podem assumir valores positivos, considerou-se -1 como indicador de inexistência de dado.

Por fim, destaca-se a derivação dos atributos de renda bruta mensal per capita (atributo 94, da Tabela 11) e de indicação de minoria social (atributo 61, da Tabela 10). Enquanto o primeiro se refere à divisão da renda bruta mensal pela quantidade de membros do grupo familiar (atributos 92 e 93, da Tabela 11), o segundo indica se o registro se relaciona a um estudante que se enquadre em uma das seguintes condições:

- tenha ingressado por reserva de vaga relacionada a Pessoa com Deficiência (PcD), indicada entre os atributos de 57 a 60, da Tabela 10;

[10] Registros de admissão (de período 0), no qual o valor faltante da média de notas acumulada foi preenchido por -1, são um exemplo de inexistência de base de cálculo.

- possua raça/cor indígena, parda, ou preta, indicada entre os atributos de 36 a 40, da Tabela 10);
- apresente renda bruta mensal per capita, indicada no atributo 94 da Tabela 11, entre 0 e 1440. Este último limite se refere ao valor médio de um salário mínimo e meio, considerando as variações do salário mínimo brasileiro entre os anos de 2014 a 2022 (Ipea, 2023).

5.2.4 Transformações

Dentre as transformações, foram realizadas, inicialmente, **adequações/padronizações nos valores de alguns atributos categóricos**, a fim de torná-los mais abrangentes e, dessa forma, favorecer a identificação de padrões comuns. Diferentes subcategorias de reserva de vaga (atributo 21, da Tabela 6) relacionadas a estudantes egressos de escola pública, com renda familiar bruta per capita igual ou inferior (i), e superior (ii) a um salário-mínimo e meio, foram agrupadas nessas duas categorias ("EP $\leq$ 1.5" e "EP $>$ 1.5", respectivamente). Procedimento semelhante foi realizado sobre diferentes tipos de forma de ingresso e estado civil (atributos 4 e 14, da Tabela 6, respectivamente). Ingressos por processo seletivo, ENEM, e vestibular, por exemplo, foram agrupados em uma única categoria, denominada "Processo Seletivo". Já em relação ao estado civil, foram unidas as categorias "Casado(a)" e "União Estável", assim como as categorias "Divorciado(a)", "Separado(a)" e "Viúvo(a)", considerando que esses dois grupos apresentam significados comuns: a atual ou a anterior convivência com um(a) cônjuge/companheiro(a), respectivamente.

Adicionalmente, como os algoritmos de aprendizado de máquina da biblioteca *scikit-learn* só aceitam atributos numéricos, **os atributos categóricos foram binarizados** por meio da estratégia de codificação *one-hot*. Nessa abordagem, cada valor de um atributo categórico é transformado em um novo atributo binário, que indica a ocorrência da respectiva categoria nos registros (Raschka; Mirjalili, 2017). Por exemplo, ao ser codificado pela estratégia *one-hot*, o atributo categórico de tipo de curso (atributo 7, da Tabela 6) foi substituído por três novos atributos binários (atributos 19-21, da Tabela 10), que indicam se um estudante está vinculado a um curso de bacharelado, licenciatura, ou tecnologia. Considerando também a existência de atributos multi-categóricos, que podem apresentar múltiplas categorias por registro[11], a transformação *one-hot* foi implementada a partir das classes de pré-processamento *LabelBinarizer* e *MultiLabelBinarizer*, da biblioteca *scikit-learn*.

Por fim, como será explicado no Capítulo 6, dependendo do algoritmo de classificação utilizado como base do MPC-SDP ou como método de comparação/*baseline*, foi aplicada a normalização/padronização dos dados numéricos. Essa técnica não foi

[11] Exemplo de um atributo multi-categórico é o de turno(s) do curso (atributo 8, da Tabela 6), já que um curso pode ter disciplinas ministradas em mais de um turno.

considerada padrão por não beneficiar todos os algoritmos de classificação. Para algoritmos de DTs, que são invariantes às escalas dos atributos (Raschka; Mirjalili, 2017), por exemplo, além de não trazer benefício ao desempenho, essa transformação impediria uma interpretação direta das regras de decisão, já que os dados deixariam de apresentar escalas reais.

5.2.5 Descrição estatística dos dados pré-processados

Nas Tabelas 12 e 13 são apresentadas as descrições estatísticas dos principais atributos resultantes da etapa de pré-processamento, considerando seus valores máximo, médio, mínimo, de desvio padrão, e dos percentis 25, 50 (mediana), e 75. Note que as descrições são apresentadas para as classes positiva e negativa de evasão, separadamente (Sim = Evadidos, Não = Concluídos). Embora essas informações considerem, unificadamente, todos os conjuntos de dados pré-processados (com registros de estudantes concluídos e evadidos de 2017 a 2022), faz-se necessário pontuar que foram desconsiderados os valores preenchidos com "-1" no tratamento de dados faltantes, para que eles não interfiram indevidamente nas médias dos atributos. Por isso, as tabelas apresentam colunas de contagem, indicando a quantidade de valores válidos nos atributos, para cada classe. Para facilitar a visualização das diferenças entre as classes, também foi adicionada uma coluna com a diferença percentual das médias, com destaque às variações superiores a 20%. Além disso, os atributos estão agrupados por aspecto (social, econômico, contextual, acadêmico, e interacional) e referenciam, entre parênteses, suas correspondências com as Tabelas 10 e 11.

Dentre os atributos socioeconômicos da Tabela 12, pode-se observar que o público dos cursos de graduação presenciais do IFFar concentra mais estudantes jovens[12], oriundos de escola pública[13], brancos, solteiros, com grupos familiares pequenos (médias e medianas de três membros) e renda per capita inferior a um salário mínimo brasileiro (Ipea, 2023), mesmo no percentil 75. Observando as variações entre as médias, percebe-se que a classe de estudantes evadidos (Sim), quando comparada à de concluídos (Não), apresenta uma proporção significativamente maior de registros de estudantes do sexo masculino, pardos, e pretos, além de uma maior média de pontuação no Cadastro Único, sugerindo uma maior vulnerabilidade social entre os estudantes evadidos. Embora apresente uma maior proporção de ingressantes por vaga de Ampla Concorrência Geral (ACG), pode-se observar que a classe de estudantes evadidos, quando comparada à de concluídos, também demonstra um aumento na

[12] Embora as médias de idade de ingresso possam ser bastante afetadas por uma minoria idosa, as medianas são de 20 e 22 anos, para as classes negativa e positiva de evasão, respectivamente.

[13] Apesar de o conjunto de dados conter múltiplos registros por estudante (registros semestrais), como a descrição foi separada por classe, a média dos atributos binários, além de indicar a proporção de registros na classe, também se aproxima da proporção de estudantes na classe. Isso porque, em geral, estudantes evadidos costumam cursar quantidades próximas de semestres, assim como os concluídos.

proporção de registros de estudantes vinculados à reserva de vaga de oriundos de escola pública com renda per capita igual ou inferior a um salário mínimo e meio (EP <= 1.5), e uma redução nos vinculados à reserva de vaga de oriundos de escola pública com renda per capita superior a um salário mínimo e meio (EP > 1.5). Isso sugere, novamente, uma maior vulnerabilidade social entre os evadidos.

Em contrapartida, a classe de evasão também apresenta uma menor concentração de registros de estudantes em minoria social, que representa estudantes pretos, pardos, ou indígenas, que tenham ingressado por reserva de PcD, ou que tenham renda familiar per capita igual ou inferior a um salário mínimo e meio. Embora, em um primeiro momento, essa descrição pareça divergente, é preciso observar que menos da metade dos registros de estudantes evadidos possui informação de renda. Mais especificamente, nas colunas de contagem da Tabela 12, nota-se que 73% (14639/20005) dos registros de estudantes concluídos possui informação de renda, enquanto essa proporção é de apenas 47% (10931/23049) na classe de evadidos. Ou seja, muitos estudantes evadidos não foram contabilizados no atributo de minoria social por deixarem o IFFar sem registrar suas informações econômicas no SIGAA.

Considerando os atributos contextuais, nas Tabelas 12 e 13, pode-se observar que a classe positiva de evasão apresentou médias/proporções significativamente maiores nos atributos vinculados à área de Ciências Exatas e da Terra, às modalidades de ingresso de Portador de Diploma e Reingresso, a cursos de Tecnologia, e aos *campi* Alegrete, Frederico Westphalen, Panambi, Santo Ângelo, e São Borja. Complementarmente, além de uma menor média no atributo de período acadêmico[14], a classe de evasão demonstrou, de forma significativa, menores médias/proporções de registros nas áreas de Ciências Biológicas e Ciências Sociais Aplicadas, em cursos de Bacharelado, nos turnos Matutino e Vespertino, e nos *campi* Jaguari, Santa Rosa, e São Vicente do Sul[15].

Por fim, na Tabela 13, é possível visualizar que a classe positiva de evasão apresenta piores médias para todos os atributos acadêmicos e para a maioria dos interacionais[16]. Isso indica que, em geral, os estudantes que evadiram estavam associados a um pior desempenho acadêmico e uma menor interação com as turmas virtuais do SIGAA (AVA).

[14]Consequência natural do menor número de períodos/semestres cursados por estudantes evadidos.

[15]Note que, além de considerar dados de múltiplos anos (de 2017 a 2022), essas proporções de evasão por *campi*, diferentemente das apresentadas na Seção 2.1.1, são sobre o total de registros vinculados aos estudantes evadidos do IFFar, e não sobre o total de matrículas de cada unidade.

[16]Embora chame atenção, nos atributos acadêmicos, as grandes diferenças nas médias de porcentagens de matrículas/turmas interrompidas, de dias de faltas justificadas e estudos domiciliares, e de dias em trancamento, é possível observar, por meio do percentil 75, que a grande maioria dos estudantes possui valor “0” nesses atributos, em ambas as classes. Ou seja, as médias da classe de evadidos foi bastante influenciada por uma minoria de estudantes. Entre os atributos interacionais, situação semelhante pode ser observada na média diária de acessos e na média semestral de contatos, que parecem ter sido determinadas por cerca de 25% dos registros (percentil 75).

Tabela 12 – Descrição estatística dos principais atributos pré-processados, por classe de evasão

Atributo	Contagem		Máximo		Média		Δ Média	Mínimo		Desvio Padrão		25%		50%		75%	
	Não	Sim	Não	Sim	Não	Sim		Não	Sim	Não	Sim	Não	Sim	Não	Sim	Não	Sim
s_sexo_M (35)	20005	23049	1	1	0.41	0.53	**29%**	0	0	0.49	0.50	0	0	0	1	1	1
s_idade_ingresso (33)	20005	23049	61	70	23.80	25.31	6%	16	16	8.05	8.19	19	19	20	22	26	28
s_anos_em_ingresso (45)	20005	23049	41	42	5.27	6	14%	0	0	6.18	6.34	1	1	3	3	7	8
s_reside_cidade_campus (47)	20005	23049	1	1	0.51	0.59	16%	0	0	0.50	0.49	0	0	1	1	1	1
s_minoria_social (61)	20005	23049	1	1	0.69	0.55	**-20%**	0	0	0.46	0.50	0	0	1	1	1	1
s_escola_em_publica (44)	20005	23049	1	1	0.95	0.94	-1%	0	0	0.22	0.23	1	1	1	1	1	1
s_raca_Branco (36)	20005	23049	1	1	0.82	0.76	-7%	0	0	0.38	0.43	1	1	1	1	1	1
s_raca_Indígena (37)	20005	23049	0	1	0	0	-	0	0	0	0.05	0	0	0	0	0	0
s_raca_Pardo (38)	20005	23049	1	1	0.12	0.17	**42%**	0	0	0.33	0.38	0	0	0	0	0	0
s_raca_Preto (39)	20005	23049	1	1	0.03	0.04	**33%**	0	0	0.18	0.21	0	0	0	0	0	0
s_ec_Casado/União Estável (41)	20005	23049	1	1	0.13	0.13	0%	0	0	0.33	0.34	0	0	0	0	0	0
s_ec_Divorciado/Separado/Viúvo (42)	20005	23049	1	1	0.02	0.02	0%	0	0	0.13	0.13	0	0	0	0	0	0
s_ec_Solteiro (43)	20005	23049	1	1	0.78	0.74	-5%	0	0	0.41	0.44	1	0	1	1	1	1
s_reserva_vaga_ACG (57)	20005	23049	1	1	0.25	0.54	**116%**	0	0	0.43	0.50	0	0	0	1	0	1
s_reserva_vaga_EP <= 1.5 (58)	20005	23049	1	1	0.09	0.10	11%	0	0	0.29	0.30	0	0	0	0	0	0
s_reserva_vaga_EP > 1.5 (59)	20005	23049	1	1	0.08	0.07	-13%	0	0	0.27	0.26	0	0	0	0	0	0
s_reserva_vaga_PcD (60)	20005	23049	1	1	0	0	-	0	0	0.06	0.07	0	0	0	0	0	0
s_pos_idhm (53)	19933	22988	253	279	91.74	87.39	-5%	1	3	34.57	34.11	64	64	87	86	113	109
s_pos_idhm_educacao (54)	19933	22988	338	351	139.45	131.69	-6%	2	4	53.50	50.27	95	86	134	124	181	163
s_pos_idhm_longevidade (55)	19933	22988	324	343	105.80	103.51	-2%	1	2	37.42	38.85	86	85	105	116	121	118
s_pos_idhm_renda (56)	19933	22988	142	176	49.80	48.32	-3%	5	1	19	17.56	35	35	47	46	63	61
e_membros_familia (93)	14699	10968	12	10	3.06	3.11	2%	1	1	1.24	1.31	2	2	3	3	4	4
e_renda_familiar (92)	14642	10944	28001	28000	2180	1889	-13%	0	0	2026	1867	1000	998	1600	1500	3000	2500
e_renda_per_capta (94)	14639	10931	20000	14000	789	696	-12%	0	0	797	727	348	300	619	518	1000	950
e_pontuacao_cadu (96)	2719	1345	441	419	94	142	**51%**	0	0	99	101	0	73	77	119	147	211
e_media_auxilios (97)	17844	17227	1	1	0	0	-	0	0	0.02	0.02	0	0	0	0	0	0
c_area_Ciências Agrárias (25)	20005	23049	1	1	0.19	0.20	5%	0	0	0.39	0.40	0	0	0	0	0	0
c_area_Ciências Biológicas (26)	20005	23049	1	1	0.19	0.13	**-32%**	0	0	0.39	0.34	0	0	0	0	0	0
c_area_Ciências Exatas e da Terra (27)	20005	23049	1	1	0.26	0.40	**54%**	0	0	0.44	0.49	0	0	0	0	1	1
c_area_Ciências Sociais Aplicadas (28)	20005	23049	1	1	0.31	0.23	**-26%**	0	0	0.46	0.42	0	0	0	0	1	0
c_forma_ingresso_Portador Diploma (2)	20005	23049	1	1	0.02	0.04	**100%**	0	0	0.13	0.20	0	0	0	0	0	0
c_forma_ingresso_Processo Seletivo (3)	20005	23049	1	1	0.90	0.87	-3%	0	0	0.30	0.34	1	1	1	1	1	1
c_forma_ingresso_Reingresso (4)	20005	23049	1	1	0.01	0.03	**200%**	0	0	0.11	0.16	0	0	0	0	0	0
c_forma_ingresso_Transferência (5)	20005	23049	1	1	0.06	0.06	0%	0	0	0.25	0.24	0	0	0	0	0	0
c_periodo_atual (32)	20005	23049	16	16	4.35	2.23	**-49%**	0	0	3.02	2.38	2	0	4	1	7	3
c_tipo_curso_Bacharelado (19)	20005	23049	1	1	0.33	0.26	**-21%**	0	0	0.47	0.44	0	0	0	0	1	1
c_tipo_curso_Licenciatura (20)	20005	23049	1	1	0.37	0.36	-3%	0	0	0.48	0.48	0	0	0	0	1	1
c_tipo_curso_Tecnologia (21)	20005	23049	1	1	0.31	0.38	**23%**	0	0	0.46	0.49	0	0	0	0	1	1

Fonte: Elaborada pela autora.

Tabela 13 – Descrição estatística dos principais atributos pré-processados, por classe de evasão (Continuação da Tabela 12)

Atributo	Contagem		Máximo		Média		Δ Média	Mínimo		Desvio Padrão		25%		50%		75%	
	Não	Sim	Não	Sim	Não	Sim		Não	Sim	Não	Sim	Não	Sim	Não	Sim	Não	Sim
c_turno_Matutino (22)	20005	23049	1	1	0.23	0.14	**-39%**	0	0	0.42	0.34	0	0	0	0	0	0
c_turno_Noturno (23)	20005	23049	1	1	0.78	0.85	9%	0	0	0.41	0.36	1	1	1	1	1	1
c_turno_Vespertino (24)	20005	23049	1	1	0.21	0.14	**-33%**	0	0	0.41	0.35	0	0	0	0	0	0
c_unidade_Alegrete (9)	20005	23049	1	1	0.13	0.20	**54%**	0	0	0.33	0.40	0	0	0	0	0	0
c_unidade_F. Westphalen (10)	20005	23049	1	1	0.05	0.09	**80%**	0	0	0.22	0.29	0	0	0	0	0	0
c_unidade_Jaguari (11)	20005	23049	1	1	0.04	0.02	**-50%**	0	0	0.20	0.14	0	0	0	0	0	0
c_unidade_Júlio de Castilhos (12)	20005	23049	1	1	0.12	0.11	-8%	0	0	0.33	0.31	0	0	0	0	0	0
c_unidade_Panambi (13)	20005	23049	1	1	0.08	0.10	**25%**	0	0	0.26	0.30	0	0	0	0	0	0
c_unidade_Santa Rosa (14)	20005	23049	1	1	0.16	0.07	**-56%**	0	0	0.37	0.26	0	0	0	0	0	0
c_unidade_Santo Augusto (15)	20005	23049	1	1	0.11	0.10	-9%	0	0	0.31	0.29	0	0	0	0	0	0
c_unidade_Santo Ângelo (16)	20005	23049	1	1	0.04	0.07	**75%**	0	0	0.19	0.25	0	0	0	0	0	0
c_unidade_São Borja (17)	20005	23049	1	1	0.09	0.12	**33%**	0	0	0.28	0.33	0	0	0	0	0	0
c_unidade_São Vicente do Sul (18)	20005	23049	1	1	0.19	0.13	**-32%**	0	0	0.39	0.33	0	0	0	0	0	0
a_media_notas_us (63)	17844	17227	10	10	7.84	3.67	**-53%**	0	0	1.74	3.35	7.5	0	8.2	3.2	8.8	7.1
a_media_freq_us (68)	17844	17227	100	100	92.91	67.09	**-28%**	0	0	14.16	33.29	92.1	42.6	96.3	81.9	99.1	94.5
a_media_sem_projetos (86)	17844	17227	5	4	0.10	0.03	**-70%**	0	0	0.27	0.15	0	0	0	0	0.1	0
a_pct_aprovacoes (70)	17844	17227	100	100	89.57	52.49	**-41%**	0	0	15.68	34.78	85.7	22.6	96.0	55.6	100	84.3
a_pct_exames (72)	17844	17227	85.71	100	9.73	16.18	**66%**	0	0	11.25	16.66	0	0	6.1	12.5	15.4	26.2
a_pct_mat_interrompidas (73)	17844	17227	100	100	0.74	6.76	**814%**	0	0	3.96	19.96	0	0	0	0	0	0
a_pct_ch_integralizada_us (76)	17844	17227	100	87.47	11.95	5.17	**-57%**	0	0	5.30	6.26	9.8	0	11.6	2.5	13.9	10.1
a_pct_dias_FJED (87)	17844	17227	45.45	190.65	0.27	1.47	**444%**	0	0	1.73	6.18	0	0	0	0	0	0
a_pct_dias_trancamento (88)	17844	17227	84.94	99.36	0.26	1.86	**615%**	0	0	2.53	7.86	0	0	0	0	0	0
i_media_diaria_acessos (109)	17844	17227	10.02	255.50	0.45	0.65	**44%**	0	0	0.44	2.92	0.16	0.12	0.34	0.29	0.61	0.62
i_media_diaria_acessos_us (110)	17844	17227	186	367	0.88	0.44	**-50%**	0	0	4.88	3.55	0.18	0.04	0.42	0.18	0.76	0.48
i_media_diaria_acoes (106)	17844	17227	22.16	26.87	1.11	0.96	-14%	0	0	1.26	1.24	0.17	0.10	0.78	0.57	1.63	1.36
i_media_diaria_acoes_us (107)	17844	17227	605	234	2.49	0.94	**-62%**	0	0	14.97	2.48	0.16	0.01	1.09	0.34	2.18	1.29
i_media_colegas_grupos (105)	17844	17227	34	38	0.55	0.41	**-25%**	0	0	1.48	1.76	0	0	0	0	0.40	0
i_media_semestral_contatos (120)	17844	17227	6.33	44	0.13	0.21	**62%**	0	0	0.31	0.72	0	0	0	0	0.17	0.17
i_pct_contatos_docentes (121)	17844	17227	100	100	26.82	22.52	-16%	0	0	42.96	40.97	0	0	0	0	66.67	0
i_pct_msg_env_docente (118)	17844	17227	100	100	14.47	7.53	**-48%**	0	0	34.28	25.90	0	0	0	0	0	0
i_pct_msg_rec_docente (119)	17844	17227	100	100	19.04	19	0%	0	0	38.52	38.69	0	0	0	0	0	0
i_pct_enquetes_respondidas (98)	3052	2824	100	100	43.91	22.04	**-50%**	0	0	43.41	36.46	0	0	40	0	100	35.99
i_pct_foruns_respondidos (99)	7479	8198	100	100	24.95	13.07	**-48%**	0	0	30.60	25.97	0	0	13.21	0	42.86	13.64
i_pct_quest_respondidos (100)	7048	8354	100	100	59.95	33.35	**-44%**	0	0	30.85	32.04	47.06	0	66	28.57	80.82	58.69
i_pct_tarefas_respondidas (101)	932	1161	100	100	55.62	26.62	**-52%**	0	0	43.78	41.66	0	0	50	0	100	50
i_pct_geral_respostas (102)	8913	9923	100	100	41.39	24.73	**-40%**	0	0	30.48	27.85	12.50	0	44.44	15.09	62.92	43.92

Fonte: Elaborada pela autora.

5.3 MPC-SDP

As atividades de extração e pré-processamento de dados são essenciais para o desenvolvimento de qualquer modelo preditivo. O MPC-SDP, inclusive, devido suas particularidades, requer registros de dados que contemplem, separadamente, cada semestre cursado pelos estudantes, e que representem, preferencialmente, atributos relativos, conforme explicado nas Seções 5.1 e 5.2. Porém, a extração e o pré-processamento dos dados são etapas específicas a cada instituição/estudo de caso, e não compõem o método de classificação/predição em si, cuja implementação é descrita nesta Seção. Mais especificamente, nas Seções 5.3.1 e 5.3.2, são descritas as implementações da seleção de atributos orientada por aspecto e do comitê de classificação/predição de evasão com subclassificadores especializados por semestre, conforme proposto no MPC-SDP. Embora a aplicação de tais soluções tenha sido proposta em conjunto, as implementações foram realizadas isoladamente, considerando que as técnicas não precisam ser, necessariamente, associadas, e que também é desejável avaliá-las separadamente. As implementações apresentadas nesta Seção também foram disponibilizadas em um *notebook* do *Google Colab*, em: `http://www.bit.ly/mpc_sdp`.

5.3.1 Seleção de atributos orientada por aspectos

Como descrito no Capítulo 4, um dos diferenciais propostos na construção do MPC-SDP é a seleção de atributos orientada por aspectos, que visa oportunizar, de forma mais contundente, a participação de atributos relacionados a diferentes aspectos dos estudantes nos padrões preditivos. Esse método de seleção, nomeado de FSA – *Feature Selection by Aspect*, foi implementado em uma classe Python, conforme código apresentado nas Figuras 16 e 17. Para que a classe FSA pudesse ser utilizada de forma integrada/compatível com os métodos da biblioteca *scikit-learn*, sua implementação herdou, como indicado na linha 8 da Figura 16, o funcionamento e/ou métodos das classes *BaseEstimator* e *SelectorMixin*, disponibilizadas na API (*Application Programming Interface*) do *scikit-learn* (Buitinck et al., 2013). Enquanto a primeira é a classe base para todos os estimadores do *scikit-learn*[17], a segunda provê métodos de transformações a serem reutilizados pelos algoritmos de seleção de atributos, no *scikit-learn* (Scikit-Learn, 2023).

Em relação ao seu funcionamento, a partir do construtor, na linha 8 da Figura 16, é possível observar que a classe FSA requer: (i) obrigatoriamente, um método de seleção de atributos do *scikit-learn* (*estimator*), a ser utilizado para selecionar os atributos de cada aspecto, separadamente; e (ii) opcionalmente, uma lista com os descritores que identificam cada aspecto dos dados no início dos nomes dos atributos (*aspects*).

[17] Estimadores são objetos básicos do *scikit-learn*, que implementam um método "*fit*" para aprender a partir dos dados (Buitinck et al., 2013).

Figura 16 – Implementação da classe FSA

```python
import numpy as np
import pandas as pd
from sklearn.base import BaseEstimator, clone
from sklearn.feature_selection import SelectorMixin
from sklearn.utils.validation import check_is_fitted


class FSA(SelectorMixin, BaseEstimator):
    """
    Feature Selection by Aspect (FSA): Class used to select features from data,
    considering different aspects separately.

    Parameters
    ----------
    estimator: object (sklearn feature selector)
        Used to select features for each aspect.
    aspects: list of str objects, optional, default=None
        List containing the descriptors that identify each aspect, at the beginning
        of the feature names. This list can be:
        (i) passed as a parameter; or
        (ii) automatically identified, if the feature names follow the format:
        <aspect descriptor>_<feature descriptor>.

    Attributes
    -------
    estimator: object (sklearn feature selector)
        The estimator used to select features in each aspect.
    aspects: list of str objects
        The descriptors that identify each aspect in the feature names.
    n_features_in_: int
        Number of features seen during fit.
    feature_names_in_: ndarray of shape (`n_features_in_`,)
        Names of features seen during fit.
    n_features_: int
        The number of selected features.
    feature_names_: ndarray of shape (`n_features_`,)
        Names of selected features.
    support_: ndarray of shape (`n_features_in_`,)
        The mask of selected features.
    """

    def __init__(self, estimator, aspects=None):
        self.estimator = estimator
        self.aspects = aspects

    def _extract_aspects(self):
        """
        Extract the list of descriptors that identify each aspect of the data, if
        the feature names follow the following format:
            <aspect descriptor>_<feature descriptor>.

        Returns
        -------
        aspects : list of str objects
            The descriptors that identify each aspect, at the beginning of the
            feature names.
        """
        aspects = np.unique(
            [(c.split("_")[0]) for c in self.feature_names_in_ if "_" in c]
        ).tolist()
        if len(aspects) == 0:
            raise ValueError(
                "Feature names do not follow the format <attribute descriptor>_<
                    attribute descriptor>."
            )
        return aspects
```

Fonte: Elaborada pela autora.

Figura 17 – Implementação da classe FSA (Continuação)

```
67      def fit(self, X, y):
68          """
69          This method splits the data into subsets, according to the aspect/perspective
70          represented by the attributes. Then, it trains the feature selection model
71          for each subset, selecting attributes for each aspect. The selection result
72          corresponds to the union of the attributes selected for each aspect.
73
74          Parameters
75          ----------
76          X : pandas DataFrame of shape (n_samples, n_features)
77              The training input samples, with named columns.
78          y : array-like of shape (n_samples,)
79              The target values (class labels) as integers or strings.
80          Returns
81          -------
82          self : FSA
83              Fitted feature estimator.
84          """
85          if not isinstance(X, pd.DataFrame) or X.columns is None:
86              raise TypeError(
87                  "FSA requires X to be a DataFrame object, with named columns."
88              )
89          self.feature_names_in_ = X.columns.values
90          self.n_features_in_ = self.feature_names_in_.size
91          if self.aspects is None:
92              self.aspects = self._extract_aspects()
93          self.feature_names_ = np.array([])
94
95          for aspect in self.aspects:
96              X_aspect = X[
97                  [c for c in self.feature_names_in_ if c.startswith(aspect + "_")]
98              ]
99              selector = clone(self.estimator)
100             pct_features = selector.get_params()["n_features_to_select"]
101             if X_aspect.shape[1] * pct_features < 1:
102                 selector.set_params(n_features_to_select=1)
103             selector = selector.fit(X_aspect, y)
104             self.feature_names_ = np.append(
105                 self.feature_names_, selector.get_feature_names_out()
106             )
107
108         self.n_features_ = self.feature_names_.size
109         self.support_ = np.in1d(self.feature_names_in_, self.feature_names_)
110         return self
111
112     def _get_support_mask(self):
113         """
114         Get a boolean mask indicating which features are selected.
115         This implementation is required by SelectorMixin.
116
117         Returns
118         -------
119         support_ : boolean array of size `n_features_in_`
120             An element is True if its corresponding feature was selected by the model.
121         """
122         check_is_fitted(self)
123         return self.support_
```

Fonte: Elaborada pela autora.

Essa lista é necessária para possibilitar a separação dos subconjuntos de dados por aspecto, de modo que possa ser conduzido um processo de seleção de atributos em cada subconjunto, isoladamente. Caso não seja informada na instanciação, essa lista será extraída automaticamente, durante o treinamento do seletor FSA, por meio do método *_extract_aspects* (linha 46), desde que os nomes dos atributos dos dados de treinamento sigam o padrão *<descritor de aspecto>_<descritor do atributo>*.

Conforme o padrão dos estimadores do *scikit-learn*, o treinamento do seletor FSA é implementado no método *fit*, que recebe os dados de treino (*X*) e seus respectivos rótulos/classes (*y*) – linha 67 da Figura 17. Porém, na implementação da FSA, exige-se que os dados de treinamento estejam no formato/tipo *DataFrame* (biblioteca pandas)[18] e que apresentem colunas nomeadas, conforme validação da linha 85. O uso de uma tabela estruturada com colunas nomeadas, nesse caso, decorre da necessidade de identificação e separação das colunas/atributos, conforme o aspecto representado em seus nomes. Como já mencionado, no método *fit* são extraídos os descritores dos aspectos representados nos dados de treino (linha 92), caso esses descritores não tenham sido informados na instanciação da FSA, mas sigam um padrão pré-determinado. Após, para cada descritor/aspecto, são filtrados/obtidos os seus respectivos atributos e dados de treino, que são, então, utilizados para treinar o modelo de seleção de atributos do aspecto (linhas 95-103). Lembrando que o modelo de seleção de cada aspecto é criado/treinado a partir de um algoritmo de seleção do *scikit-learn*, passado por parâmetro (*estimator*) na instanciação da classe FSA. Os atributos selecionados pelo modelo de cada aspecto são, então, adicionados ao resultado da seleção FSA (*features_names_*, linhas 104-106).

Assim, ao invés de considerar apenas os atributos de maior capacidade preditiva geral, o conjunto final de atributos selecionados pela FSA contempla os atributos mais importantes de cada aspecto. Com isso, espera-se aumentar a representatividade dos diferentes aspectos dos estudantes entre os padrões de evasão, e possibilitar uma percepção mais abrangente acerca dos fatores relacionados ao abandono estudantil.

Por fim, além dos métodos *_extract_aspects* e *fit*, pode-se observar a existência do método *_get_support_mask* na classe FSA (linha 112 da Figura 17). Esse é um método abstrato da classe *SelectorMixin* do *scikit-learn*, sendo sua implementação necessária para o correto funcionamento dos métodos herdados (de transformação e transformação inversa, por exemplo). O método tem a função de retornar uma máscara (*support_*) que indique, por meio de seus valores lógicos/booleanos, se cada um dos atributos de entrada (dos dados de treinamento) foi ou não selecionado pela FSA.

[18] Os estimadores do *scikit-learn* aceitam qualquer tipo tabular/matricial que possa ser convertido em um *array* n-dimensional da biblioteca *numpy* (*ndarray*).

5.3.2 Comitê de classificação orientado por semestres

Com o intuito de considerar a variabilidade dos padrões de evasão perante diferentes períodos letivos, o MPC-SDP foi proposto, no Capítulo 4, como um modelo de *ensemble* cujo comitê de decisão é formado por subclassificadores especializados por semestres acadêmicos. A proposta do MPC-SDP foi implementada em uma classe Python, conforme código apresentado nas Figuras 18, 19, 20, e 21. A implementação da classe MPCSDP herdou o funcionamento e/ou métodos das classes *BaseEstimator* e *ClassifierMixin* da API do *scikit-learn* (Buitinck et al., 2013), como indicado na linha 9 da Figura 18. Enquanto a primeira, já utilizada na implementação da FSA (Subseção 5.3.1), é a classe base de todos os estimadores do *scikit-learn*, a segunda é o *mixin* que provê métodos a serem reutilizados pelos classificadores do *scikit-learn* (Scikit-Learn, 2023).

A partir do construtor, linhas 62-86 da Figura 19, pode-se observar que, como parâmetros, a implementação da classe MPCSDP requer, obrigatoriamente: (i) um objeto (*estimator*) que seja um classificador do *scikit-learn*, ou um *pipeline* do *scikit-learn* ou do *Imbalanced-learn* (Lemaître; Nogueira; Aridas, 2017)[19], a ser utilizado como base no treinamento de cada subclassificador (especializado por semestre) do comitê de decisão/*ensemble*; e (ii) o nome do atributo que identifica o semestre acadêmico nos registros de dados (*semester_feature*), a ser utilizado para separar os subconjuntos de dados de treino de cada subclassificador especializado por semestre. Note que, ao invés de definir parâmetros para a especificação de métodos de reamostragem/balanceamento e seleção de atributos, e montar um *pipeline* de classificação internamente, optou-se por oportunizar o recebimento desse objeto, a fim de flexibilizar/oportunizar a aplicação de outras transformações, se desejáveis ou necessárias. Opcionalmente, na instanciação da classe MPCSDP, podem ser informados: (i) uma métrica avaliativa (*metric*) e (ii) um conjunto de parâmetros (*param_grid*), a serem considerados na otimização de hiperparâmetros de cada subclassificador especializado por semestre; (iii) uma quantidade mínima de amostras por classe (*min_samples_class*), a ser considerada em cada subconjunto de treinamento; e (iv) um valor/estado de controle de

[19]Um *pipeline* permite que sejam especificadas, ordenadamente, diversas etapas de transformação (como, por exemplo, reamostragem/balanceamento, escalonamento/normalização/padronização, e seleção de atributos) a serem consideradas (treinadas e/ou aplicadas) sobre os dados, antes do treinamento ou do posterior uso de um estimador/classificador, que compõe a última etapa do *pipeline*. Ou seja, o uso de um *pipeline* garante que todas as transformações treinadas e aplicadas sobre os dados de treino e, consequentemente, consideradas no ajuste/treinamento do classificador também serão devidamente aplicadas sobre os dados de teste, antes da utilização/aplicação desse classificador (Raschka; Mirjalili, 2017). A única exceção se aplica à etapa de reamostragem/balanceamento, que nunca deve ser aplicada sobre os dados de teste. Por isso, quando essa transformação for desejável, pode-se optar por um método de reamostragem/balanceamento do pacote *Imbalanced-learn* (Lemaître; Nogueira; Aridas, 2017) e instanciar a classe *Pipeline* implementada por esse pacote, que é compatível com as funcionalidades do *scikit-learn* e garante que a reamostragem será devidamente desconsiderada na testagem/predição.

Figura 18 – Implementação da classe MPC

```
import numpy as np
import pandas as pd
from sklearn.model_selection import GridSearchCV, StratifiedKFold
from sklearn.base import BaseEstimator, ClassifierMixin, clone
from sklearn.preprocessing import LabelEncoder
from sklearn.utils.validation import check_X_y, check_is_fitted


class MPCSDP(ClassifierMixin, BaseEstimator):
    """
    Multi-Perpective Classifier for Student Dropout Prediction (MPC-SDP)

    Parameters
    ----------
    estimator: object (sklearn classifier, sklearn Pipeline, or imblearn Pipeline)
        The base estimator/classifier or a pipeline (with transforms - such as
        sampling, scaling, feature selection - and a final estimator) to be fitted
        on the subsets of data from each academic term/semester.
    semester_feature: String
        Feature name that indicates the semester/term in data samples.
        Feature used to separate the subset of training data for each subclassifier,
        but not seen during `fit`.
    param_grid: dict or list of dictionaries, optional, default=None
        param_grid to be evaluated by sklearn.model_selection.GridSearchCV when
        optimizing model hyperparameters
    metric: String, optional, default=None
        Name of the sklearn metric to be evaluated by GridSearchCV when optimizing
        model hyperparameters
    min_samples_class: int, optional, default=300
        minimum number of samples per class to be considered in the training of the
        subclassifier of each academic term/semester. If the subset of data for a
        given term/semester does not reach this number of samples in any of the
        classes, all the data from this and the next semesters are combined into a
        single dataset that will be used to fit the last subclassifier.
    random_state: int, default=None
        controls the randomness of the estimator.

    Attributes
    ----------
    estimator : object (sklearn classifier, sklearn Pipeline, or imblearn Pipeline)
        The base estimator/classifier or a pipeline (with transforms - such as
        sampling, scaling, feature selection - and a final estimator) to be fitted
        on the subsets of data from each academic term/semester.
    le_ : object (sklearn LabelEncoder)
        Transformer used to encode the labels during `fit` and decode in `predict`.
    classes_ : ndarray
        The class labels.
    semesters_ : list of int
        Academic terms/semesters existing in the training data and for which
        subclassifiers specialized per term/semester will be built.
    n_estimators_ : int
        Number of subclassifiers specialized per term/semester will be built
        (size of `semesters_` list).
    estimators_ : list of estimators
        The collection of fitted subclassifiers/pipelines.
    n_features_in_ : int
        Number of features seen during `fit`.
    feature_names_in_ : ndarray of shape (`n_features_in_`,)
        Names of features seen during `fit`.
    """
```

Fonte: Elaborada pela autora.

Figura 19 – Implementação da classe MPC (Continuação 1)

```
62      def __init__(
63          self, estimator, semester_feature, param_grid=None, metric=None,
64          min_samples_class=300, random_state=None,
65      ):
66          self.estimator = estimator
67          self.semester_feature = semester_feature
68          self.param_grid = param_grid
69          self.metric = metric
70          self.min_samples_class = min_samples_class
71          self.random_state = random_state
72          if hasattr(self.estimator, "random_state"):
73              self.estimator = self.estimator.set_params(
74                  **dict(
75                      random_state=self.random_state,
76                  )
77              )
78          # Sets the total number of processors to be used, if the estimator allows
79          # parallel processing and no hyperparameter optimization will be performed.
80          # Otherwise, parallelism will be exploited in hyperparameter optimization.
81          if hasattr(self.estimator, "n_jobs"):
82              self.estimator = self.estimator.set_params(
83                  **dict(
84                      n_jobs=(-1 if self.param_grid is None else None),
85                  )
86              )
87
88      def _get_semester_data(self, X, y, semester):
89          """
90          Get the subset of training data for a given academic term/semester. If the
91          data subset does not reach the `min_samples_class` for any of the classes,
92          samples from subsequent semesters are included in the data subset, so that
93          it is used to train a last subclassifier.
94
95          Parameters
96          ----------
97          X : pandas Dataframe of shape (n_samples, n_features)
98              The training data.
99          y : array-like of shape (n_samples,)
100             The target values.
101         semester : int
102             Semester number to get the data.
103
104         Returns
105         -------
106         X_semester : pandas Dataframe
107             The training data related to `semester`, without the `semester_feature`.
108         y_semester : array-like of shape
109             The target values related to `X_semester` samples.
110         exit : boolean
111             A True value indicates that the obtained subset of data should be used
112             to train one last subclassifier. It will not be possible to separate
113             other subsets because the `semester` training samples did not reach
114             `min_samples_class` in any of the classes.
115         """
116         df = X.copy()
117         df["y"] = y.copy()
118         df_semester = df[df[self.semester_feature] == semester]
119         sample_counts = df_semester["y"].value_counts()
120         exit = False
121         if sample_counts.lt(self.min_samples_class).any():
122             exit = True
123             self.semesters_ = [s for s in self.semesters_ if s <= semester]
124             self.n_estimators = len(self.semesters_)
125             df_semester = df[df[self.semester_feature] >= semester]
126         y_semester = df_semester["y"]
127         X_semester = df_semester.drop(columns=["y", self.semester_feature])
128         return X_semester, y_semester, exit
```

Fonte: Elaborada pela autora.

Figura 20 – Implementação da classe MPC (Continuação 2)

```
130     def fit(self, X, y):
131         """
132         Build/fit the MPC-SDP from the training set (X, y).
133         Initially, the academic semesters/terms present in the training dataset are
134         identified. For each semester/term, the related training data is obtained
135         and a subclassifier/pipeline specialized on the academic semester/term
136         is trained, considering the Pipeline's transforms (such as sampling, scaling,
137         and attribute selection) and hyperparameter optimization, if they
138         were specified in the MPCSDP instantiation.
139 
140 
141         Parameters
142         ----------
143         X : pandas Dataframe of shape (n_samples, n_features)
144             The training input samples.
145         y : array-like, shape (n_samples,)
146             The target values.
147 
148         Returns
149         -------
150         self : object
151            The fitted MPCSDP.
152         """
153         check_X_y(X, y)
154         self.le_ = LabelEncoder().fit(y)
155         self.classes_ = self.le_.classes_
156         self.semesters_ = np.unique(X[self.semester_feature]).tolist()
157         self.n_estimators = len(self.semesters_)
158         self.estimators_ = []
159         self.feature_names_in_ = X.columns
160         self.n_features_in_ = len(self.feature_names_in_)
161         if not self.feature_selector is None:
162             self.feature_selectors_ = []
163         y = self.le_.transform(y)
164 
165         for semester in self.semesters_:
166             x_semester, y_semester, exit = self._get_semester_data(X, y, semester)
167             clf = clone(self.estimator)
168             if self.param_grid is None:
169                 clf.fit(x_semester, y_semester)
170                 self.estimators_.append(clf)
171             else:
172                 cv = StratifiedKFold(
173                     n_splits=5,
174                     random_state=self.random_state,
175                     shuffle=(True if not self.random_state is None else False),
176                 )
177                 gs = GridSearchCV(
178                     estimator=clf,
179                     param_grid=self.param_grid,
180                     cv=cv,
181                     scoring=self.metric,
182                     n_jobs=-1,
183                     verbose=1,
184                 ).fit(x_semester, y_semester)
185                 self.estimators_.append(gs.best_estimator_)
186             if exit:
187                 break
188         return self
```

Fonte: Elaborada pela autora.

Figura 21 – Implementação da classe MPC (Continuação 3)

```
190     def predict_proba(self, X):
191         """
192         Predict class probabilities for X.
193         The predicted class probabilities of an input sample are computed as
194         the mean predicted class probabilities of the MPC-SDP' subclassifiers.
195 
196         Parameters
197         ----------
198         X : pandas Dataframe of shape (n_samples, n_features)
199             The input samples.
200         Returns
201         -------
202         probas : ndarray of shape (n_samples, n_classes)
203             The class probabilities.
204         """
205         check_is_fitted(self)
206         if self.semester_feature in X.columns:
207             X = X.drop(columns=[self.semester_feature])
208         probas = [est.predict_proba(X) for est in self.estimators_]
209         probas = np.mean(np.asarray(probas), axis=0, dtype=np.float64)
210         return probas
211 
212     def predict(self, X):
213         """
214         Predict class for X.
215         The predicted class of an input sample is a vote by the subclassifiers in
216         the MPC-SDP. When there is a tie in the hard vote, a soft vote is applied
217         so that the probability estimates of the classifiers are considered. In this
218         case, the predicted class is the one with the highest mean probability
219         estimate across all subclassifiers.
220 
221         Parameters
222         ----------
223         X : pandas Dataframe of shape (n_samples, n_features)
224             The input samples.
225         Returns
226         -------
227         predictions : ndarray, shape (n_samples,)
228             The predicted classes.
229         """
230         check_is_fitted(self)
231         if self.semester_feature in X.columns:
232             X = X.drop(columns=[self.semester_feature])
233         hard_votes = [est.predict(X) for est in self.estimators_]
234         hard_votes = np.asarray(hard_votes).T
235         predictions = []
236         for i in range(0, X.shape[0]):
237             hard_voting = pd.Series(hard_votes[i, :]).mode()
238             if hard_voting.size == 1:  # there was a majority vote
239                 predictions.append(hard_voting[0])
240             else:
241                 x = pd.DataFrame(X.iloc[i, :]).transpose()
242                 soft_voting = np.argmax(self.predict_proba(x))
243                 predictions.append(soft_voting)
244         predictions = self.le_.inverse_transform(predictions)
245         return predictions
```

Fonte: Elaborada pela autora.

aleatoriedade (*random_state*), a ser considerado na construção de cada subclassificador especializado por semestre.

Faz-se importante observar que não foi explorado o paralelismo entre os subclassificadores, na implementação do MPC-SDP. Optou-se por utilizar o processamento paralelo no desenvolvimento de cada subclassificador, preferencialmente durante a otimização de hiperparâmetros, quando informado, na instanciação, um conjunto de parâmetros a ser avaliado (*param_grid*). Na falta dessa informação e, consequentemente, da otimização de hiperparâmetros, explora-se o paralelismo no treinamento do modelo especialista (subclassificador), se permitido pelo algoritmo de classificação base (*estimator*), conforme linhas 78-86 da Figura 19.

Por ser um modelo de *ensemble* cujo comitê de decisão é formado por subclassificadores especializados por semestres acadêmicos, o treinamento do MPC-SDP, naturalmente, exige que os dados de treino sejam divididos conforme os períodos letivos que eles representam. Essa separação é viabilizada pelo método *_get_semester_data* (linhas 88-128 da Figura 19), que recebe, como parâmetros, o conjunto completo de treinamento (*X* e *y*), e o semestre a ter os registros separados (*semester*). A partir do conjunto de treinamento recebido, são filtrados os registros cujo valor do atributo *semester_feature* corresponda ao semestre desejado (*semester*). Caso o subconjunto de dados resultante não satisfaça o mínimo de amostras em alguma das classes (*min_samples_class*[20]), alcança-se a condição de parada da separação (*exit = True*, linha 122 da Figura 19). Nesse caso, as amostras de todos os semestres subsequentes são incluídas no subconjunto de treinamento, de modo que não seja realizada nenhuma nova separação dos dados; e, que esse subconjunto seja utilizado para treinar um último subclassificador[21]. Ao final, o método retorna o subconjunto de dados obtido (*X_semester* e *y_semester*), além do indicativo de ter ou não alcançado o critério de parada (*exit*).

Como em todo estimador do *scikit-learn*, o treinamento do classificador MPCSDP é desencadeado pelo método *fit* (linhas 130-188 da Figura 20), a partir de um conjunto de treinamento (*X, y*), passado por parâmetro. No treinamento, basicamente, é obtida uma lista com os diferentes semestres representados no atributo *semester_feature* do conjunto de dados (linha 156). Após, para cada semestre da lista (bloco das linhas 165-187), é feita a separação do seu respectivo subconjunto de dados, a partir do já descrito método *_get_semester_data* (linha 166), e, então, conduzido o treinamento do seu subclassificador, até que não seja mais possível continuar a divisão do conjunto

[20] Por padrão, este parâmetro considera um mínimo de 300 amostras (linha 64 da Figura 19).

[21] Note que as quantidades de amostras/registros das classes evadido e concluído (não evadido) diminuem ao longo dos semestres acadêmicos, já que os registros semestrais dos estudantes cessam após o semestre de evasão ou conclusão. Por isso, ao não ser atingido o mínimo de amostras por classe em um determinado semestre, interrompe-se a separação dos dados e utiliza-se, como último subconjunto de treinamento, os registros desse e dos próximos semestres.

de treinamento (critério de parada, linha 186). Quando informado, na instanciação, um conjunto de hiperparâmetros (*param_grid*) a ser considerado na construção/avaliação de modelos candidatos, o treinamento de cada subclassificador ocorre a partir do processo de otimização de hiperparâmetros (linhas 171-185), realizado por meio da classe *GridSearchCV* do *scikit-learn* e com validação cruzada de 5-*folds*. Caso contrário, cada subclassificador é construído a partir de um treinamento simples/sem otimização (linhas 168-170). Note que, após treinado, cada subclassificador especializado por semestre é adicionado à lista do comitê de classificação (*estimators_*, nas linhas 170 ou 185), que será utilizada, posteriormente, nos métodos de predição.

Faz-se importante destacar que a construção de cada um dos subclassificadores especializados por semestre considera as estratégias definidas nos parâmetros de instanciação da classe MPCSDP. Considerando as especificidades e o natural desbalanceamento entre as classes de evasão dos registros de treino de cada semestre, conforme proposta apresentada no Capítulo 4, espera-se que cada subclassificador do MPC-SDP seja treinado após os processos de balanceamento/reamostragem, seleção de atributos, e otimização de hiperparâmetros. Ou seja, na instanciação da classe MPCSDP, embora não seja obrigatório, espera-se receber, no parâmetro *estimator*, um objeto *Pipeline* do *Imbalanced-learn*, contendo os métodos de reamostragem/balanceamento, seleção de atributos, e classificação. Nesse caso, a execução em *pipeline* é fundamental, para restringir a reamostragem apenas aos dados de treinamento, quando executado o processo de otimização de hiperparâmetros (linhas 171-185). Também é importante mencionar que, se um *pipeline* for utilizado como *estimator*, a lista *estimators_* será composta pelos *pipelines* treinados para cada semestre, permitindo que suas etapas sejam aplicadas, posteriormente, nos métodos de predição da classe MPCSDP. Porém, na prática, por conterem e desempenharem o processo de classificação especializado em cada semestre, eles continuem sendo tratados/considerados como subclassificadores do comitê de decisão.

Seguindo o padrão do *scikit-learn*, na classe MPCSDP, a implementação dos processos de predição de probabilidades e de classes, dado um conjunto de dados de teste/avaliação, é realizada, respectivamente, nos métodos *predict_proba* e *predict*, apresentados na Figura 21. No primeiro (linhas 190-210), para cada registro/amostra do conjunto de dados recebido para teste/avaliação (X), são estimadas as probabilidades de classe a partir da média das estimativas de probabilidade dos subclassificadores (especializados por semestre) que compõem o comitê do MPC-SDP (linhas 208 e 209). No segundo (linhas 212-245), é realizada a predição/decisão do rótulo de classe de cada amostra do conjunto de teste/avaliação, a partir de uma votação mista entre as predições/decisões dos subclassificadores. Essa votação mista consiste em aplicar, inicialmente, uma votação pela maioria (*hard voting*, linha 237), e, caso haja empate nos rótulos, considerar as estimativas de probabilidades dos subclassificado-

res do MPC-SDP (*soft voting*, linha 242). Nesse último caso, para cada registro do conjunto de dados de teste/avaliação, é predita a classe que tiver a maior estimativa de probabilidade, considerando a média das estimativas de todos os subclassificadores (calculadas no método *predict_proba*).

Por fim, cabe destacar que, se utilizado um *pipeline* no treinamento, os métodos *predict_proba* e *predict* de cada componente do comitê de decisão (lista *estimators_*, nas linhas 208 e 233 da Figura 21) serão executados pelos *pipelines* treinados para cada semestre, e, portanto, incluirão a execução de suas etapas de transformação/-seleção de atributos, além da predição/classificação.

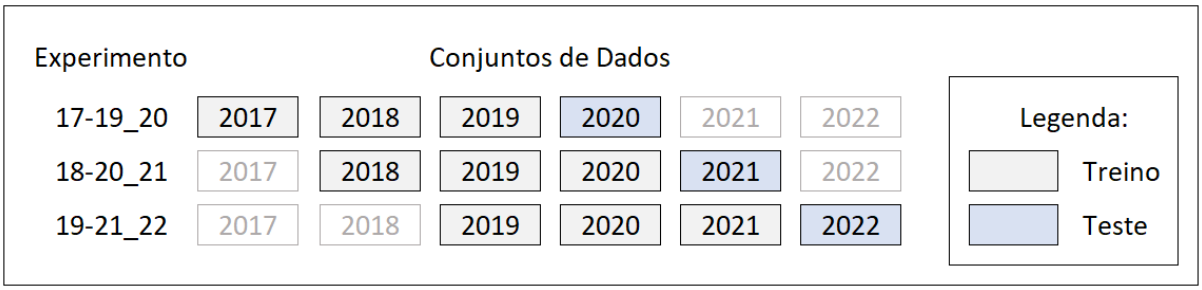
6 AVALIAÇÃO EXPERIMENTAL

A fim de analisar o potencial preditivo do MPC-SDP, foram realizadas avaliações experimentais, considerando modelos genéricos de predição tradicionais como *baselines*. Para compor o MPC-SDP, como base para o treinamento de cada subclassificador especializado por semestre do comitê de decisão, foram considerados os algoritmos de classificação DT e LR, disponíveis no *scikit-learn*. A escolha desses algoritmos partiu de suas características de simplicidade e interpretabilidade, além dos resultados da RSL apresentada no Capítulo 3, que os indicou como de ampla utilização no contexto da predição de evasão. Como *baselines*, também foram desenvolvidos modelos preditivos genéricos a partir desses algoritmos tradicionais, empregando-os individualmente ou na composição de *ensembles*. No caso do algoritmo DT, foram considerados os *ensembles* RF e GBT, que implementam, respectivamente, técnicas de *bagging* e *boosting*, especificamente com subclassificadores DT. Para o algoritmo LR também foram desenvolvidos *ensembles* tradicionais, a partir das técnicas de *bagging* e *boosting*, mas considerando os algoritmos *Bagging* e *AdaBoost*, todos disponibilizados pelo *scikit-learn*.

Para o desenvolvimento e a avaliação dos modelos preditivos, ao invés de dividir estratificadamente os conjuntos de dados de treino e teste, decidiu-se separá-los temporalmente, a fim de aproximar a avaliação da realidade de predição futura, onde dados do passado são utilizados para treinar um modelo a ser aplicado a dados de um contexto temporal futuro. Isso porque contextos ou exceções temporais (a exemplo da pandemia de COVID-19 e, consequentemente, do ERE), podem influenciar/alterar os padrões de evasão. E, dessa forma, manter uma sequência temporal na avaliação favorece a identificação de possíveis oscilações no desempenho preditivo. Como neste estudo foram utilizados dados de estudantes de cursos de graduação presenciais do IFFar, evadidos e concluídos entre os anos de 2017 a 2022 (veja Seção 5.1), foram separados seis conjuntos de dados, conforme o ano de conclusão ou abandono dos estudantes (de 2017 a 2022). A partir desses conjuntos de dados foram conduzidos três experimentos, considerando uma estratégia de validação *walk-forward*, que, para garantir uma quantidade representativa de dados, utilizou uma janela deslizante de

três conjuntos de treinamento. Ou seja, a validação avançou ao longo do tempo, atualizando sequencialmente os três conjuntos que compõem os dados de treinamento e utilizando o conjunto subsequente para teste, conforme demonstrado na Figura 22.

Figura 22 – Relação dos experimentos, e seus conjuntos de treino e teste



Fonte: Elaborada pela autora.

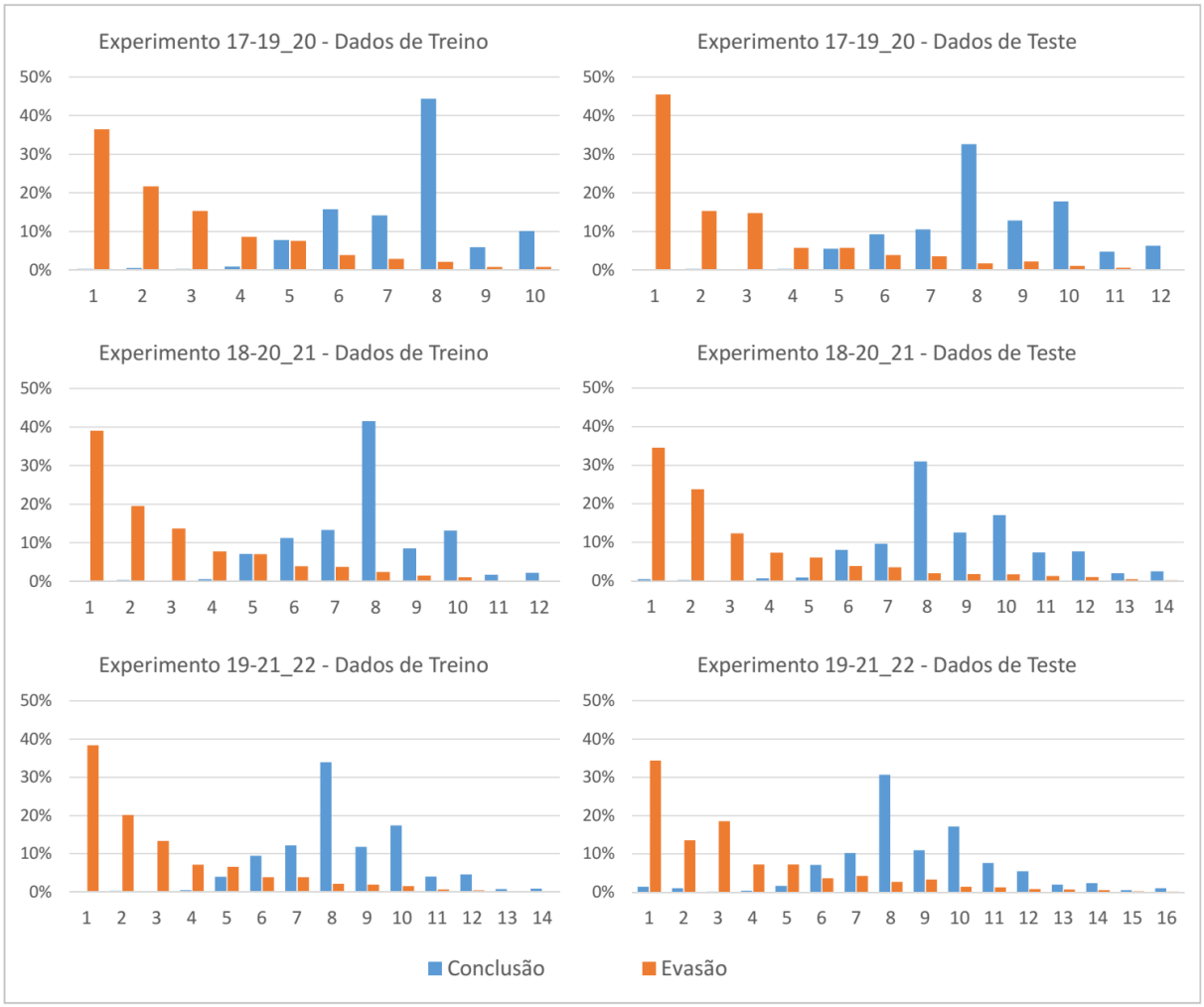
Após estabelecidos, os conjuntos de treino e teste foram pré-processados, considerando o processo descrito na Seção 5.2. Faz-se importante mencionar que, seguindo a lógica de aplicação real, o pré-processamento foi executado separadamente para os conjuntos de treino e teste de cada experimento. Após pré-processado, cada conjunto de treino foi utilizado no pré-processamento do seu respectivo conjunto de teste, servindo como referência para a adição/derivação dos atributos que posicionam os dados dos estudantes em relação aos seus pares.

Na Figura 23, são apresentadas as concentrações de estudantes concluídos e evadidos em cada semestre dos conjuntos de treino e teste de cada experimento. De forma geral, pode-se observar que, no IFFar, as evasões ocorrem, principalmente, no primeiro semestre, e apresentam progressiva redução ao longo do tempo, sendo que cerca de 70% das ocorrências se concentram nos três primeiros semestres. Já as conclusões passam a ocorrer a partir do quinto semestre[1], coincidindo com a duração de alguns cursos de tecnologia, e atingem o ápice no oitavo semestre, quando também se encerram diversos cursos de licenciatura e bacharelado.

Na Tabela 14, são apresentados os quantitativos de registros considerados nos conjuntos de treino e teste de cada experimento. Os quantitativos estão separados por classe (evasão e conclusão) e período/semestre, e acompanhados pelas proporções/percentuais que representam em cada período/semestre. Pode-se observar que há um grande desbalanceamento de registros entre as classes nos semestres iniciais, especialmente no de admissão e no primeiro (0 e 1). Porém, a classe predominante, que inicialmente é a de estudantes evadidos, se inverte após o terceiro semestre, passando a ser a de estudantes concluídos. Isso porque, como apresentado na Figura 23,

[1] Embora existam algumas ocorrências de conclusão anteriores, elas se relacionam a casos excepcionais, de alunos que tiveram aproveitamento de disciplinas/componentes curriculares (situações de reingresso, transferência, etc.).

Figura 23 – Concentrações de estudantes concluídos e evadidos por semestre, considerando os conjuntos de dados de treino e teste de cada experimento.



Fonte: Elaborada pela autora.

grande parte das evasões se concentra nos períodos iniciais, cessando a geração de novos registros semestrais. Como cada combinação de treino envolve diferentes proporções de registros relacionados a alunos evadidos e concluídos, foi considerado o balanceamento completo dos dados de treino na construção dos modelos de cada experimento. Para isso, utilizou-se a técnica de subamostragem aleatória *RandomUnderSampler* da biblioteca *Imbalanced-learn* (Lemaître; Nogueira; Aridas, 2017). Cabe mencionar que a técnica de sobreamostragem SMOTE também foi testada, considerando sua maior frequência entre os trabalhos analisados na RSL do Capítulo 3. Porém, além de mais lenta, neste estudo de caso, ela não demonstrou ganho de desempenho considerável.

Ainda em relação aos dados, para a aplicação dos classificadores LR, e *ensembles* de LR (*Bagging*, *AdaBoost*, e MPC-SDP), os dados numéricos foram escalonados/padronizados, por meio do método *StandardScaler* do *scikit-learn*. Essa técnica não foi

Tabela 14 – Quantitativos de registros envolvidos em cada experimento, considerando diferentes classes (conclusão e evasão), períodos/semestres, e conjuntos de dados (treino e teste)

Experimento	Semestre	Registros de Treino				Registros de Teste			
		Conclusão		Evasão		Conclusão		Evasão	
		Total	%	Total	%	Total	%	Total	%
17-19_20	0 e 1	762	24.1	2403	75.9	399	27.5	1053	72.5
	2	760	33.2	1527	66.8	399	41.0	574	59.0
	3	756	42.9	1006	57.1	398	49.1	413	50.9
	4	754	54.2	638	45.8	398	60.7	258	39.3
	5	747	63.4	432	36.6	397	66.7	198	33.3
	6	688	73.3	251	26.7	375	73.1	138	26.9
	7	568	78.2	158	21.8	338	77.7	97	22.3
	8	460	83.8	89	16.2	296	83.2	60	16.9
	9	122	76.3	38	23.8	166	79.8	42	20.2
	10	77	80.2	19	19.8	115	85.8	19	14.2
	11	-	-	-	-	44	84.6	8	15.4
	12	-	-	-	-	25	92.6	2	7.4
18-20_21	0 e 1	1126	28.4	2836	71.6	446	28.8	1102	71.2
	2	1125	39.4	1728	60.6	444	38.1	721	61.9
	3	1121	48.8	1175	51.2	443	49.1	459	50.9
	4	1119	58.7	786	41.3	443	57.8	323	42.2
	5	1113	66.3	566	33.7	440	64.5	242	35.5
	6	1033	73.8	367	26.2	436	71.4	175	28.6
	7	906	78.0	255	22.0	400	75.2	132	24.8
	8	756	83.5	149	16.5	357	79.3	93	20.7
	9	288	78.3	80	21.7	219	75.5	71	24.5
	10	192	83.5	38	16.5	163	76.2	51	23.8
	11	44	84.6	8	15.4	87	73.1	32	26.9
	12	25	92.6	2	7.4	54	75.0	18	25.0
	13	-	-	-	-	20	74.1	7	25.9
	14	-	-	-	-	11	84.6	2	15.4
19-21_22	0 e 1	1288	28.7	3201	71.3	548	30.3	1263	69.7
	2	1286	39.5	1972	60.5	540	39.4	829	60.6
	3	1282	49.1	1328	50.9	534	44.8	658	55.2
	4	1280	58.7	901	41.3	533	55.7	424	44.3
	5	1274	65.4	674	34.6	531	61.5	332	38.5
	6	1223	72.5	464	27.5	522	68.4	241	31.6
	7	1101	76.4	341	23.7	483	71.2	195	28.8
	8	944	81.2	219	18.8	427	75.2	141	24.8
	9	507	77.1	151	23.0	259	70.8	107	29.2
	10	355	80.0	89	20.1	199	75.4	65	24.6
	11	131	76.6	40	23.4	105	69.1	47	30.9
	12	79	79.8	20	20.2	63	67.0	31	33.0
	13	20	74.1	7	25.9	33	61.1	21	38.9
	14	11	84.6	2	15.4	22	64.7	12	35.3
	15	-	-	-	-	9	64.3	5	35.7
	16	-	-	-	-	6	75.0	2	25.0

Fonte: Elaborada pela autora.

considerada para algoritmos baseados em DTs pois eles são invariantes às escalas dos atributos (Raschka; Mirjalili, 2017), e, além de não trazer benefício ao desempenho, essa transformação impediria uma interpretação direta das regras de decisão, já que os dados deixariam de apresentar escalas reais. Além disso, considerando que a multicolinearidade afeta direta e negativamente a estimativa dos coeficientes de regressão, impedindo a real distinção/interpretação do impacto de cada atributo sobre a variável alvo/classe[2] (Sundus et al., 2022), para modelos baseados em LR, também foi realizada a remoção de atributos multicolineares[3], antes da aplicação dos métodos de seleção de atributos avaliados nos experimentos[4]. Os atributos multicolineares foram identificados, iterativamente, a partir de uma busca gulosa, baseada no fator de inflação de variância (*Variation Inflation Factor* – VIF) de cada atributo. A busca foi encerrada após nenhum atributo apresentar VIF igual ou superior a 5. Esse é o limite determinante de multicolinearidade recomendado na documentação do pacote *statsmodels* (statsmodels, 2023), que provê a implementação do cálculo de VIF utilizado neste trabalho.

Para compor e também servir como *baseline* para a seleção de atributos FSA, foram considerados os métodos *Select Percentile* (SP) e *Recursive Feature Elimination* (RFE), disponíveis no *scikit-learn*. Enquanto o primeiro filtra/seleciona uma certa quantidade (percentil) de atributos a partir de testes estatísticos uni-variados, o segundo é um *wrapper* que usa um algoritmo de aprendizado de máquina para definir a importância dos atributos de treinamento e remover, recursivamente, os menos importantes (Scikit-Learn, 2023). Ambos os métodos foram configurados para selecionar 20% dos atributos. Na seleção SP, foi utilizada informação mútua (*mutual_info_classif*) para medir a dependência entre cada atributo e a classe, e, na RFE, foi utilizado um classificador DT, com configuração padrão, para definir a importância dos atributos em cada etapa da eliminação. Esses métodos foram escolhidos por, além de rápidos, não serem paramétricos[5], e, dessa forma, não terem tendência a beneficiar um determinado tipo de classificador.

Com o intuito de evitar o *overfitting*, os modelos também tiveram seus hiperparâmetros otimizados, por meio do método *GridSearchCV*, do *scikit-learn*, com validação

[2]Além de um grande erro padrão dos coeficientes, o sinal atribuído a um coeficiente pode ser o oposto do que seu atributo/variável realmente representa/significa, gerando uma explicação/interpretação enganosa (Chan et al., 2022).

[3]Note que, como foram derivados diversos atributos (média, desvio da média, porcentagem, diferença, etc.) de uma mesma informação, claramente existem atributos multicolineares nos conjuntos de dados.

[4]Observe que, em geral, os métodos de seleção tradicionais selecionam os atributos com base na influência que eles exercem sobre a variável alvo/classe ou sobre o resultado preditivo dos modelos. Ou seja, eles não avaliam as relações entre as variáveis preditoras (atributos) e, dessa forma, não tratam o problema da multicolinearidade diretamente.

[5]Enquanto *mutual_info_classif* se baseia em cálculos de entropia, modelos DT, na configuração padrão do *scikit-learn*, se baseiam em cálculos da impureza Gini para estabelecer a estrutura da árvore de decisão (Scikit-Learn, 2023).

cruzada de 5-*folds*. Na Tabela 15, são apresentados os valores de hiperparâmetros avaliados para os modelos, com destaque aos valores-padrão dos algoritmos do *scikit-learn*. No processo de otimização, foi considerada a métrica UAR, que corresponde a *"recall_macro"*, no *scikit-learn*, e, como visto na RSL do Capítulo 3, foi considerada por Barros et al. (2019) adequada para o contexto de dados desbalanceados. Essa métrica foi utilizada com o intuito de penalizar grandes desequilíbrios entre as taxas de verdadeiros positivos (TVP, sensibilidade, ou revocação da classe de interesse/positiva para evasão) e de verdadeiros negativos (TVN, especificidade, ou revocação da classe negativa). Isso porque, embora seja desejável recuperar o máximo possível de estudantes em risco de evasão (TVP alta), uma baixa TVN implica em um excesso de falsos positivos, prejudicando a precisão das predições, que também é crucial para a otimização dos recursos institucionais, no direcionamento de políticas preventivas.

Tabela 15 – Valores considerados no processo de otimização de hiperparâmetros dos modelos.

Modelos baseados em DT	**Hiperparâmetros**	**Valores Considerados**
DT, RF, GBT, MPC_DT	max_depth	**None**, 3, 4, 5, 7, 10
DT, RF, GBT, MPC_DT	min_samples_leaf	**1**, 0.01, 0.025, 0.05
DT, RF, MPC_DT	criterion	**gini**, entropy, log_loss
RF, GBT	n_estimators	**100**, 50, 10
RF	bootstrap	**true**, false
GBT	n_iter_no_change	**None**, 2
GBT	learning_rate	1.0, **0.1**, 0.05, 0.01
Modelos baseados em LR	**Hiperparâmetros**	**Valores Considerados**
LR, BAG_LR, ADA_LR, MPC_LR	solver (LR)	**lbfgs**, liblinear
LR, BAG_LR, ADA_LR, MPC_LR	C (LR)	0.01, 0.1, **1.0**, 10
LR, BAG_LR, ADA_LR, MPC_LR	penalty (LR)	l1, **l2**
BAG_LR	bootstrap	**true**, false
BAG_LR	n_estimators	100, 50, **10**
ADA_LR	n_estimators	100, **50**, 10
ADA_LR	learning_rate	**1.0**, 0.1, 0.05, 0.01

Fonte: Elaborada pela autora.

Faz-se importante destacar que, para restringir o balanceamento apenas aos dados de treino, a aplicação das técnicas de escalonamento (nos modelos baseados em LR), reamostragem/balanceamento, seleção de atributos, e classificação foram executadas em *Pipeline* (do pacote *Imbalanced-learn*), na construção de cada modelo candidato do processo de otimização de hiperparâmetros (*GridSearchCV*). Cabe também observar que, ao contrário dos classificadores tradicionais/*baselines*, no MPC-SDP o *Pipeline* foi passado por parâmetro e sua aplicação, assim como a otimização de hiperparâmetros, foi realizada internamente, considerando o fluxo de construção de cada subclassificador especializado por semestre, conforme detalhado na Seção 5.3.2.

Como resultado do processo de otimização de hiperparâmetros, foi obtido o mo-

delo preditivo otimizado para cada combinação de experimento e técnica considerada (seleção de atributos e algoritmo de classificação). Posteriormente, considerando que os percentuais de evasão costumam ser maiores em períodos iniciais da trajetória acadêmica dos estudantes, a avaliação dos modelos foi conduzida de modo a observar seus desempenhos preditivos frente a diferentes semestres letivos. Ou seja, na avaliação, cada conjunto de dados de teste foi subdividido conforme os semestres representados nos registros (atributo 32 da Tabela 10)[6], de modo que cada modelo pudesse ser avaliado sobre cada subconjunto de teste/período letivo, separadamente. Nessa avaliação, foram consideradas as métricas padrão de precisão, revocação, acurácia, UAR, e F1 (com média "macro"). Porém, como os conjuntos de teste refletem o desbalanceamento natural entre as classes de estudantes evadidos e formados/concluídos por semestre, optou-se por também avaliar a precisão e a revocação da classe negativa de evasão, a fim de favorecer a observação de possíveis discrepâncias de desempenho entre as classes. Adicionalmente, foram geradas curvas PR, para verificar o comportamento dos modelos em diferentes limiares de decisão.

Por fim, também foram mapeados os principais padrões/regras de evasão que compõem os subclassificadores especializados por semestre dos modelos MPC-SDP, a fim de verificar se a seleção FSA oportunizou uma maior variabilidade, entre os padrões preditivos, dos aspectos acadêmico, social, contextual, econômico, e interacional. Faz-se necessário mencionar que, para os modelos baseados em DT, quando possível, foi realizada a simplificação das regras, para facilitar a interpretação, da seguinte forma: quando todos os ramos de uma subárvore levavam a folhas com a mesma decisão/classe, as condições relacionadas a esses ramos foram podadas/-suprimidas, permitindo a unificação de várias regras específicas em uma mais abrangente. Assim, a regra resultante deixa de relatar o caminho para um nó folha e relata o caminho para uma subárvore cujos nós folhas compartilham a mesma decisão/classe. Além disso, durante a extração das regras dos subclassificadores de DT, foram desconsiderados os padrões de evasão que (i) abrangessem menos de 0.5% das amostras de treinamento ou (ii) apresentassem probabilidade (proporção de amostras de estudantes evadidos que confirmam a regra, no nó decisório) inferior a 60%.

[6]Embora tenham sido utilizados na segmentação dos dados, os atributos que identificam/distinguem o período/semestre e o estudante nos registros (atributos 32 e 1, respectivamente, na Tabela 10) não foram utilizados/considerados no treinamento dos modelos (nem nos subclassificadores do MPC-SDP, nem nos modelos tradicionais). O segundo por não ser, efetivamente, um atributo preditivo, e o primeiro por poder impactar negativamente na generalização (a quantidade padrão de semestres varia entre os diferentes cursos/tipos de curso).

7 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados e discutidos os resultados da avaliação experimental descrita no Capítulo 6. Nas Seções 7.1 e 7.2, os desempenhos e padrões preditivos dos modelos são apresentados, separadamente. E, em cada Seção, os resultados estão organizados/agrupados segundo os algoritmos de classificação utilizados pelos modelos. Nas Seções 7.1.1 e 7.1.2 são apresentados e discutidos os desempenhos preditivos dos modelos baseados em DT e LR, respectivamente, enquanto os padrões preditivos desses mesmos grupos de modelos são analisados e descritos nas Seções 7.2.1 e 7.2.2.

7.1 Desempenho Preditivo

7.1.1 Modelos baseados em DT

Nas Tabelas 16, 17 e 18 são apresentados os resultados da avaliação dos classificadores baseados em DT, desenvolvidos nos experimentos 17-19_20, 18-20_21, e 19-21_22, respectivamente. Note que as tabelas estão divididas em oito blocos, separando os resultados de avaliação dos modelos preditivos para os dados de teste dos semestres de 0 (admissão) a 7. Embora existam outros semestres entre os dados de teste, optou-se por avaliar os resultados nos dados do momento de admissão e dos sete semestres subsequentes, considerando que esses semestres concentram a maior quantidade dos registros e são suficientes para demonstrar o comportamento preditivo dos modelos, ao longo do tempo. Em cada bloco, as colunas representam as métricas avaliadas (precisão e revocação das classes positiva e negativa de evasão, acurácia, UAR, e F1), e as linhas representam os resultados dos modelos desenvolvidos, considerando cada combinação dos algoritmos de classificação e seleção de atributos. Além disso, note que os resultados estão apresentados em forma de mapa de calor. Ou seja, quanto mais fria a cor (mais próxima de verde), maior o valor e, consequentemente, melhor o resultado.

Tabela 16 – Resultados dos modelos baseados em DT no experimento 17-19_20, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	75.13	65.96	96.96	15.50	74.54	56.23	54.88	81.86	71.21	92.97	45.75	79.97	69.36	71.38	76.52	78.03	87.98	61.25	77.00	74.61	75.24	73.57	83.28	86.92	67.59	77.44	77.26	77.15
DT	SP	72.47	0.00	100.0	0.00	72.47	50.00	42.02	84.42	63.54	87.46	57.50	79.22	72.48	73.14	85.45	73.52	79.79	80.50	80.08	80.15	79.69	81.13	82.17	83.29	79.90	81.63	81.60	81.61
DT	FSASP	82.74	55.81	83.76	54.00	75.57	68.88	69.07	85.57	72.12	91.26	59.50	82.52	75.38	76.77	75.27	73.99	85.37	59.75	74.85	72.56	73.06	72.28	78.36	82.08	67.34	74.85	74.71	74.65
DT	RFE	76.83	61.08	93.83	25.50	75.02	59.66	60.23	82.96	76.56	94.30	49.00	81.83	71.65	74.01	74.18	79.42	90.07	55.00	75.67	72.53	73.17	78.90	81.60	83.29	76.88	80.15	80.09	80.10
DT	FSARFE	79.21	65.90	92.97	35.75	77.22	64.36	65.95	78.91	78.53	96.68	32.00	78.87	64.34	66.18	75.25	83.39	92.16	56.50	77.52	74.33	75.11	70.62	83.16	87.89	62.06	75.22	74.98	74.70
RF	None	78.60	74.40	95.92	31.25	78.11	63.58	65.21	80.41	81.28	96.68	38.00	80.52	67.34	69.79	77.91	82.89	90.94	63.00	79.47	76.97	77.76	79.29	84.25	86.20	76.63	81.50	81.42	81.43
RF	SP	72.47	0.00	100.0	0.00	72.47	50.00	42.02	83.50	63.91	88.41	54.00	78.94	71.21	72.21	82.72	75.06	82.58	75.25	79.57	78.91	78.90	80.94	82.12	83.29	79.65	81.50	81.47	81.48
RF	FSASP	73.00	91.67	99.91	2.75	73.16	51.33	44.85	81.14	79.71	96.01	41.25	80.94	68.63	71.16	74.35	79.29	89.90	55.50	75.77	72.70	73.34	77.65	81.25	83.29	75.13	79.28	79.21	79.22
RF	RFE	78.61	51.34	86.23	38.25	73.02	62.24	63.04	81.88	81.11	96.11	44.00	81.76	70.05	72.74	78.17	85.14	92.33	63.00	80.29	77.67	78.54	77.04	84.35	86.92	73.12	80.15	80.02	80.01
RF	FSARFE	78.41	58.30	90.69	34.25	75.15	62.47	63.63	81.76	84.73	97.06	43.00	82.17	70.03	72.90	76.58	85.25	92.86	59.25	79.06	76.05	76.92	76.41	85.84	88.62	71.61	80.27	80.11	80.07
GBT	None	78.21	67.58	94.40	30.75	76.88	62.57	63.91	80.16	82.58	97.06	36.75	80.45	66.90	69.33	70.91	80.26	92.16	45.75	73.10	68.96	69.22	72.53	84.92	88.86	65.08	77.19	76.97	76.78
GBT	SP	72.47	0.00	100.0	0.00	72.47	50.00	42.02	83.26	66.04	89.74	52.50	79.49	71.12	72.44	78.61	77.26	86.41	66.25	78.13	76.33	76.83	78.22	83.10	85.23	75.38	80.39	80.30	80.31
GBT	FSASP	78.23	62.02	92.50	32.25	75.91	62.37	63.60	80.83	81.12	96.49	39.75	80.87	68.12	70.66	71.23	77.87	90.59	47.50	72.90	69.05	69.38	75.57	84.64	87.65	70.60	79.28	79.13	79.08
GBT	RFE	78.64	54.21	88.13	37.00	74.05	62.56	63.55	81.88	79.73	95.73	44.25	81.56	69.99	72.59	75.50	84.70	92.86	56.75	78.03	74.80	75.62	74.15	86.22	89.59	67.59	78.79	78.59	78.46
GBT	FSARFE	79.02	56.09	88.70	38.00	74.74	63.35	64.44	80.99	80.60	96.30	40.50	80.94	68.40	70.95	73.15	83.61	93.03	51.00	75.77	72.02	72.63	72.46	85.95	89.83	64.57	77.44	77.20	76.98
MPCDT	None	72.47	0.00	100.0	0.00	72.47	50.00	42.02	94.37	58.57	76.35	88.00	79.56	82.18	77.37	86.15	73.15	79.09	81.75	80.18	80.42	79.84	78.91	82.43	84.26	76.63	80.52	80.45	80.46
MPCDT	SP	87.65	53.96	76.83	71.50	75.36	74.16	71.69	95.40	59.67	76.83	90.25	80.52	83.54	78.48	85.66	72.06	78.05	81.25	79.36	79.65	79.03	82.18	84.70	85.96	80.65	83.35	83.30	83.32
MPCDT	FSASP	72.47	0.00	100.0	0.00	72.47	50.00	42.02	93.99	57.78	75.78	87.25	78.94	81.52	76.72	85.14	71.71	77.87	80.50	78.95	79.19	78.60	80.69	83.51	84.99	78.89	82.00	81.94	81.96
MPCDT	RFE	72.47	0.00	100.0	0.00	72.47	50.00	42.02	91.51	65.30	83.95	79.50	82.73	81.73	79.64	85.93	73.70	79.79	81.25	80.39	80.52	80.02	82.78	82.95	83.78	81.91	82.86	82.84	82.85
MPCDT	FSARFE	72.47	0.00	100.0	0.00	72.47	50.00	42.02	93.33	62.89	81.01	84.75	82.04	82.88	79.47	85.44	72.58	78.75	80.75	79.57	79.75	79.20	82.19	82.82	83.78	81.16	82.49	82.47	82.48

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	69.70	85.79	80.23	77.39	78.51	78.81	77.98	71.05	90.19	81.82	83.38	82.86	82.60	81.35	68.86	93.35	83.33	86.13	85.38	84.73	82.50	64.34	95.42	85.57	86.39	86.21	85.98	82.07
DT	SP	81.43	84.49	74.81	88.94	83.38	81.88	82.32	85.63	88.36	75.25	93.70	87.56	84.48	85.53	80.00	91.12	75.36	93.07	88.30	84.21	84.85	78.95	93.53	77.32	94.08	90.34	85.70	85.97
DT	FSASP	77.39	85.82	78.29	85.18	82.47	81.74	81.67	80.95	88.92	77.27	90.93	86.39	84.10	84.49	77.86	90.58	73.91	92.27	87.33	83.09	83.62	78.49	92.98	75.26	94.08	89.89	84.67	85.19
DT	RFE	73.55	85.53	78.68	81.66	80.49	80.17	79.79	75.47	90.08	80.81	86.90	84.87	83.85	83.26	70.55	90.46	74.64	88.53	84.80	81.59	81.01	74.53	94.53	81.44	92.01	89.66	86.73	85.54
DT	FSARFE	71.23	86.26	80.62	78.89	79.57	79.76	79.03	73.17	87.69	75.76	86.15	82.69	80.95	80.68	72.06	89.39	71.01	89.87	84.80	80.44	80.58	80.00	92.75	74.23	94.67	90.11	84.45	85.35
RF	None	79.77	86.72	79.46	86.93	83.99	83.20	83.22	80.79	91.33	82.83	90.18	87.73	86.50	86.27	75.51	92.62	80.43	90.40	87.72	85.42	84.70	69.17	95.56	85.57	89.05	88.28	87.31	84.34
RF	SP	80.82	85.40	76.74	88.19	83.69	82.47	82.75	81.35	89.80	79.29	90.93	87.06	85.11	85.33	78.42	92.25	78.99	92.00	88.50	85.49	85.41	75.96	94.56	81.44	92.60	90.11	87.02	86.09
RF	FSASP	78.46	86.36	79.07	85.93	83.23	82.50	82.46	78.92	90.54	81.31	89.17	86.55	85.24	84.97	73.97	91.83	78.26	89.87	86.74	84.06	83.45	72.07	94.75	82.47	90.83	88.97	86.65	84.84
RF	RFE	78.49	87.21	80.62	85.68	83.69	83.15	82.99	77.83	91.38	83.33	88.16	86.55	85.75	85.12	73.65	92.05	78.99	89.60	86.74	84.29	83.52	75.00	95.98	86.60	91.72	90.57	89.16	87.09
RF	FSARFE	76.28	87.17	81.01	83.67	82.62	82.34	81.98	76.42	90.60	81.82	87.41	85.55	84.61	84.00	74.32	92.33	79.71	89.87	87.13	84.79	84.00	74.31	95.09	83.51	91.72	89.89	87.61	86.01
GBT	None	68.21	88.86	85.66	74.12	78.66	79.89	78.38	71.37	92.66	86.87	82.62	84.03	84.74	82.86	68.42	93.86	84.78	85.60	85.38	85.19	82.63	64.62	95.74	86.60	86.39	86.44	86.49	82.42
GBT	SP	77.41	87.31	81.01	84.67	83.23	82.84	82.57	80.00	90.38	80.81	89.92	86.89	85.37	85.28	76.39	92.41	79.71	90.93	87.91	85.32	84.84	78.79	94.35	80.41	93.79	90.80	87.10	86.83
GBT	FSASP	71.72	87.47	82.56	78.89	80.34	80.73	79.86	74.34	91.87	84.85	85.39	85.21	85.12	83.88	70.19	92.90	81.88	87.20	85.77	84.54	82.77	74.07	94.80	82.47	91.72	89.66	87.10	85.64
GBT	RFE	71.91	87.96	83.33	78.89	80.64	81.11	80.19	74.22	91.62	84.34	85.39	85.04	84.87	83.68	71.79	92.72	81.16	88.27	86.35	84.71	83.31	66.41	96.09	87.63	87.28	87.36	87.45	83.51
GBT	FSARFE	68.67	87.94	84.11	75.13	78.66	79.62	78.32	71.06	91.39	84.34	82.87	83.36	83.61	82.03	68.67	93.08	82.61	86.13	85.19	84.37	82.24	69.83	94.98	83.51	89.64	88.28	86.58	84.15
MPCDT	None	68.57	87.68	83.72	75.13	78.51	79.42	78.16	67.95	93.45	88.89	79.09	82.35	83.99	81.35	54.22	94.44	88.41	72.53	76.80	80.47	74.63	53.80	98.11	94.85	76.63	80.69	85.74	77.35
MPCDT	SP	72.30	87.78	82.95	79.40	80.79	81.17	80.32	64.96	93.77	89.90	75.82	80.50	82.86	79.63	52.36	94.29	88.41	70.40	75.24	79.40	73.19	52.57	98.08	94.85	75.44	79.77	85.14	76.47
MPCDT	FSASP	73.04	87.88	82.95	80.15	81.25	81.55	80.76	71.37	93.95	89.39	82.12	84.54	85.75	83.50	61.11	94.60	87.68	79.47	81.68	83.57	79.20	58.97	98.21	94.85	81.07	84.14	87.96	80.77
MPCDT	RFE	74.65	87.63	82.17	81.91	82.01	82.04	81.45	75.78	92.20	85.35	86.40	86.05	85.88	84.75	60.71	94.01	86.23	79.47	81.29	82.85	78.69	61.49	97.91	93.81	83.14	85.52	88.48	82.10
MPCDT	FSARFE	76.53	87.86	82.17	83.67	83.08	82.92	82.48	71.61	91.92	85.35	83.12	83.87	84.24	82.59	60.20	94.55	87.68	78.67	81.09	83.17	78.63	55.09	98.13	94.85	77.81	81.61	86.33	78.25

Fonte: Elaborada pela autora.

Tabela 17 – Resultados dos modelos baseados em DT no experimento 18-20_21, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	81.53	73.47	92.92	48.21	80.00	70.57	72.54	78.87	81.00	96.55	36.24	79.15	66.40	68.45	78.72	74.26	86.69	62.11	77.29	74.40	75.08	72.74	80.65	84.31	67.42	76.00	75.86	75.77
DT	SP	71.10	0.00	100.0	0.00	71.10	50.00	41.55	80.77	51.52	79.67	53.24	72.05	66.46	66.29	77.48	66.03	80.17	62.33	73.35	71.25	71.46	75.41	78.64	80.83	72.81	76.88	76.82	76.82
DT	FSASP	76.10	55.02	89.84	30.58	72.71	60.21	60.85	80.35	77.64	95.01	42.73	79.92	68.87	71.10	76.96	79.81	91.26	55.83	77.72	73.55	74.60	77.11	79.71	81.48	75.06	78.32	78.27	78.28
DT	RFE	78.20	61.82	90.47	37.95	75.29	64.21	65.46	81.85	76.95	94.10	48.55	80.96	71.32	73.54	74.61	82.76	93.76	48.43	76.44	71.09	72.10	66.29	83.63	89.98	52.81	71.68	71.39	70.54
DT	FSARFE	77.67	60.22	90.29	36.16	74.65	63.23	64.35	82.23	76.19	93.65	50.11	81.08	71.88	74.01	75.00	82.77	93.62	49.55	76.78	71.59	72.64	66.24	82.81	89.32	53.03	71.46	71.18	70.36
RF	None	76.66	68.23	94.46	29.24	75.61	61.85	62.79	79.34	81.73	96.55	38.03	79.66	67.29	69.51	79.10	79.65	90.29	61.43	79.26	75.86	76.85	74.14	81.75	84.97	69.44	77.32	77.20	77.14
RF	SP	71.10	0.00	100.0	0.00	71.10	50.00	41.55	80.89	51.17	79.13	53.91	71.85	66.52	66.25	83.05	71.02	81.55	73.09	78.32	77.32	77.17	79.52	78.88	79.52	78.88	79.20	79.20	79.20
RF	FSASP	76.61	52.56	87.39	34.38	72.06	60.88	61.61	81.03	77.19	94.56	45.41	80.37	69.98	72.23	75.40	81.14	92.65	51.12	76.78	71.89	72.93	70.94	82.22	86.71	63.37	75.22	75.04	74.81
RF	RFE	76.72	54.06	88.20	34.15	72.58	61.18	61.96	81.98	77.11	94.10	48.99	81.08	71.55	73.77	77.53	78.83	90.43	57.62	77.89	74.03	75.03	71.96	83.72	87.80	64.72	76.44	76.26	76.05
RF	FSARFE	76.19	51.40	87.39	32.81	71.61	60.10	60.73	82.19	76.63	93.83	49.89	81.15	71.86	74.03	77.31	80.77	91.68	56.50	78.23	74.09	75.19	73.94	84.40	87.80	68.09	78.10	77.94	77.83
GBT	None	77.17	62.71	92.01	33.04	74.97	62.53	63.61	79.56	78.92	95.74	39.37	79.47	67.55	69.72	72.83	82.76	94.45	43.05	74.81	68.75	69.44	64.97	85.16	91.72	48.99	70.69	70.35	69.13
GBT	SP	71.10	0.00	100.0	0.00	71.10	50.00	41.55	82.29	51.57	77.59	58.84	72.18	68.21	67.42	77.34	70.82	84.74	59.87	75.24	72.30	72.88	73.94	80.31	83.44	69.66	76.66	76.55	76.51
GBT	FSASP	75.98	58.11	91.56	28.79	73.42	60.18	60.78	81.22	77.44	94.56	46.09	80.57	70.32	72.58	74.08	84.49	94.73	46.41	76.26	70.57	71.53	70.07	85.13	89.76	60.45	75.33	75.10	74.70
GBT	RFE	76.83	60.50	91.47	32.14	74.32	61.81	62.75	81.75	77.34	94.28	48.10	80.96	71.19	73.44	75.94	81.03	92.37	52.69	77.21	72.53	73.61	69.78	84.76	89.54	60.00	75.00	74.77	74.35
GBT	FSARFE	76.82	51.27	86.03	36.16	71.61	61.09	61.79	82.08	76.74	93.92	49.44	81.08	71.68	73.87	70.61	51.16	67.96	54.26	62.72	61.11	60.96	62.63	63.07	66.45	59.10	62.83	62.77	62.75
MPCDT	None	71.10	0.00	100.0	0.00	71.10	50.00	41.55	89.50	69.50	86.66	74.94	83.28	80.80	80.09	85.07	69.62	79.06	77.58	78.49	78.32	77.67	77.11	81.53	83.66	74.38	79.09	79.02	79.02
MPCDT	SP	71.10	0.00	100.0	0.00	71.10	50.00	41.55	87.10	64.51	84.57	69.13	80.12	76.85	76.28	86.22	69.26	78.09	79.82	78.75	78.95	78.06	76.44	81.70	84.10	73.26	78.76	78.68	78.67
MPCDT	FSASP	71.10	0.00	100.0	0.00	71.10	50.00	41.55	87.87	71.56	88.75	69.80	83.28	79.27	79.49	84.65	70.60	80.31	76.46	78.83	78.38	77.92	75.65	80.15	82.57	72.58	77.65	77.58	77.57
MPCDT	RFE	71.10	0.00	100.0	0.00	71.10	50.00	41.55	88.23	66.12	85.03	72.04	81.28	78.53	77.77	83.87	69.28	79.33	75.34	77.81	77.34	76.86	74.71	80.77	83.66	70.79	77.32	77.22	77.19
MPCDT	FSARFE	71.10	0.00	100.0	0.00	71.10	50.00	41.55	85.02	71.39	90.11	60.85	81.67	75.48	76.59	84.02	69.48	79.47	75.56	77.98	77.52	77.04	74.32	80.26	83.22	70.34	76.88	76.78	76.75

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	71.39	85.79	82.66	75.96	78.78	79.31	78.59	74.32	88.03	78.93	85.03	82.87	81.98	81.53	75.90	89.01	72.00	90.85	85.46	81.42	81.91	75.35	93.61	81.06	91.27	88.74	86.17	85.26
DT	SP	78.55	83.59	77.09	84.72	81.51	80.90	80.98	78.10	87.98	78.10	87.98	84.48	83.04	83.04	75.90	89.01	72.00	90.85	85.46	81.42	81.91	71.63	92.09	76.52	90.02	86.68	83.27	82.52
DT	FSASP	83.39	84.58	77.71	88.76	84.11	83.24	83.54	78.09	89.35	80.99	87.53	85.21	84.26	83.97	75.58	89.77	74.29	90.39	85.78	82.34	82.50	72.00	93.73	81.82	89.53	87.62	85.67	84.09
DT	RFE	71.28	85.97	82.97	75.73	78.78	79.35	78.60	78.08	90.78	83.88	87.07	85.94	85.48	84.88	76.02	89.80	74.29	90.62	85.95	82.45	82.67	72.79	93.52	81.06	90.02	87.80	85.54	84.22
DT	FSARFE	72.90	86.47	83.28	77.53	79.95	80.40	79.75	80.31	91.14	84.30	88.66	87.12	86.48	86.07	79.01	89.56	73.14	92.22	86.76	82.68	83.42	75.18	92.68	78.03	91.52	88.18	84.78	84.34
RF	None	78.26	87.47	83.59	83.15	83.33	83.37	83.05	78.38	90.80	83.88	87.30	86.09	85.59	85.03	77.71	91.08	77.71	91.08	87.25	84.39	84.39	75.69	94.09	82.58	91.27	89.12	86.92	85.82
RF	SP	79.45	85.52	80.19	84.94	82.94	82.56	82.52	80.58	89.34	80.58	89.34	86.24	84.96	84.96	78.82	90.72	76.57	91.76	87.42	84.17	84.46	78.68	93.70	81.06	92.77	89.87	86.91	86.54
RF	FSASP	76.44	86.43	82.35	81.57	81.90	81.96	81.61	78.80	89.61	81.40	87.98	85.65	84.69	84.43	79.88	90.18	74.86	92.45	87.42	83.65	84.29	75.00	92.44	77.27	91.52	87.99	84.40	84.05
RF	RFE	75.21	85.85	81.73	80.45	80.99	81.09	80.70	78.16	91.00	84.30	87.07	86.09	85.69	85.05	78.70	90.52	76.00	91.76	87.25	83.88	84.23	78.52	93.47	80.30	92.77	89.68	86.54	86.26
RF	FSARFE	77.12	87.92	84.52	81.80	82.94	83.16	82.70	78.87	92.11	86.36	87.30	86.97	86.83	86.04	76.97	91.24	78.29	90.62	87.09	84.45	84.28	77.70	93.91	81.82	92.27	89.68	87.04	86.39
GBT	None	63.52	89.14	89.47	62.70	73.96	76.09	73.95	67.41	92.10	88.02	76.64	80.67	82.33	80.00	66.51	91.50	80.57	83.75	82.84	82.16	80.16	68.45	95.34	87.12	86.78	86.87	86.95	83.76
GBT	SP	75.56	86.89	83.28	80.45	81.64	81.87	81.39	78.49	89.58	81.40	87.76	85.51	84.58	84.29	78.31	89.91	74.29	91.76	86.76	83.02	83.54	78.29	92.33	76.52	93.02	88.93	84.77	85.03
GBT	FSASP	73.16	88.40	86.07	77.08	80.86	81.57	80.72	76.95	91.55	85.54	85.94	85.80	85.74	84.84	76.84	91.03	77.71	90.62	86.93	84.17	84.05	74.64	92.66	78.03	91.27	87.99	84.65	84.13
GBT	RFE	70.40	89.07	87.62	73.26	79.30	80.44	79.23	75.81	92.12	86.78	84.81	85.51	85.79	84.62	75.84	90.78	77.14	90.16	86.44	83.65	83.48	76.47	92.95	78.79	92.02	88.74	85.40	85.05
GBT	FSARFE	60.50	74.38	67.80	67.87	67.84	67.83	67.46	59.59	82.61	71.90	73.24	72.77	72.57	71.41	53.69	83.86	62.29	78.49	73.86	70.39	69.38	53.12	87.40	64.39	81.30	77.11	72.85	71.23
MPCDT	None	72.94	87.72	85.14	77.08	80.47	81.11	80.31	69.71	92.55	88.43	78.91	82.28	83.67	81.57	57.09	93.60	87.43	73.68	77.61	80.56	75.77	51.04	96.92	93.18	70.57	76.17	81.88	73.81
MPCDT	SP	73.28	88.21	85.76	77.30	80.86	81.53	80.71	71.57	92.71	88.43	80.73	83.46	84.58	82.71	58.37	92.96	85.71	75.51	78.43	80.61	76.39	56.67	95.98	90.15	77.31	80.49	83.73	77.61
MPCDT	FSASP	71.47	88.13	86.07	75.06	79.69	80.56	79.58	67.90	93.87	90.91	76.42	81.55	83.66	80.99	56.55	93.04	86.29	73.46	77.12	79.87	75.21	52.59	96.68	92.42	72.57	77.49	82.50	74.97
MPCDT	RFE	71.39	87.89	85.76	75.06	79.56	80.41	79.44	69.65	93.51	90.08	78.46	82.58	84.27	81.94	57.79	93.41	86.86	74.60	78.10	80.73	76.18	53.98	96.74	92.42	74.06	78.61	83.24	76.03
MPCDT	FSARFE	72.49	87.44	84.83	76.63	80.08	80.73	79.93	68.35	92.92	89.26	77.32	81.55	83.29	80.91	56.67	93.57	87.43	73.23	77.29	80.33	75.46	52.40	96.05	90.91	72.82	77.30	81.86	74.66

Fonte: Elaborada pela autora.

Tabela 18 – Resultados dos modelos baseados em DT no experimento 19-21_22, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	73.59	61.69	93.91	22.55	72.27	58.23	57.77	78.98	67.40	90.56	44.73	76.64	67.65	69.07	73.57	71.72	86.51	52.40	73.03	69.45	70.03	76.70	74.35	80.42	69.91	75.71	75.17	75.29
DT	SP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	83.25	56.73	78.83	63.64	74.21	71.23	70.48	77.76	58.80	67.83	70.30	68.80	69.06	68.25	84.59	72.95	74.96	83.18	78.64	79.07	78.61
DT	FSASP	74.57	50.14	85.84	32.73	69.74	59.28	59.71	78.67	64.30	89.21	44.55	75.65	66.88	68.12	73.49	70.00	85.18	52.95	72.45	69.07	69.60	81.88	75.63	79.51	78.32	78.98	78.92	78.81
DT	RFE	76.12	52.58	84.73	38.91	70.84	61.82	62.46	79.56	65.74	89.21	47.45	76.53	68.33	69.62	79.04	72.47	83.61	66.05	76.68	74.83	75.19	81.75	74.47	78.15	78.50	78.31	78.33	78.17
DT	FSARFE	74.08	51.10	87.74	29.45	70.07	58.60	58.85	78.24	62.57	88.66	43.45	74.93	66.06	67.21	73.60	68.71	83.98	53.87	72.08	68.93	69.42	77.37	73.75	79.36	71.40	75.80	75.38	75.45
RF	None	73.88	59.39	92.64	24.73	72.05	58.68	58.56	78.28	78.06	95.16	39.45	78.24	67.31	69.16	79.19	76.69	87.11	64.94	78.35	76.03	76.65	82.15	77.02	81.03	78.32	79.82	79.67	79.63
RF	SP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	78.58	60.48	86.99	45.64	74.43	66.32	67.30	80.03	64.19	73.86	71.77	73.03	72.81	72.29	85.59	73.81	75.72	84.30	79.56	80.01	79.53
RF	FSASP	73.72	48.48	86.55	29.09	69.13	57.82	57.99	76.91	69.86	93.26	35.82	75.81	64.54	65.83	73.99	71.50	86.02	53.69	73.25	69.86	70.44	78.77	76.32	81.64	72.90	77.72	77.27	77.37
RF	RFE	74.10	47.81	84.89	31.82	68.80	58.35	58.67	79.04	74.53	93.58	43.09	78.24	68.33	70.15	77.39	76.51	87.83	60.70	77.11	74.27	74.99	79.00	76.41	81.64	73.27	77.89	77.45	77.55
RF	FSARFE	74.47	49.86	85.84	32.36	69.63	59.10	59.50	77.50	73.05	93.97	37.45	76.81	65.71	67.23	75.79	73.93	86.75	57.56	75.22	72.16	72.81	81.13	76.01	80.27	77.01	78.81	78.64	78.61
GBT	None	73.99	54.35	90.03	27.27	71.00	58.65	58.77	77.30	76.15	95.08	36.00	77.14	65.54	67.08	72.65	77.88	91.20	47.42	73.91	69.31	69.91	73.01	80.05	87.86	60.00	75.38	73.93	74.17
GBT	SP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	78.14	62.53	88.74	43.09	74.88	65.92	67.06	78.72	63.51	73.98	69.37	72.16	71.67	71.29	83.33	73.81	76.63	81.12	78.64	78.88	78.57
GBT	FSASP	73.00	47.60	87.90	25.27	68.91	56.58	56.39	76.19	63.04	91.12	34.73	73.99	62.92	63.89	71.62	66.92	84.22	48.89	70.26	66.55	66.96	74.02	75.83	83.46	63.93	74.71	73.69	73.92
GBT	RFE	74.52	53.33	88.37	30.55	70.84	59.46	59.85	78.28	71.43	92.86	40.91	77.08	66.89	68.49	72.23	73.37	88.67	47.79	72.52	68.23	68.74	72.19	76.01	84.67	59.81	73.53	72.24	72.44
GBT	FSARFE	73.28	47.04	86.55	27.45	68.63	57.00	57.02	77.12	69.66	93.02	36.73	75.92	64.87	66.21	73.65	72.26	86.87	52.40	73.25	69.63	70.23	74.69	76.03	83.31	65.23	75.21	74.27	74.49
MPCDT	None	69.68	0.00	100.0	0.00	69.68	50.00	41.07	92.88	51.91	64.16	88.73	71.62	76.44	70.70	87.46	62.94	67.23	85.24	74.34	76.23	74.22	83.01	75.26	78.60	80.19	79.31	79.40	79.20
MPCDT	SP	77.91	49.63	78.40	48.91	69.46	63.66	63.71	93.94	49.95	60.19	91.09	69.57	75.64	68.94	89.78	62.06	64.58	88.75	74.13	76.66	74.08	86.83	74.39	76.02	85.79	80.40	80.91	80.38
MPCDT	FSASP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	92.90	51.37	63.28	88.91	71.07	76.10	70.20	87.18	61.24	64.70	85.42	72.89	75.06	72.81	84.41	73.34	75.57	82.80	78.81	79.19	78.77
MPCDT	RFE	0.00	30.32	0.00	100.0	30.32	50.00	23.27	91.98	51.73	64.55	87.09	71.40	75.82	70.38	86.30	61.74	66.02	83.95	73.10	74.99	72.98	83.55	75.44	78.60	80.93	79.65	79.77	79.54
MPCDT	FSARFE	78.13	49.20	77.45	50.18	69.18	63.82	63.74	94.90	49.66	59.00	92.73	69.24	75.86	68.72	89.88	61.22	63.13	89.11	73.40	76.12	73.37	87.30	74.64	76.18	86.36	80.74	81.27	80.72

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
DT	None	76.19	80.52	75.29	81.27	78.62	78.28	78.32	70.89	83.20	73.87	81.02	78.27	77.44	77.22	81.02	87.80	72.31	92.16	85.88	82.24	83.17	76.92	88.73	71.43	91.30	85.57	81.37	82.04
DT	SP	81.84	78.04	68.94	87.83	79.46	78.38	78.74	76.64	84.01	73.87	85.90	81.27	79.89	80.09	77.93	87.29	71.49	90.63	84.58	81.06	81.75	79.27	91.15	78.06	91.72	87.78	84.89	85.05
DT	FSASP	82.13	77.12	67.06	88.39	78.94	77.72	78.10	76.31	84.26	74.47	85.53	81.27	80.00	80.13	85.35	87.13	69.83	94.46	86.67	82.14	83.73	83.24	89.72	73.47	94.00	88.07	83.73	84.93
DT	RFE	80.95	77.41	68.00	87.27	78.73	77.63	77.98	76.73	83.73	73.27	86.09	81.16	79.68	79.93	78.18	87.16	71.07	90.82	84.58	80.95	81.71	80.79	89.44	72.96	92.96	87.19	82.96	83.92
DT	FSARFE	77.38	78.25	70.82	83.52	77.89	77.17	77.38	75.65	82.05	69.97	85.90	79.77	77.94	78.31	81.40	87.82	72.31	92.35	86.01	82.33	83.31	81.93	88.30	69.39	93.79	86.75	81.59	83.05
RF	None	80.68	79.86	72.71	86.14	80.19	79.42	79.68	80.00	85.27	75.68	88.16	83.35	81.92	82.23	82.65	88.83	74.79	92.73	87.06	83.76	84.63	82.98	91.85	79.59	93.37	89.40	86.48	86.93
RF	SP	84.23	77.21	66.59	90.07	79.67	78.33	78.76	80.27	83.57	72.07	88.91	82.43	80.49	81.05	84.73	87.54	71.07	94.07	86.80	82.57	84.00	84.66	90.66	76.02	94.41	89.10	85.22	86.30
RF	FSASP	80.99	80.17	73.18	86.33	80.50	79.75	80.01	79.15	83.87	72.97	87.97	82.20	80.47	80.90	82.73	88.99	75.21	92.73	87.19	83.97	84.81	82.26	91.28	78.06	93.17	88.81	85.61	86.16
RF	RFE	78.19	80.76	75.06	83.33	79.67	79.20	79.31	79.30	84.75	74.77	87.78	82.77	81.28	81.61	81.06	89.22	76.03	91.78	86.80	83.91	84.47	81.68	91.80	79.59	92.75	88.95	86.17	86.45
RF	FSARFE	82.01	80.21	72.94	87.27	80.92	80.10	80.40	77.71	84.87	75.38	86.47	82.20	80.92	81.09	82.57	88.67	74.38	92.73	86.93	83.56	84.46	82.70	91.30	78.06	93.37	88.95	85.72	86.32
GBT	None	70.76	83.19	81.41	73.22	76.85	77.32	76.80	68.03	85.86	79.88	76.50	77.80	78.19	77.20	71.43	89.58	78.51	85.47	83.27	81.99	81.14	71.05	92.46	82.65	86.34	85.27	84.49	82.85
GBT	SP	81.99	78.43	69.65	87.83	79.77	78.74	79.09	78.18	83.33	72.07	87.41	81.50	79.74	80.16	83.50	87.48	71.07	93.50	86.41	82.29	83.59	83.52	91.15	77.55	93.79	89.10	85.67	86.44
GBT	FSASP	69.68	79.55	76.24	73.60	74.77	74.92	74.63	73.05	83.24	73.27	83.08	79.31	78.18	78.16	74.17	87.81	73.55	88.15	83.53	80.85	80.92	75.90	90.08	75.51	90.27	86.01	82.89	82.94
GBT	RFE	67.56	79.37	76.94	70.60	73.41	73.77	73.34	69.68	85.48	78.68	78.57	78.61	78.63	77.89	71.48	88.41	75.62	86.04	82.75	80.83	80.35	72.20	92.32	82.14	87.16	85.71	84.65	83.26
GBT	FSARFE	71.17	78.83	74.35	76.03	75.29	75.19	75.07	72.11	84.90	76.88	81.39	79.65	79.13	78.76	73.71	88.91	76.45	87.38	83.92	81.91	81.59	76.53	90.48	76.53	90.48	86.45	83.50	83.50
MPCDT	None	79.90	82.03	76.71	84.64	81.13	80.68	80.79	69.39	87.10	81.68	77.44	79.08	79.56	78.51	67.67	91.61	83.88	81.45	82.22	82.67	80.57	60.69	94.86	89.80	76.40	80.27	83.10	78.53
MPCDT	SP	82.81	81.39	74.82	87.64	81.96	81.23	81.51	73.15	86.80	80.18	81.58	81.04	80.88	80.31	69.05	91.72	83.88	82.60	83.01	83.24	81.33	65.30	94.89	89.29	80.75	83.21	85.02	81.34
MPCDT	FSASP	80.05	81.36	75.53	85.02	80.81	80.27	80.44	71.08	85.86	78.98	79.89	79.54	79.43	78.79	67.12	90.64	81.82	81.45	81.57	81.64	79.77	61.32	94.90	89.80	77.02	80.71	83.41	78.95
MPCDT	RFE	77.36	81.87	77.18	82.02	79.87	79.60	79.61	70.51	87.79	82.58	78.38	80.00	80.48	79.45	66.02	91.67	84.30	79.92	81.31	82.11	79.72	61.28	96.34	92.86	76.19	81.00	84.52	79.46
MPCDT	FSARFE	82.59	80.69	73.65	87.64	81.44	80.64	80.94	73.58	85.58	77.78	82.52	80.69	80.15	79.82	72.76	90.54	80.58	86.04	84.31	83.31	82.35	68.48	95.26	89.80	83.23	85.13	86.51	83.27

Fonte: Elaborada pela autora.

Embora tenham sido apresentados resultados de avaliação sobre registros de até o sétimo semestre, faz-se importante destacar que, na análise, é preciso priorizar os desempenhos preditivos de períodos iniciais. Isso porque, conforme apresentado na Figura 23, cerca de 70% das evasões ocorre nos três primeiros semestres. Observando os resultados apresentados nas Tabelas 16, 17 e 18, percebe-se que nenhum classificador demonstrou comportamento confiável na predição da evasão em dados de admissão (semestre 0), independente do experimento e do método de seleção de atributos considerado. O único modelo que conseguiu identificar uma proporção aceitável de registros da classe negativa foi o MPC-SDP com seleção SP, no experimento 17-19_20 (primeiro bloco da Tabela 16), que apresentou uma revocação negativa de 71.5%[1]. Porém, é preciso observar que, para as predições baseadas em dados de admissão (semestre 0), os modelos MPC-SDP se mostraram bastante instáveis, perante os diferentes métodos de seleção de atributos. Nesse contexto, a maioria dos modelos MPC-SDP apresentou revocação da classe negativa igual ou próxima a 0. Ou seja, nesses casos, todos (ou praticamente todos) os registros de estudantes concluídos/-diplomados foram classificados/preditos, indevidamente, como sendo de estudantes evadidos. Em relação aos *baselines*, seus melhores resultados, em geral associados ao uso das seleções RFE e FSA+RFE, também apresentaram revocações bastante baixas para a classe negativa de evasão, geralmente menores que 40%. Na prática, isso significaria prever mais de 60% dos potenciais concluintes como estudantes em risco de evasão, indevidamente. Considerando a limitação dos recursos institucionais (financeiros e humanos) e que o momento de admissão envolve o maior quantitativo de estudantes (todos os ingressantes/calouros), esses desempenhos preditivos seriam insuficientes para o direcionamento das estratégias preventivas.

Cabe, aqui, destacar a importância de se analisar o desempenho preditivo por período/semestre letivo, além do uso de métricas que evidenciem os resultados por classe, no contexto da predição genérica de evasão, onde os registros de teste costumam ser, naturalmente, desbalanceados entre os semestres. Caso só tivessem sido consideradas a precisão e a revocação da classe de interesse (positiva para evasão), o baixo desempenho da classe sub-representada (negativa para evasão, nos semestres iniciais) não seria evidenciado. Semelhantemente, caso não tivessem sido analisados os resultados por período letivo, baixos desempenhos preditivos em determinados semestres seriam provavelmente mascarados pelos resultados gerais.

Voltando à análise dos resultados, agora considerando as avaliações nos demais períodos, é perceptível que os desempenhos preditivos dos *baselines* (DT, RF, GBT) melhoraram (apresentam cores mais frias) com o avançar dos semestres de avaliação.

[1] Observe que o modelo MPC-SDP com seleção RFE, no experimento 19-21_22, não foi citado/considerado, pois, embora tenha apresentado uma revocação máxima para a classe negativa, ele classificou todos os registros como sendo de estudantes concluídos/diplomados (revocação positiva igual a 0), sem garantir nenhuma separabilidade entre as classes.

Porém, no primeiro e no segundo semestre, os resultados de revocação da classe negativa de evasão ainda se mostraram, de forma geral, baixos. É possível observar que, nos semestres iniciais/principais, os *baselines* apresentaram melhores resultados junto à seleção de atributos SP, e tiveram seus desempenhos frequentemente prejudicados pelo uso da seleção FSA, principalmente pela FSA+SP. Já no semestre de admissão (0), a seleção FSA demonstrou beneficiar o desempenho preditivo dos modelos, o que surpreende, considerando que a FSA tende a selecionar os principais atributos de cada aspecto, e, nos dados de admissão, não existem dados representando os aspectos acadêmicos, interacional, e econômico[2]. Ou seja, alguns atributos contextuais e sociais ignorados pela FSA (para dar lugar a atributos de outros aspectos, na seleção) poderiam ser frequentemente selecionados pelos métodos padrão, oportunizando-lhes uma melhor capacidade preditiva perante os dados de admissão.

Por outro lado, os modelos MPC-SDP demonstraram desempenhos mais constantes para os diferentes métodos de seleção de atributos. Embora o uso da seleção FSA, quando comparado ao do seu respectivo método de seleção base/tradicional, tenha demonstrado beneficiar os resultados dos modelos MPC-SDP em alguns casos e piorá-los em outros, as variações entre os resultados mostraram-se bem menores, não gerando, efetivamente, um prejuízo. E, mais importante, os modelos MPC-SDP apresentaram um comportamento inverso ao dos *baselines*, em relação ao desempenho preditivo ao longo dos semestres. Ao invés de melhorar a revocação da classe negativa (de estudantes concluídos/diplomados) com o avançar dos semestres, os modelos MPC-SDP demonstraram priorizar o desempenho dessa métrica nos semestres iniciais, proporcionando-lhes, consequentemente, melhores precisões. Isso é bastante benéfico, pois no início dos cursos, quando o quantitativo de alunos matriculados é maior, é preciso priorizar a precisão das predições de risco de evasão. Caso contrário, muitos falsos positivos (potenciais concluintes) se somarão a um grande quantitativo de estudantes em risco (verdadeiros positivos)[3], prejudicando o direcionamento das medidas preventivas. Ou seja, como a limitação institucional de recursos humanos e financeiros inviabilizará provavelmente o acompanhamento/auxílio de todos os estudantes em risco, é mais vantajoso que o modelo recupere uma parcela menor (menor revocação), mas mais correta (maior precisão), de casos de evasão (classe positiva), nos semestres iniciais.

Uma lógica inversa pode ser aplicada aos semestres mais avançados. Nesse contexto, onde o quantitativo de evasões é pequeno, é aceitável relaxar um pouco a precisão das predições em prol da recuperação de um quantitativo maior de estudantes

[2]Note que os dados econômicos são informados pelos estudantes, no SIGAA, geralmente antes das solicitações de bolsas/auxílios e das matrículas semestrais, com exceção do primeiro semestre, em que a matrícula é realizada pela Coordenação de Registros Acadêmicos.

[3]Lembre-se de que, conforme apresentado na Figura 23, cerca de 70% das evasões ocorre nos três primeiros semestres.

em risco de evasão. Considerando que estudantes em semestres mais avançados já dispenderam mais recursos institucionais e, por estarem mais próximos da conclusão do curso, talvez sejam mais facilmente dissuadidos da decisão de evadir, essa lógica, inclusive, se torna bastante razoável. Pode-se observar, nas Tabelas 16, 17 e 18, que esse foi o comportamento apresentado pelos modelos MPC-SDP. Mais detalhadamente, a partir do quinto semestre, os modelos MPC-SDP passaram a apresentar a maioria das precisões inferiores a 75%, reduzindo-as progressivamente e chegando a pouco mais de 50% (experimentos 17-19_20 e 18-20_21) ou 60% (experimento 19-21_22), no sétimo semestre. Embora uma precisão de pouco mais de 50% seja bastante ruim, pois quase a metade dos estudantes preditos como evadidos correspondem a falsos positivos, é preciso observar que, no sétimo semestre, a quantidade de registros de estudantes evadidos é cerca de três vezes menor que a de concluídos (vide dados relativos aos conjuntos de teste, na Tabela 14). Isso faz com que os falsos positivos tenham um grande impacto negativo na precisão da classe positiva, mesmo que a revocação da classe negativa se mantenha relativamente boa[4].

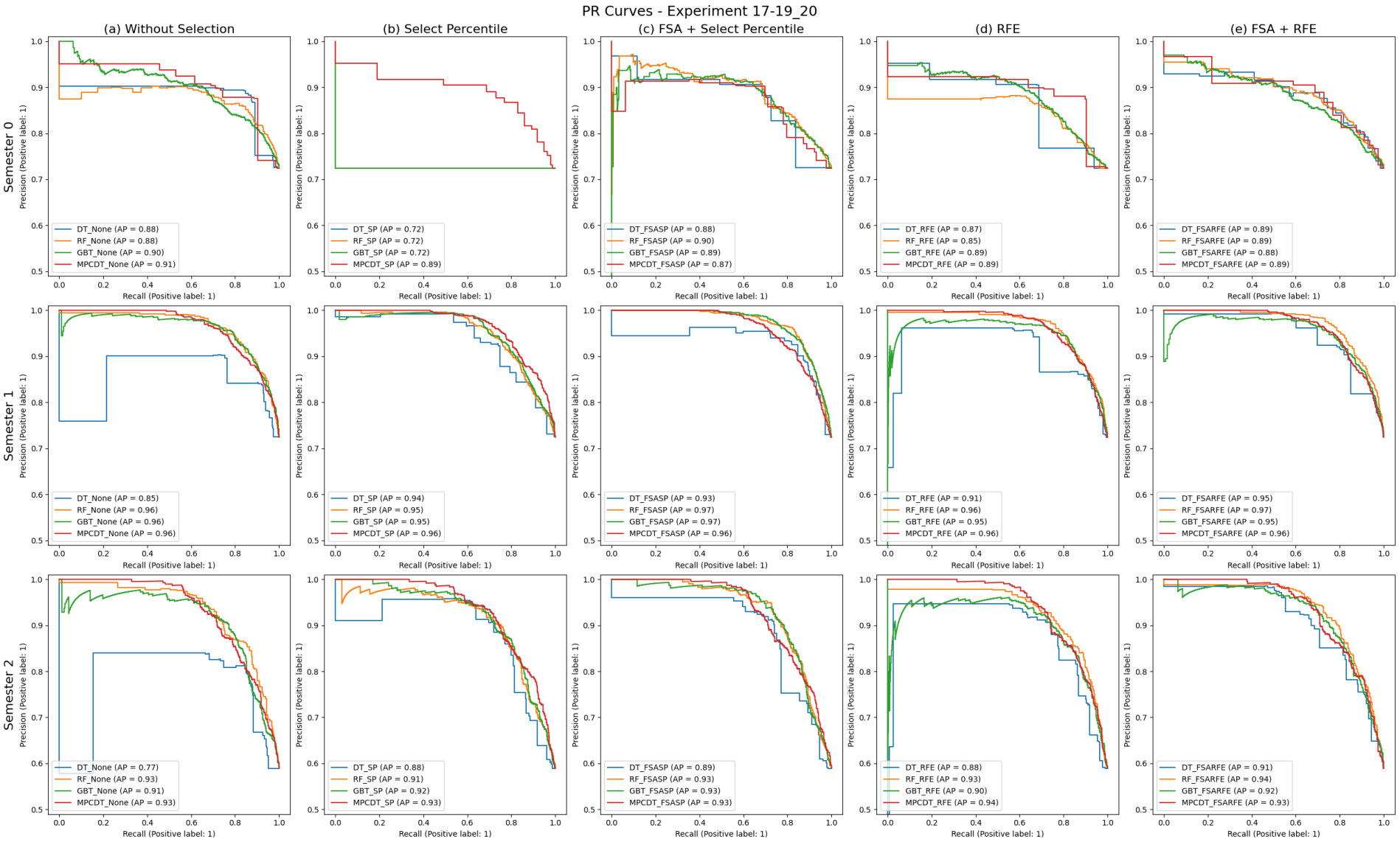
Além disso, no experimento 19-21_22 (Tabela 18), os modelos MPC-SDP apresentaram, de forma geral, maior precisão na classe positiva de evasão e, consequentemente, reduções de precisão bem menos problemáticas, ao longo dos semestres. Porém, considerando o *trade-off* precisão vs. revocação, os modelos também demonstraram maior dificuldade na identificação dos registros de estudantes evadidos (menor revocação da classe positiva). Considerando que, nesse experimento, o conjunto de treinamento dos modelos contemplou, majoritariamente, registros de estudantes diplomados e evadidos durante o ERE (anos de 2020 e 2021), é possível que os padrões de evasão tenham se mostrado mais complexos durante a pandemia de COVID-19. Ainda, o enriquecimento dos dados de interação dos estudantes nas turmas virtuais do SIGAA, durante o ERE, também pode ter contribuído para a detecção de padrões de evasão menos abrangentes, mas mais precisos.

Complementarmente, nas Figuras 24, 25, e 26, são apresentadas as curvas PR dos modelos, considerando as avaliações em registros dos semestres de 0 (admissão) a 2, para os experimentos 17-19_20, 18-20_21, e 19-21_22, respectivamente. Embora as predições dos modelos e os desempenhos avaliados anteriormente considerem apenas um limiar de classificação/decisão[5], as relações precisão vs. revocação dos modelos perante diferentes limiares podem contribuir para observações adicionais.

[4]Nas Tabelas 16, 17, e 18, é possível observar que, em geral, os modelos MPC-SDP apresentaram revocações da classe negativa superiores a 70%, quando avaliados sobre registros de sétimo semestre.

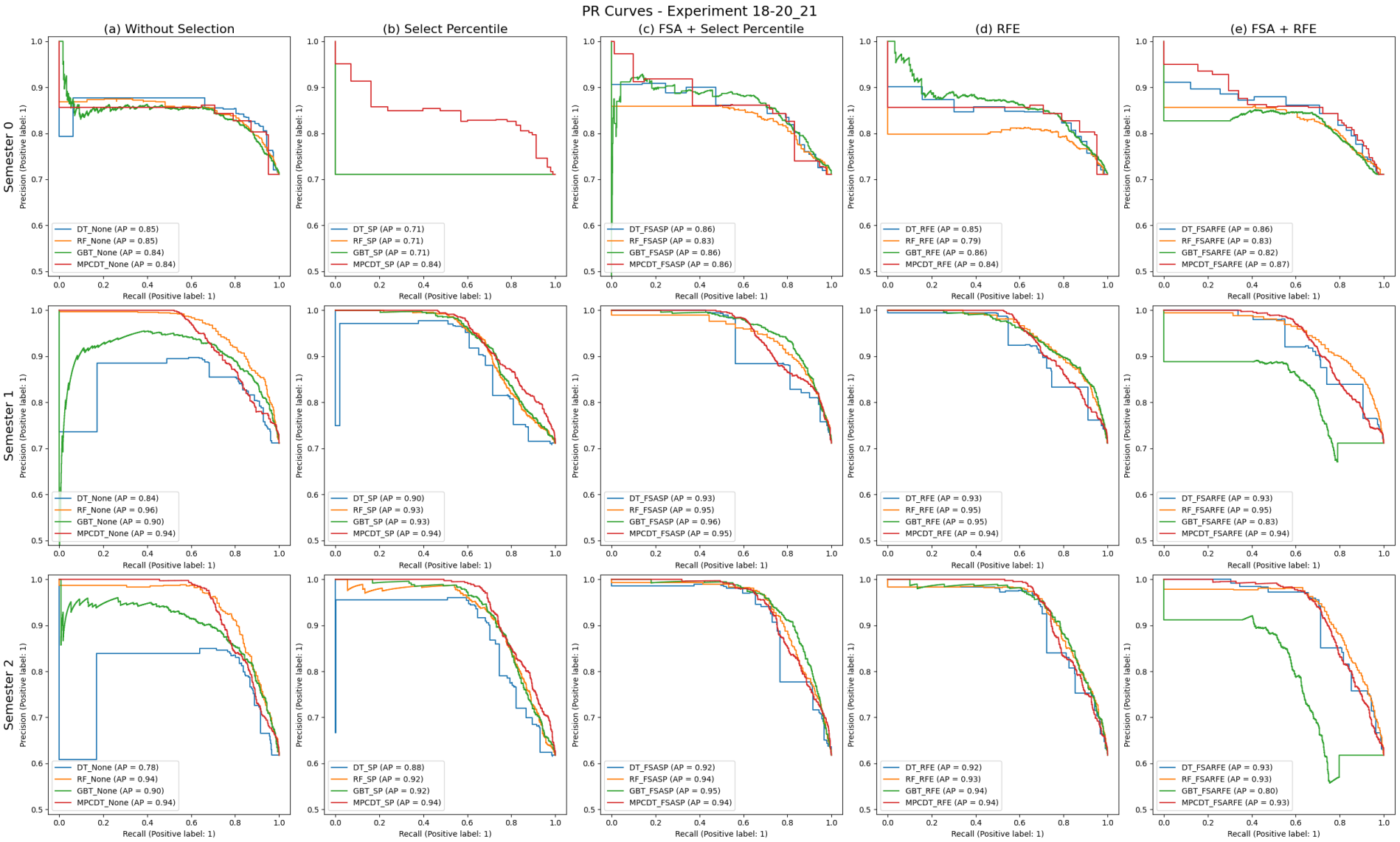
[5]Assim como os classificadores do *scikit-learn*, o MPC-SDP considera o limiar de classificação de 0.5. Ou seja, os modelos classificam/predizem como sendo da classe positiva de evasão os registros de estudantes que obtém uma probabilidade positiva maior que 0.5. Os demais registros são preditos como sendo da classe negativa de evasão.

Figura 24 – Curvas PR dos modelos baseados em DT, para o experimento 17-19_20 e os semestres de 0 a 2



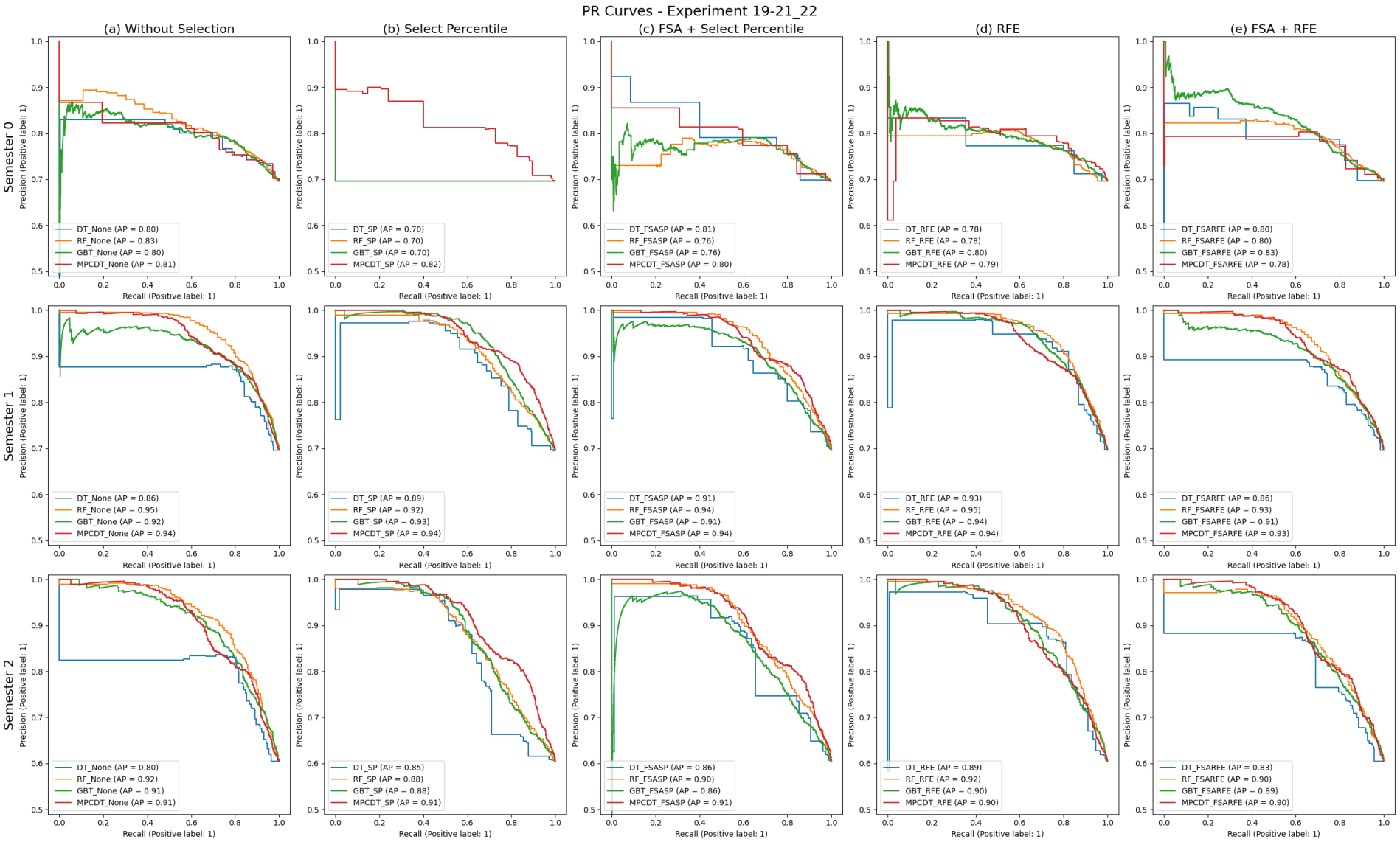
Fonte: Elaborada pela autora.

Figura 25 – Curvas PR dos modelos baseados em DT, para o experimento 18-20_21 e os semestres de 0 a 2



Fonte: Elaborada pela autora.

Figura 26 – Curvas PR dos modelos baseados em DT, para o experimento 19-21_22 e os semestres de 0 a 2



Fonte: Elaborada pela autora.

Além de confirmar que os desempenhos preditivos foram insatisfatórios perante os dados de admissão, as Figuras 24, 25, e 26 revelam que, de forma geral, nas predições de registros dos semestres 1 e 2, os modelos MPC-SDP apresentaram precisões próximas a 100% para uma maior quantidade de limiares de decisão, demonstrando que esses classificadores apresentam maior tendência em beneficiar a precisão, em detrimento da revocação, nos semestres iniciais. Também se confirma uma maior estabilidade no desempenho preditivo dos modelos MPC-SDP, perante os diferentes métodos de seleção de atributos. Porém, ao contrário do observado nos resultados preditivos (relativos ao limiar de decisão 0.5), as curvas PR demonstram que, considerando diferentes limiares, a seleção FSA frequentemente manteve ou melhorou os desempenhos preditivos dos *baselines* DT e RF, quando comparados aos correspondentes métodos de seleção tradicionais. Isso sugere que, se fosse escolhido/utilizado outro limiar de predição, a seleção FSA poderia favorecer esses modelos.

Por fim, para uma comparação mais efetiva sobre os resultados preditivos dos modelos baseados em DT, foram conduzidos testes estatísticos. Primeiramente, foi aplicado o teste de Friedman sobre os resultados de UAR, F1, e acurácia nos semestres de 1 a 7[6], para cada experimento das Tabelas 16, 17 e 18. Quando rejeitada a hipótese nula, de que os modelos demonstram resultados semelhantes, considerando um nível de significância de 5%, foi aplicado o teste *post-hoc* de Nemenyi[7], para identificar quais modelos apresentaram diferenças significativas. Os resultados dos testes são apresentados em blocos, na Figura 27, variando as métricas UAR, F1, e acurácia, verticalmente, e os experimentos 17-19_20, 18-20_21 e 19-21_22, horizontalmente. A comparação pareada (modelo vs. modelo) do teste *post-hoc* é apresentada em forma de mapa de calor, com as tonalidades de verde representando a força das probabilidades de significância (p). Faz-se importante destacar que, embora o teste *post-hoc* sinalize os modelos com resultados significativamente distintos, a relação de superioridade/inferioridade entre eles deve ser inferida a partir dos seus desempenhos, entre os semestres de 1 a 7, nas métricas e experimentos analisados (Tabelas 16, 17 e 18).

Considerando, na Figura 27, os resultados dos testes estatísticos para o experimento 17-19_20, pode-se observar que o modelo DT associado à seleção FSA+RFE demonstrou, na métrica UAR, desempenho significativamente pior que os modelos: RF sem nenhum método de seleção de atributos (None) ou associado às seleções tradicionais (SP e RFE); GBT com seleção SP; e MPC-SDP associado às seleções RFE e FSA+SP. Em especial, nessa comparação, o modelo MPC-SDP com seleção

[6]Como já foi percebida a ineficácia dos modelos genéricos para a predição em momento de admissão, o semestre 0 foi desconsiderado nos testes estatísticos, para ser possível comparar, sem ruído, o desempenho dos modelos em situações viáveis/úteis de predição (a partir dos dados de primeiro semestre).

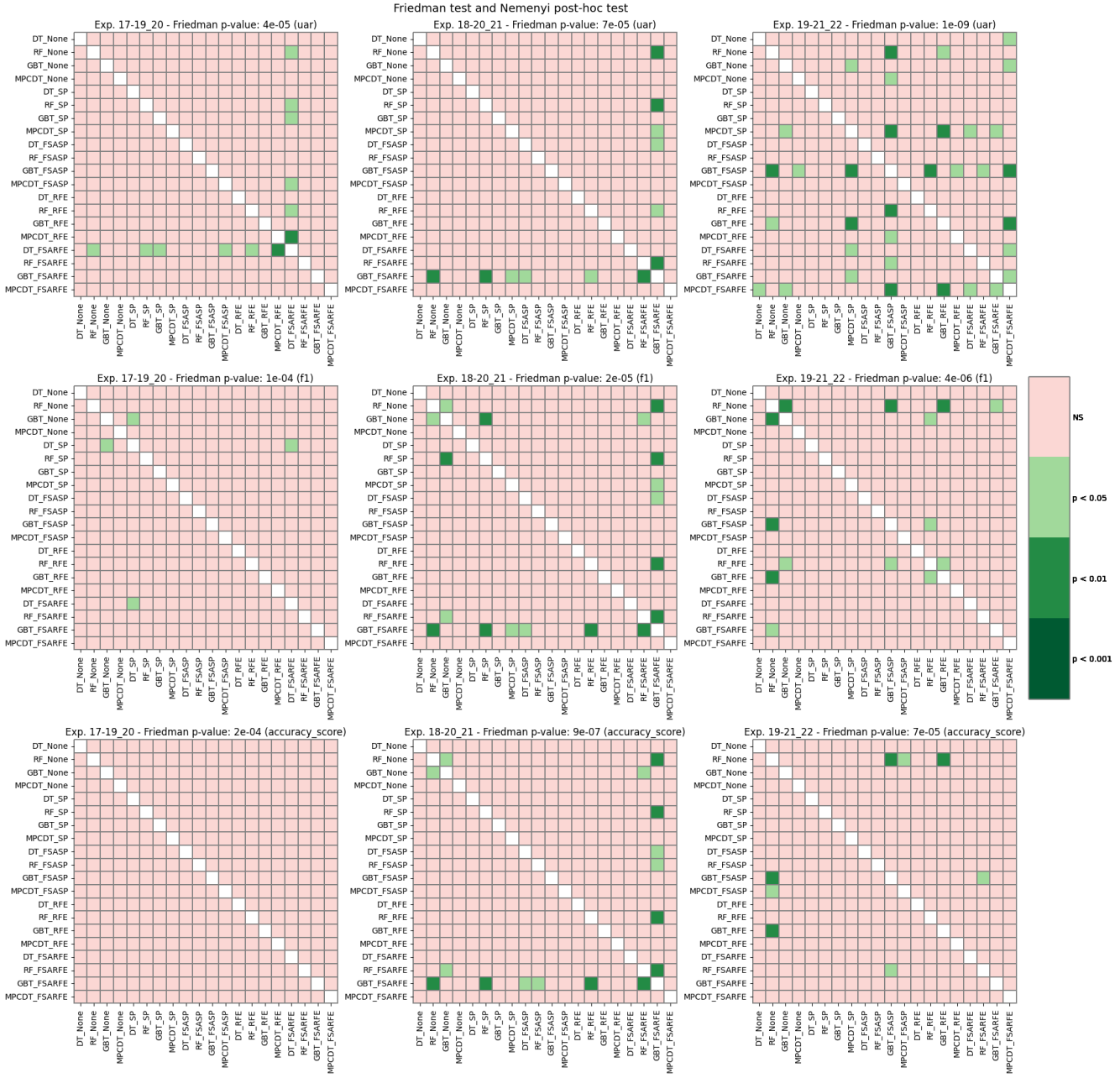
[7]Foram consideradas as implementações das bibliotecas/pacotes *SciPy* (`https://scipy.org/`) e *scikit-posthocs* (`https://scikit-posthocs.readthedocs.io/`) para os testes de Friedman e *post-hoc* de Nemenyi, respectivamente.

RFE demonstrou diferença com maior probabilidade de significância ($p < 0.01$). Já em relação à métrica F1, o modelo DT com seleção SP superou significativamente os modelos DT associado à seleção FSA+RFE, e GBT sem nenhum método de seleção de atributos. Entre os resultados de acurácia, embora o teste de Friedman tenha indicado a existência de diferença entre os desempenhos dos modelos ($p = 0.0002$), o teste *post-hoc* de Nemenyi não encontrou nenhum par de modelos com resultados significativamente distintos.

Os resultados dos testes estatísticos sobre as métricas UAR, F1, e acurácia, no experimento 18-20_21, revelam que, de forma geral, o modelo GBT associado à seleção FSA+RFE demonstrou desempenhos significativamente piores que: o DT associado à seleção FSA+SP; a maioria dos modelos RF, considerando as diferentes combinações de seleção de atributos; e o MPC-SDP com seleção SP. Além disso, o modelo GBT sem nenhum método de seleção de atributos também se mostrou significativamente pior, em F1 e acurácia, que o RF sem seleção de atributos ou associado, principalmente, à seleção FSA+RFE. Em todos os casos, modelos RF apresentaram as diferenças de desempenho com maior probabilidade de significância ($p < 0.01$).

Para os testes envolvendo o experimento 19-21_22, de forma geral e considerando todas as métricas, os modelos RF sem seleção de atributos ou associados às seleções RFE e FSA+RFE frequentemente demonstraram desempenhos significativamente melhores que modelos GBT, especialmente quando associados à seleção FSA+SP. Porém, na métrica UAR, os modelos MPC-SDP se destacaram mais, principalmente os associados às seleções FSA+RFE e SP, que superaram, significativamente, todos os modelos GBT, com exceção do associado à seleção SP (que também não demonstrou perda significativa perante nenhum outro modelo ou métrica). Além disso, os modelos MPC-SDP com FSA+RFE e SP também superaram, em UAR, o modelo DT associado à seleção FSA+RFE e, ainda, o DT sem seleção de atributos, no caso do MPC-SDP com FSA+RFE. Inclusive, na métrica UAR, os modelos MPC-SDP com seleção FSA+RFE ou SP apresentaram a maioria das diferenças com probabilidade de significância superior ($p < 0.01$). Por outro lado, na métrica de acurácia, pela primeira vez um modelo MPC-SDP demonstrou resultado significativamente pior. Mais especificamente, o MPC-SDP com seleção FSA+SP foi superado pelo RF sem seleção de atributos. Analisando os desempenhos desses modelos nas diferentes métricas da Tabela 18, pode-se observar que a superioridade em acurácia do modelo RF ocorre por ele beneficiar a revocação da classe majoritária/mais representativa (veja Tabela 14), ao longo dos semestres. Porém, como discutido anteriormente, esse comportamento não significa, necessariamente, uma vantagem preditiva para a realidade/prática educacional, já que, nos semestres iniciais, beneficiar a revocação da classe majoritária – positiva para evasão – implica em reduzir sua precisão.

Figura 27 – Testes estatísticos sobre os desempenhos de UAR, F1, e acurácia dos modelos baseados em DT, para cada experimento



Fonte: Elaborada pela autora.

Resumindo, os testes estatísticos evidenciaram melhores desempenhos dos modelos RF e MPC-SDP, e piores performances dos modelos GBT. Estes, além de terem se mostrado frequentemente piores que os demais modelos de *ensemble*, chegaram a ser superados, em algumas ocasiões, por modelos DT associados, principalmente, à seleção FSA+SP. Em relação aos melhores classificadores, os modelos RF demonstraram melhores desempenhos sem seleção de atributos e com as seleções RFE e FSA+RFE, nessa ordem. Já os modelos MPC-SDP demonstraram superioridade com as seleções SP e FSA+RFE, com igual frequência. Considerando que os modelos MPC-SDP foram construídos com até 9 subclassificadores[8], enquanto os demais *ensembles* (RF e GBT) foram construídos com 100 subclassificadores[9], seus resultados se mostraram bastante positivos/favoráveis, principalmente por beneficiarem, em semestres iniciais, a precisão das predições de evasão.

7.1.2 Modelos baseados em LR

Seguindo o mesmo padrão de apresentação da Seção 7.1.1, as Tabelas 19, 20 e 21 mostram os resultados da avaliação dos classificadores baseados em LR, desenvolvidos nos experimentos 17-19_20, 18-20_21, e 19-21_22, respectivamente. Assim como verificado nos modelos baseados em DT, pode-se observar que nenhum classificador demonstrou comportamento confiável na predição da evasão em dados de admissão (semestre 0). Embora os modelos MPC-SDP com seleção FSA+RFE, no experimento 17-19_20 (primeiro bloco da Tabela 19), e com seleção SP, no experimento 18-20_21 (primeiro bloco da Tabela 20), tenham identificado boas proporções de registros da classe negativa (revocações negativas de 75.5% e 75.0%, respectivamente), novamente os modelos MPC-SDP demonstraram instabilidade preditiva, perante os diferentes métodos de seleção de atributos, no momento de admissão (semestre 0)[10].

Em relação às avaliações dos demais períodos, também é possível verificar comportamentos semelhantes aos dos modelos baseados em DT. Ou seja, embora tenham melhorado as revocações da classe negativa com o avançar dos semestres, de forma geral, os *baselines* (LR, *Bagging*, e *AdaBoost*) apresentaram resultados insatisfatórios nos semestres iniciais, recuperando uma quantidade maior de falsos positivos e, dessa forma, prejudicando a precisão das predições da classe positiva de evasão. Já os modelos MPC-SDP novamente demonstraram beneficiar a identificação

[8]Lembrando que, no MPC-SDP, é treinado um subclassificador especializado para cada semestre cujo conjunto de treinamento satisfaça um limite mínimo de amostras por classe (condição de parada).

[9]Embora o número de subclassificadores (*n_estimators*) tenha sido um dos hiperparâmetros otimizados para os *baselines* de *ensemble*, verificou-se que o quantitativo padrão de 100 subclassificadores foi escolhido como melhor configuração, na construção de todos os modelos RF e GBT.

[10]Observe que os modelos MPC-SDP com seleção SP e RFE obtiveram, no experimento 17-19_20, revocações máximas (100%) para a classe negativa de evasão, perante os registros de admissão. Porém, esses resultados estão associados a revocações positivas mínimas (0%), o que indica que os modelos classificaram todos os registros como sendo da classe "negativa", sem demonstrar capacidade preditiva/discriminativa.

de estudantes concluídos/diplomados (revocação da classe negativa) nos semestres iniciais, favorecendo as precisões das predições de evasão (classe positiva), durante os três primeiros semestres da trajetória acadêmica dos estudantes. Essa, inclusive, é a maior vantagem demonstrada pelos modelos MPC-SDP, já que, além de concentrar a maioria das evasões, os semestres iniciais detém um volume maior de estudantes. E, nesse cenário, um bom aproveitamento/direcionamento das medidas preventivas depende da precisão das predições da classe positiva.

Nas Figuras 28, 29, e 30, também são apresentadas as curvas PR dos modelos baseados em LR, considerando as avaliações em registros dos semestres de 0 (admissão) a 2, para os experimentos 17-19_20, 18-20_21, e 19-21_22, respectivamente. Novamente, as curvas PR demonstram que (i) os resultados preditivos foram insatisfatórios nos dados de admissão; (ii) considerando os diferentes limiares de decisão, a seleção FSA frequentemente manteve ou melhorou os desempenhos preditivos dos modelos, quando comparados aos correspondentes métodos de seleção tradicionais (SP e RFE); e (iii) os modelos MPC-SDP baseados em LR também demonstraram estabilidade, perante os diferentes métodos de seleção de atributos, a partir do primeiro semestre.

Por fim, também seguindo os padrões de avaliação e apresentação considerados na Seção 7.1.1, na Figura 27, são apresentados os resultados do teste estatístico de Friedman e do teste *post-hoc* de Nemenyi, aplicados sobre os desempenhos de UAR, F1, e acurácia dos experimentos 17-19_20, 18-20_21 e 19-21_22, separadamente. Pode-se observar que o teste de Friedman não indicou diferença significativa ($p > 0.05$) entre os desempenhos dos modelos do experimento 17-19_20. E, apesar de ter indicado entre os resultados das métricas do experimento 18-20_21, o teste *post-hoc* de Nemenyi só identificou diferença nos desempenhos de UAR do modelo MPC-SDP com seleção RFE, que superou significativamente o *Bagging* com seleção FSA+SP.

Já os resultados dos testes estatísticos do experimento 19-21_22 demonstraram, na Figura 31, que o modelo MPC-SDP com seleção SP apresentou resultados de UAR significativamente superiores aos do modelo *AdaBoost* com seleção FSA+RFE e aos de todos os *baselines* (LR, *Bagging*, e *AdaBoost*) sem seleção de atributos ou com seleção FSA+SP. Além disso, os modelos *AdaBoost* sem seleção de atributos e com seleção FSA+SP também foram superados pelos modelos LR e/ou *Bagging* com seleção RFE. Nas demais métricas analisadas, os *baselines* sem seleção de atributos foram superados pelos modelos LR e *Bagging* com seleção RFE. Especificamente em F1, o *AdaBoost* sem seleção de atributos também foi superado pelos modelos MPC-SDP, LR, e *Bagging* com seleção SP. Já em acurácia, o *AdaBoost* com seleção FSA+SP também apresentou desempenhos significativamente inferiores aos dos modelos LR e *Bagging* com seleção RFE.

Tabela 19 – Resultados dos modelos baseados em LR no experimento 17-19_20, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	76.04	88.75	99.15	17.75	76.74	58.45	57.83	84.06	81.96	95.63	52.25	83.69	73.94	76.64	76.79	86.23	93.38	59.50	79.47	76.44	77.35	75.05	90.60	93.22	67.84	80.76	80.53	80.37
LR	SP	72.55	66.67	99.91	0.50	72.54	50.20	42.53	84.32	73.24	92.40	54.75	82.04	73.58	75.42	74.71	80.58	90.59	56.00	76.39	73.30	73.98	73.33	87.04	90.56	65.83	78.42	78.19	78.00
LR	FSASP	72.74	53.33	99.34	2.00	72.54	50.67	43.92	84.77	69.39	90.41	57.25	81.28	73.83	75.12	82.19	75.90	83.62	74.00	79.67	78.81	78.92	81.27	80.25	80.87	80.65	80.76	80.76	80.76
LR	RFE	73.45	64.86	98.77	6.00	73.23	52.38	47.61	85.52	72.48	91.45	59.25	82.59	75.35	76.79	76.15	79.21	89.02	60.00	77.10	74.51	75.18	73.84	85.35	88.86	67.34	78.30	78.10	77.97
LR	FSARFE	72.47	0.00	100.0	0.00	72.47	50.00	42.02	88.27	68.56	87.94	69.25	82.79	78.59	78.51	79.90	75.00	83.80	69.75	78.03	76.77	77.04	78.18	81.40	83.29	75.88	79.65	79.59	79.60
BAGLR	None	76.26	89.29	99.15	18.75	77.01	58.95	58.60	83.87	82.40	95.82	51.50	83.62	73.66	76.42	76.68	86.18	93.38	59.25	79.36	76.31	77.22	74.61	90.51	93.22	67.09	80.39	80.15	79.97
BAGLR	SP	72.51	40.00	99.72	0.50	72.40	50.11	42.48	84.39	72.22	91.93	55.25	81.83	73.59	75.30	74.64	80.00	90.24	56.00	76.18	73.12	73.79	72.62	86.82	90.56	64.57	77.81	77.56	77.33
BAGLR	FSASP	72.53	50.00	99.81	0.50	72.47	50.16	42.50	83.26	69.00	91.17	51.75	80.32	71.46	73.09	80.87	75.66	83.97	71.50	78.85	77.74	77.96	83.17	82.04	82.57	82.66	82.61	82.61	82.61
BAGLR	RFE	73.50	65.79	98.77	6.25	73.30	52.51	47.85	85.52	72.48	91.45	59.25	82.59	75.35	76.79	76.15	79.21	89.02	60.00	77.10	74.51	75.18	73.69	85.30	88.86	67.09	78.18	77.97	77.84
BAGLR	FSARFE	72.47	0.00	100.0	0.00	72.47	50.00	42.02	88.37	68.81	88.03	69.50	82.93	78.77	78.68	80.13	74.93	83.62	70.25	78.13	76.94	77.18	78.00	81.35	83.29	75.63	79.53	79.46	79.47
ADALR	None	75.85	87.18	99.05	17.00	76.46	58.03	57.18	83.76	74.20	93.07	52.50	81.90	72.78	74.83	77.79	82.84	90.94	62.75	79.36	76.85	77.63	75.95	89.10	91.77	69.85	81.01	80.81	80.71
ADALR	SP	72.57	100.0	100.0	0.50	72.61	50.25	42.55	84.16	72.82	92.31	54.25	81.83	73.28	75.11	75.00	80.50	90.42	56.75	76.59	73.58	74.28	72.90	86.91	90.56	65.08	78.05	77.82	77.60
ADALR	FSASP	74.19	68.97	98.29	10.00	73.98	54.15	51.01	85.73	68.84	89.55	60.75	81.62	75.15	76.07	79.22	71.83	81.01	69.50	76.28	75.26	75.38	79.24	79.34	80.39	78.14	79.28	79.26	79.27
ADALR	RFE	73.30	73.08	99.34	4.75	73.30	52.04	46.64	86.18	71.39	90.60	61.75	82.66	76.17	77.28	76.96	77.40	87.28	62.50	77.10	74.89	75.48	75.42	84.59	87.65	70.35	79.16	79.00	78.95
ADALR	FSARFE	72.47	0.00	100.0	0.00	72.47	50.00	42.02	88.38	68.98	88.13	69.50	83.00	78.81	78.75	80.10	75.47	84.15	70.00	78.34	77.07	77.35	78.72	81.55	83.29	76.63	80.02	79.96	79.98
MPCLR	None	88.10	35.22	40.08	85.75	52.65	62.91	52.51	93.92	62.96	80.72	86.25	82.24	83.49	79.80	87.14	77.96	83.80	82.25	83.16	83.02	82.74	80.58	85.67	87.41	78.14	82.86	82.77	82.80
MPCLR	SP	0.00	27.53	0.00	100.0	27.53	50.00	21.59	93.51	61.18	79.39	85.50	81.07	82.45	78.60	86.75	75.11	81.01	82.25	81.52	81.63	81.15	78.34	84.46	86.68	75.13	81.01	80.90	80.91
MPCLR	FSASP	76.27	60.26	94.30	22.75	74.60	58.53	58.68	91.26	71.49	88.22	77.75	85.34	82.99	82.10	82.80	79.20	86.41	74.25	81.42	80.33	80.61	81.59	83.51	84.75	80.15	82.49	82.45	82.47
MPCLR	RFE	0.00	27.53	0.00	100.0	27.53	50.00	21.59	92.00	60.56	79.77	81.75	80.32	80.76	77.51	82.99	75.88	83.28	75.50	80.08	79.39	79.41	73.24	84.39	88.14	66.58	77.56	77.36	77.22
MPCLR	FSARFE	88.02	47.56	68.38	75.50	70.34	71.94	67.66	94.40	54.62	71.98	88.75	76.60	80.37	74.65	88.12	69.43	73.69	85.75	78.64	79.72	78.50	82.06	80.45	80.87	81.66	81.26	81.26	81.26

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	68.67	90.74	88.37	73.87	79.57	81.12	79.36	67.44	92.88	87.88	78.84	81.85	83.36	80.80	63.33	92.79	82.61	82.40	82.46	82.50	79.49	50.65	93.24	80.41	77.51	78.16	78.96	73.40
LR	SP	67.72	87.06	82.95	74.37	77.74	78.66	77.39	72.12	90.51	82.32	84.13	83.53	83.23	82.05	77.54	91.73	77.54	91.73	87.91	84.63	84.63	78.89	92.46	73.20	94.38	89.66	83.79	84.67
LR	FSASP	78.80	84.98	76.36	86.68	82.62	81.52	81.69	79.27	88.81	77.27	89.92	85.71	83.60	83.81	72.99	89.89	72.46	90.13	85.38	81.30	81.37	75.90	90.34	64.95	94.08	87.59	79.52	81.09
LR	RFE	67.92	87.57	83.72	74.37	78.05	79.05	77.72	71.00	90.66	82.83	83.12	83.03	82.98	81.59	71.81	91.48	77.54	88.80	85.77	83.17	82.34	76.47	94.29	80.41	92.90	90.11	86.66	85.99
LR	FSARFE	73.58	83.89	75.58	82.41	79.73	79.00	78.86	77.47	86.20	71.21	89.67	83.53	80.44	81.06	79.65	88.00	65.22	93.87	86.16	79.54	81.28	86.15	88.92	57.73	97.34	88.51	77.53	81.04
BAGLR	None	68.56	90.99	88.76	73.62	79.57	81.19	79.38	67.44	92.88	87.88	78.84	81.85	83.36	80.80	63.54	93.07	83.33	82.40	82.65	82.87	79.76	50.65	93.24	80.41	77.51	78.16	78.96	73.40
BAGLR	SP	67.29	87.46	83.72	73.62	77.59	78.67	77.28	72.05	90.98	83.33	83.88	83.70	83.61	82.29	76.26	91.44	76.81	91.20	87.33	84.01	83.93	78.02	92.44	73.20	94.08	89.43	83.64	84.39
BAGLR	FSASP	78.35	85.32	77.13	86.18	82.62	81.66	81.74	76.38	88.38	76.77	88.16	84.37	82.46	82.42	70.14	89.97	73.19	88.53	84.41	80.86	80.44	78.82	91.43	69.07	94.67	88.97	81.87	83.32
BAGLR	RFE	67.92	87.57	83.72	74.37	78.05	79.05	77.72	71.00	90.66	82.83	83.12	83.03	82.98	81.59	71.33	91.46	77.54	88.53	85.58	83.03	82.14	76.47	94.29	80.41	92.90	90.11	86.66	85.99
BAGLR	FSARFE	72.93	83.59	75.19	81.91	79.27	78.55	78.39	76.92	85.96	70.71	89.42	83.19	80.06	80.67	78.26	87.94	65.22	93.33	85.77	79.28	80.85	86.15	88.92	57.73	97.34	88.51	77.53	81.04
ADALR	None	69.14	89.76	86.82	74.87	79.57	80.85	79.31	69.51	92.26	86.36	81.11	82.86	83.74	81.68	66.27	92.44	81.16	84.80	83.82	82.98	80.71	66.96	93.75	79.38	88.76	86.67	84.07	81.91
ADALR	SP	67.30	86.98	82.95	73.87	77.44	78.41	77.10	71.24	89.97	81.31	83.63	82.86	82.47	81.31	75.52	91.89	78.26	90.67	87.33	84.46	84.07	78.89	92.46	73.20	94.38	89.66	83.79	84.67
ADALR	FSASP	73.23	84.24	76.36	81.91	79.73	79.13	78.91	70.70	87.89	76.77	84.13	81.68	80.45	79.79	67.09	90.99	76.81	86.13	83.63	81.47	80.06	68.27	92.15	73.20	90.24	86.44	81.72	80.91
ADALR	RFE	69.23	87.79	83.72	75.88	78.96	79.80	78.60	70.39	90.61	82.83	82.62	82.69	82.72	81.27	69.74	91.14	76.81	87.73	84.80	82.27	81.25	73.33	93.94	79.38	91.72	88.97	85.55	84.53
ADALR	FSARFE	74.24	84.18	75.97	82.91	80.18	79.44	79.32	77.09	85.58	69.70	89.67	83.03	79.68	80.39	80.91	87.84	64.49	94.40	86.35	79.45	81.39	84.85	88.89	57.73	97.04	88.28	77.39	80.75
MPCLR	None	69.47	89.55	86.43	75.38	79.73	80.91	79.44	61.89	93.20	89.39	72.54	78.15	80.97	77.36	51.68	94.55	89.13	69.33	74.66	79.23	72.71	45.59	98.27	95.88	67.16	73.56	81.52	70.79
MPCLR	SP	66.77	89.66	87.21	71.86	77.90	79.53	77.70	63.12	93.61	89.90	73.80	79.16	81.85	78.35	48.65	95.28	91.30	64.53	71.73	77.92	70.21	49.47	98.38	95.88	71.89	77.24	83.88	74.17
MPCLR	FSASP	75.27	87.27	81.40	82.66	82.16	82.03	81.56	74.65	90.48	81.82	86.15	84.71	83.98	83.17	64.37	92.33	81.16	83.47	82.85	82.31	79.73	65.12	95.75	86.60	86.69	86.67	86.64	82.67
MPCLR	RFE	61.52	89.20	87.98	64.32	73.63	76.15	73.58	57.84	92.73	89.39	67.51	74.79	78.45	74.19	46.49	95.04	91.30	61.33	69.40	76.32	68.08	47.94	98.34	95.88	70.12	75.86	83.00	72.89
MPCLR	FSARFE	71.29	88.10	83.72	78.14	80.34	80.93	79.91	60.20	93.02	89.39	70.53	76.81	79.96	76.09	50.00	95.06	90.58	66.67	73.10	78.62	71.40	49.21	98.77	96.91	71.30	77.01	84.10	74.05

Fonte: Elaborada pela autora.

Tabela 20 – Resultados dos modelos baseados em LR no experimento 18-20_21, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	75.79	71.43	96.01	24.55	75.35	60.28	60.63	79.77	78.60	95.55	40.27	79.60	67.91	70.10	72.72	86.92	96.12	41.70	75.32	68.91	69.58	63.61	87.28	93.68	44.72	69.58	69.20	67.45
LR	SP	71.99	62.50	98.64	5.58	71.74	52.11	46.74	81.67	74.74	93.38	48.32	80.37	70.85	72.91	74.70	84.65	94.59	48.21	76.86	71.40	72.45	66.09	85.19	91.29	51.69	71.79	71.49	70.50
LR	FSASP	72.31	58.06	97.64	8.04	71.74	52.84	48.60	81.69	69.38	91.11	49.66	79.15	70.39	72.02	77.90	79.09	90.43	58.52	78.23	74.48	75.48	70.73	80.23	84.75	63.82	74.45	74.28	74.10
LR	RFE	73.89	54.61	93.74	18.53	72.00	56.13	55.15	81.66	75.70	93.74	48.10	80.57	70.92	73.05	74.70	82.82	93.76	48.65	76.52	71.21	72.22	66.67	85.20	91.07	53.03	72.35	72.05	71.18
LR	FSARFE	74.04	56.76	94.19	18.75	72.39	56.47	55.55	81.44	74.91	93.56	47.43	80.25	70.49	72.58	73.76	80.00	92.79	46.64	75.15	69.71	70.56	66.56	83.51	89.76	53.48	71.90	71.62	70.82
BAGLR	None	75.84	71.61	96.01	24.78	75.42	60.39	60.78	79.82	78.35	95.46	40.49	79.60	67.98	70.17	72.72	86.92	96.12	41.70	75.32	68.91	69.58	63.61	87.28	93.68	44.72	69.58	69.20	67.45
BAGLR	SP	72.58	51.58	95.83	10.94	71.29	53.38	50.32	81.49	74.74	93.47	47.65	80.25	70.56	72.63	74.37	83.46	94.17	47.53	76.35	70.85	71.84	65.98	85.13	91.29	51.46	71.68	71.37	70.37
BAGLR	FSASP	71.97	50.85	97.37	6.70	71.16	52.03	47.30	81.00	68.73	91.29	47.20	78.57	69.25	70.90	77.15	79.25	90.85	56.50	77.72	73.67	74.70	69.53	79.48	84.53	61.80	73.34	73.16	72.92
BAGLR	RFE	73.89	54.61	93.74	18.53	72.00	56.13	55.15	81.66	75.70	93.74	48.10	80.57	70.92	73.05	74.70	82.82	93.76	48.65	76.52	71.21	72.22	66.67	85.20	91.07	53.03	72.35	72.05	71.18
BAGLR	FSARFE	74.04	56.76	94.19	18.75	72.39	56.47	55.55	81.44	74.91	93.56	47.43	80.25	70.49	72.58	73.76	80.00	92.79	46.64	75.15	69.71	70.56	66.56	83.51	89.76	53.48	71.90	71.62	70.82
ADALR	None	75.00	70.15	96.37	20.98	74.58	58.68	58.33	81.06	76.69	94.37	45.64	80.31	70.01	72.22	75.25	83.83	94.04	50.00	77.21	72.02	73.12	66.40	85.93	91.72	52.13	72.23	71.93	70.97
ADALR	SP	71.81	80.95	99.64	3.79	71.94	51.72	45.36	81.04	73.14	93.10	46.31	79.60	69.71	71.68	74.18	83.94	94.45	46.86	76.26	70.66	71.62	65.98	85.13	91.29	51.46	71.68	71.37	70.37
ADALR	FSASP	72.62	51.00	95.55	11.38	71.23	53.47	50.57	82.49	69.53	90.65	52.57	79.66	71.61	73.13	78.60	79.12	90.15	60.31	78.75	75.23	76.21	70.30	80.87	85.62	62.70	74.34	74.16	73.92
ADALR	RFE	73.58	56.69	95.01	16.07	72.19	55.54	53.99	81.56	73.88	93.10	48.10	80.12	70.60	72.61	74.70	82.82	93.76	48.65	76.52	71.21	72.22	66.40	85.93	91.72	52.13	72.23	71.93	70.97
ADALR	FSARFE	73.72	58.27	95.19	16.52	72.45	55.85	54.41	82.05	74.09	92.92	49.89	80.50	71.41	73.39	74.44	80.07	92.51	48.65	75.75	70.58	71.51	66.56	84.04	90.20	53.26	72.01	71.73	70.90
MPCLR	None	86.05	34.64	33.58	86.61	48.90	60.09	48.90	92.60	61.26	78.31	84.56	80.12	81.44	77.96	84.34	78.33	87.38	73.77	82.18	80.57	80.91	74.26	84.07	87.36	68.76	78.21	78.06	77.96
MPCLR	SP	86.27	45.78	63.88	75.00	67.10	69.44	65.13	92.61	56.08	72.78	85.68	76.50	79.23	74.65	81.71	75.43	86.13	68.83	79.52	77.48	77.92	69.91	82.67	87.58	61.12	74.56	74.35	74.02
MPCLR	FSASP	72.64	67.80	98.28	8.93	72.45	53.60	49.66	85.29	68.55	88.38	62.42	80.89	75.40	76.07	78.45	78.59	89.88	60.09	78.49	74.98	75.94	70.66	81.45	86.06	63.15	74.78	74.60	74.37
MPCLR	RFE	87.88	29.27	2.63	99.11	30.52	50.87	25.15	91.46	61.00	78.77	81.88	79.66	80.32	77.28	82.14	77.25	87.38	69.28	80.46	78.33	78.86	70.85	82.84	87.36	62.92	75.33	75.14	74.88
MPCLR	FSARFE	81.88	55.09	81.58	55.58	74.06	68.58	68.53	86.98	69.43	87.93	67.56	82.05	77.75	77.97	79.83	80.52	90.57	63.00	80.03	76.79	77.78	67.80	85.47	90.85	55.51	73.45	73.18	72.48

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	61.27	91.43	92.57	57.53	72.27	75.05	72.18	64.88	93.08	90.08	73.24	79.21	81.66	78.71	68.20	93.16	84.57	84.21	84.31	84.39	81.99	75.18	93.37	80.30	91.27	88.56	85.79	84.98
LR	SP	63.45	87.58	87.62	63.37	73.57	75.49	73.57	68.85	91.53	86.78	78.46	81.41	82.62	80.64	75.42	90.76	77.14	89.93	86.27	83.54	83.31	77.34	91.85	75.00	92.77	88.37	83.88	84.23
LR	FSASP	63.11	82.30	80.50	65.84	72.01	73.17	71.95	62.05	85.79	77.69	73.92	75.26	75.80	74.20	59.62	87.97	72.57	80.32	78.10	76.45	74.72	70.00	91.35	74.24	89.53	85.74	81.88	81.24
LR	RFE	64.43	86.87	86.38	65.39	74.22	75.89	74.21	70.51	91.24	85.95	80.27	82.28	83.11	81.44	75.27	91.16	78.29	89.70	86.44	83.99	83.59	80.00	92.16	75.76	93.77	89.31	84.76	85.39
LR	FSARFE	64.73	86.94	86.38	65.84	74.48	76.11	74.47	67.92	88.97	82.23	78.68	79.94	80.46	78.95	69.07	90.19	76.57	86.27	83.50	81.42	80.41	72.13	89.29	66.67	91.52	85.37	79.09	79.84
BAGLR	None	61.19	91.10	92.26	57.53	72.14	74.89	72.05	64.50	93.04	90.08	72.79	78.92	81.44	78.43	68.20	93.16	84.57	84.21	84.31	84.39	81.99	75.35	93.61	81.06	91.27	88.74	86.17	85.26
BAGLR	SP	63.09	87.23	87.31	62.92	73.18	75.11	73.18	69.21	91.34	86.36	78.91	81.55	82.64	80.75	73.63	90.47	76.57	89.02	85.46	82.79	82.40	77.34	91.85	75.00	92.77	88.37	83.88	84.23
BAGLR	FSASP	62.86	82.02	80.19	65.62	71.74	72.90	71.69	61.49	86.10	78.51	73.02	74.96	75.76	73.99	62.56	89.28	75.43	81.92	80.07	78.68	76.92	70.42	91.82	75.76	89.53	86.12	82.64	81.82
BAGLR	RFE	64.43	86.87	86.38	65.39	74.22	75.89	74.21	70.51	91.24	85.95	80.27	82.28	83.11	81.44	75.27	91.16	78.29	89.70	86.44	83.99	83.59	80.00	92.16	75.76	93.77	89.31	84.76	85.39
BAGLR	FSARFE	64.73	86.94	86.38	65.84	74.48	76.11	74.47	67.92	88.97	82.23	78.68	79.94	80.46	78.95	69.07	90.19	76.57	86.27	83.50	81.42	80.41	72.13	89.29	66.67	91.52	85.37	79.09	79.84
ADALR	None	63.58	88.89	89.16	62.92	73.96	76.04	73.96	66.36	92.48	88.84	75.28	80.09	82.06	79.49	68.98	93.43	85.14	84.67	84.80	84.91	82.53	72.67	93.99	82.58	89.78	87.99	86.18	84.57
ADALR	SP	63.37	87.31	87.31	63.37	73.44	75.34	73.44	69.23	90.89	85.54	79.14	81.41	82.34	80.57	71.66	90.35	76.57	87.87	84.64	82.22	81.56	78.29	92.33	76.52	93.02	88.93	84.77	85.03
ADALR	FSASP	65.08	82.70	80.19	68.76	73.57	74.47	73.47	62.90	87.40	80.58	73.92	76.28	77.25	75.38	59.28	88.75	74.86	79.41	78.10	77.13	74.99	68.28	91.49	75.00	88.53	85.18	81.76	80.73
ADALR	RFE	63.51	87.35	87.31	63.60	73.57	75.45	73.57	69.54	91.60	86.78	79.14	81.84	82.96	81.06	73.40	91.27	78.86	88.56	85.78	83.71	82.96	78.63	92.79	78.03	93.02	89.31	85.52	85.61
ADALR	FSARFE	64.42	86.39	85.76	65.62	74.09	75.69	74.08	67.35	88.69	81.82	78.23	79.50	80.02	78.51	68.02	90.12	76.57	85.58	83.01	81.08	79.92	71.20	89.46	67.42	91.02	85.18	79.22	79.75
MPCLR	None	68.18	89.14	88.24	70.11	77.73	79.17	77.71	65.10	94.15	91.74	73.02	79.65	82.38	79.20	60.39	94.12	88.00	76.89	80.07	82.44	78.13	56.16	97.13	93.18	76.06	80.30	84.62	77.70
MPCLR	SP	62.69	88.93	89.47	61.35	73.18	75.41	73.17	60.48	94.53	92.98	66.67	75.99	79.82	75.74	50.62	95.55	92.57	63.84	72.06	78.21	71.00	51.65	97.59	94.70	70.82	76.74	82.76	74.46
MPCLR	FSASP	67.98	87.02	85.45	70.79	76.95	78.12	76.89	65.73	91.44	87.19	75.06	79.36	81.12	78.70	61.44	92.02	82.86	79.18	80.23	81.02	77.84	56.25	95.38	88.64	77.31	80.11	82.97	77.11
MPCLR	RFE	66.27	87.79	87.00	67.87	75.91	77.43	75.89	67.08	93.30	90.08	75.74	80.82	82.91	80.25	61.85	94.21	88.00	78.26	81.05	83.13	79.07	61.73	96.74	91.67	81.30	83.86	86.48	81.06
MPCLR	FSARFE	60.95	90.14	91.33	57.53	71.74	74.43	71.67	59.57	94.14	92.56	65.53	75.11	79.05	74.88	51.11	95.29	92.00	64.76	72.55	78.38	71.41	45.83	95.91	91.67	64.34	71.11	78.00	69.06

Fonte: Elaborada pela autora.

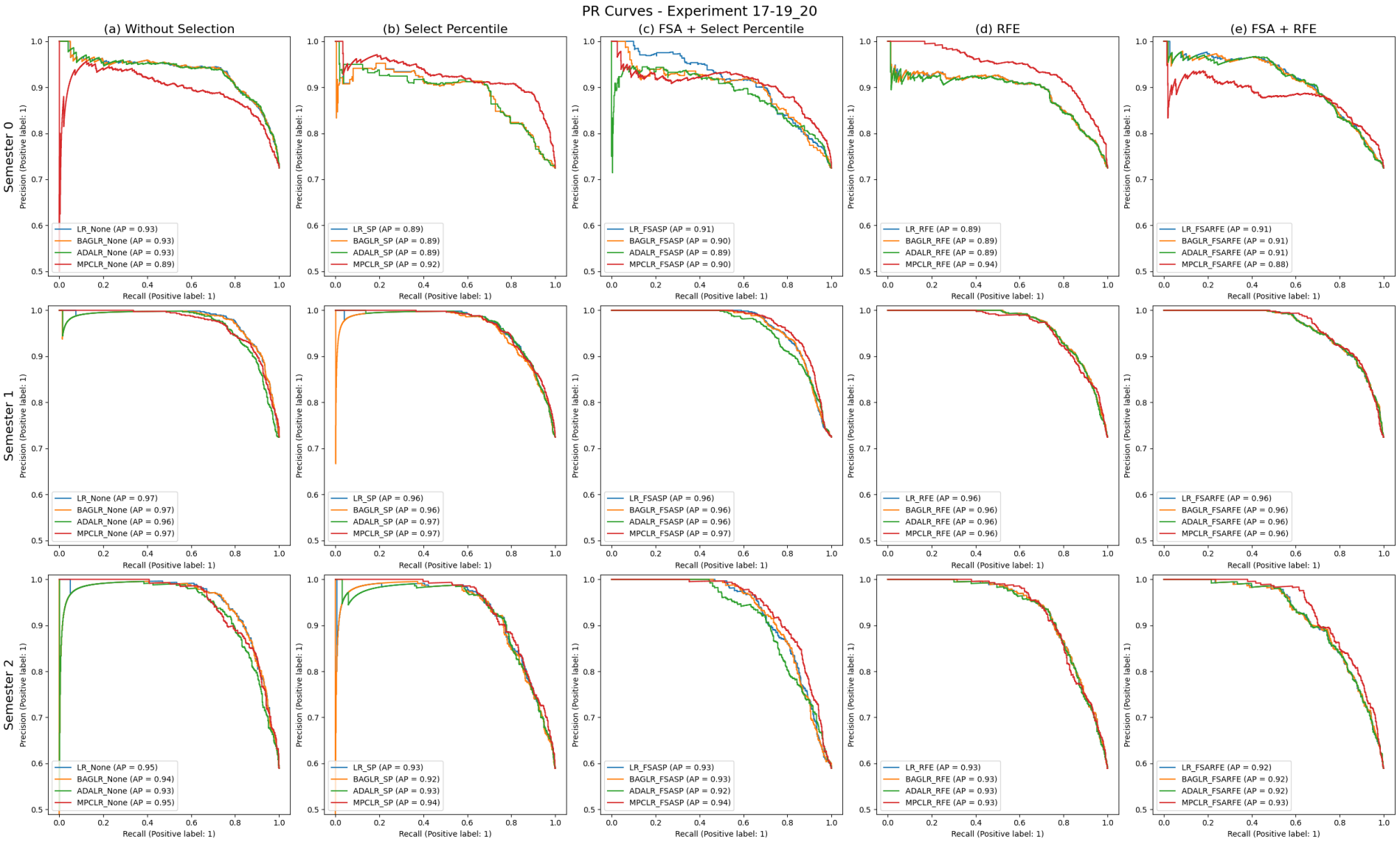
Tabela 21 – Resultados dos modelos baseados em LR no experimento 19-21_22, considerando os semestres de 0 a 7

		Semestre 0							Semestre 1							Semestre 2							Semestre 3						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	72.46	66.67	96.60	15.64	72.05	56.12	54.07	78.88	68.56	91.20	44.00	76.86	67.60	69.09	70.85	68.61	86.39	45.57	70.26	65.98	66.31	70.75	78.51	87.71	55.33	73.20	71.52	71.62
LR	SP	69.78	75.00	99.92	0.55	69.79	50.23	41.63	80.92	65.75	88.10	52.36	77.25	70.23	71.33	75.51	70.38	83.98	58.30	73.83	71.14	71.65	75.89	77.37	84.07	67.10	76.47	75.58	75.82
LR	FSASP	71.13	49.61	94.94	11.45	69.63	53.20	49.97	79.67	62.21	86.99	49.09	75.48	68.04	69.02	75.06	67.45	81.57	58.49	72.45	70.03	70.41	74.34	73.58	80.88	65.61	74.04	73.24	73.42
LR	RFE	71.43	49.03	93.75	13.82	69.51	53.78	51.32	80.72	66.98	88.98	51.27	77.53	70.12	71.37	75.96	72.71	85.66	58.49	74.93	72.07	72.67	76.14	77.19	83.76	67.66	76.55	75.71	75.94
LR	FSARFE	72.93	51.23	90.59	22.73	70.01	56.66	56.15	78.75	67.51	90.80	43.82	76.53	67.31	68.74	73.80	72.41	86.87	52.77	73.40	69.82	70.42	74.80	77.93	85.13	64.67	75.96	74.90	75.16
BAGLR	None	72.46	66.67	96.60	15.64	72.05	56.12	54.07	78.88	68.56	91.20	44.00	76.86	67.60	69.09	70.85	68.61	86.39	45.57	70.26	65.98	66.31	70.75	78.51	87.71	55.33	73.20	71.52	71.62
BAGLR	SP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	80.65	65.14	87.95	51.64	76.92	69.79	70.87	75.16	70.36	84.22	57.38	73.62	70.80	71.32	75.44	77.46	84.37	66.17	76.21	75.27	75.51
BAGLR	FSASP	69.80	33.33	96.36	4.18	68.41	50.27	44.19	81.01	61.57	85.25	54.18	75.81	69.72	70.36	75.20	64.44	78.19	60.52	71.21	69.35	69.54	75.65	72.82	79.21	68.60	74.46	73.90	74.02
BAGLR	RFE	71.27	50.38	94.78	12.18	69.74	53.48	50.49	80.72	67.55	89.29	51.09	77.69	70.19	71.48	76.06	73.38	86.14	58.49	75.22	72.32	72.94	75.89	77.37	84.07	67.10	76.47	75.58	75.82
BAGLR	FSARFE	72.93	51.23	90.59	22.73	70.01	56.66	56.15	78.71	67.61	90.88	43.64	76.53	67.26	68.70	73.75	72.52	86.99	52.58	73.40	69.79	70.39	74.80	77.93	85.13	64.67	75.96	74.90	75.16
ADALR	None	72.13	65.81	96.84	14.00	71.72	55.42	52.88	78.43	65.83	90.25	43.09	75.92	66.67	68.01	69.88	68.45	87.23	42.44	69.53	64.83	65.00	69.81	77.87	87.71	53.27	72.28	70.49	70.50
ADALR	SP	69.68	0.00	100.0	0.00	69.68	50.00	41.07	80.49	65.57	88.34	50.91	76.97	69.63	70.78	75.19	70.52	84.34	57.38	73.69	70.86	71.39	74.23	76.85	84.37	63.93	75.21	74.15	74.39
ADALR	FSASP	70.56	41.86	94.07	9.82	68.52	51.94	48.27	79.63	60.09	85.57	49.82	74.71	67.69	68.48	73.35	63.47	78.92	56.09	69.90	67.50	67.79	73.45	71.40	78.91	64.86	72.61	71.88	72.03
ADALR	RFE	70.88	48.25	95.33	10.00	69.46	52.67	48.94	80.96	66.74	88.66	52.18	77.58	70.42	71.60	75.58	73.00	86.14	57.38	74.78	71.76	72.39	75.03	76.81	83.92	65.61	75.71	74.76	75.00
ADALR	FSARFE	72.52	50.91	91.46	20.36	69.90	55.91	54.99	79.14	64.80	89.06	46.18	76.04	67.62	68.87	73.83	70.90	85.66	53.51	72.96	69.58	70.15	74.07	77.40	84.98	63.36	75.29	74.17	74.42
MPCLR	None	71.52	60.58	96.76	11.45	70.89	54.11	50.76	85.54	57.46	77.40	70.00	75.15	73.70	72.19	79.75	66.43	76.87	70.11	74.20	73.49	73.25	79.29	78.95	84.22	72.90	79.15	78.56	78.74
MPCLR	SP	74.98	51.81	86.31	33.82	70.40	60.07	60.59	87.84	52.57	69.31	78.00	71.95	73.66	70.15	84.56	62.27	67.95	81.00	73.10	74.47	72.88	83.52	75.92	79.21	80.75	79.90	79.98	79.78
MPCLR	FSASP	70.35	66.67	99.13	4.00	70.29	51.56	44.92	82.06	60.49	83.43	58.18	75.76	70.80	71.03	78.58	65.03	76.02	68.27	72.96	72.14	71.94	77.29	74.75	80.58	70.84	76.21	75.71	75.82
MPCLR	RFE	69.72	50.00	99.84	0.36	69.68	50.10	41.42	87.14	56.79	75.26	74.55	75.04	74.90	72.62	80.43	66.72	76.75	71.40	74.64	74.07	73.76	78.45	77.31	82.85	71.96	77.97	77.41	77.57
MPCLR	FSARFE	100.0	30.37	0.24	100.0	30.49	50.12	23.53	92.77	49.35	60.03	89.27	68.91	74.65	68.23	88.87	59.97	61.57	88.19	72.08	74.88	72.07	86.03	74.63	76.63	84.67	80.23	80.65	80.20

		Semestre 4							Semestre 5							Semestre 6							Semestre 7						
Modelo	Seleção	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1	Precisão		Revocação		Acur.	UAR	F1
		Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não				Sim	Não	Sim	Não			
LR	None	63.97	81.45	81.88	63.30	71.53	72.59	71.53	64.64	86.99	82.88	71.62	75.95	77.25	75.59	66.12	91.11	83.06	80.31	81.18	81.68	79.50	67.35	92.86	84.18	83.44	83.65	83.81	81.36
LR	SP	71.58	81.67	78.82	75.09	76.75	76.96	76.64	70.18	86.21	79.88	78.76	79.19	79.32	78.52	74.69	88.65	75.62	88.15	84.18	81.88	81.78	79.06	90.78	77.04	91.72	87.48	84.38	84.64
LR	FSASP	67.30	78.42	75.53	70.79	72.89	73.16	72.79	66.23	83.44	75.98	75.75	75.84	75.86	75.09	68.75	87.03	72.73	84.70	80.92	78.72	78.27	75.86	91.18	78.57	89.86	86.60	84.21	83.85
LR	RFE	72.07	82.24	79.53	75.47	77.27	77.50	77.16	72.05	86.00	78.98	80.83	80.12	79.90	79.35	75.20	89.51	77.69	88.15	84.84	82.92	82.62	74.76	91.12	78.57	89.23	86.16	83.90	83.39
LR	FSARFE	71.01	81.99	79.53	74.16	76.54	76.84	76.45	72.86	84.85	76.58	82.14	80.00	79.36	79.07	76.92	88.32	74.38	89.67	84.84	82.03	82.31	79.69	91.17	78.06	91.93	87.92	84.99	85.21
BAGLR	None	63.97	81.45	81.88	63.30	71.53	72.59	71.53	64.64	86.99	82.88	71.62	75.95	77.25	75.59	66.12	91.11	83.06	80.31	81.18	81.68	79.50	67.35	92.86	84.18	83.44	83.65	83.81	81.36
BAGLR	SP	71.49	81.80	79.06	74.91	76.75	76.98	76.64	69.90	86.34	80.18	78.38	79.08	79.28	78.43	76.35	88.93	76.03	89.10	84.97	82.57	82.60	80.42	91.02	77.55	92.34	88.07	84.95	85.32
BAGLR	FSASP	69.11	77.60	73.18	73.97	73.62	73.57	73.41	72.62	83.18	73.27	82.71	79.08	77.99	77.94	73.84	87.31	72.31	88.15	83.14	80.23	80.40	76.53	90.48	76.53	90.48	86.45	83.50	83.50
BAGLR	RFE	71.52	82.27	79.76	74.72	76.96	77.24	76.86	71.12	86.35	79.88	79.70	79.77	79.79	79.07	74.02	89.43	77.69	87.38	84.31	82.53	82.10	75.36	92.09	81.12	89.23	86.89	85.18	84.39
BAGLR	FSARFE	71.01	81.99	79.53	74.16	76.54	76.84	76.45	72.65	84.82	76.58	81.95	79.88	79.27	78.96	76.92	88.32	74.38	89.67	84.84	82.03	82.31	79.69	91.17	78.06	91.93	87.92	84.99	85.21
ADALR	None	64.78	82.97	83.53	63.86	72.58	73.69	72.57	64.92	86.32	81.68	72.37	75.95	77.03	75.54	65.78	90.52	81.82	80.31	80.78	81.06	79.02	68.38	91.91	81.63	84.68	83.80	83.16	81.28
ADALR	SP	70.33	81.97	79.76	73.22	76.12	76.49	76.05	69.82	86.16	79.88	78.38	78.96	79.13	78.30	75.20	89.02	76.45	88.34	84.58	82.39	82.25	77.49	90.16	75.51	91.10	86.60	83.30	83.56
ADALR	FSASP	65.96	76.75	73.41	69.85	71.43	71.63	71.31	68.64	82.39	72.97	79.14	76.76	76.05	75.74	71.66	87.45	73.14	86.62	82.35	79.88	79.71	76.77	90.85	77.55	90.48	86.75	84.01	83.91
ADALR	RFE	71.16	82.02	79.53	74.34	76.64	76.94	76.55	70.89	85.83	78.98	79.70	79.42	79.34	78.68	73.81	89.08	76.86	87.38	84.05	82.12	81.76	73.08	90.66	77.55	88.41	85.27	82.98	82.38
ADALR	FSARFE	69.87	81.08	78.59	73.03	75.50	75.81	75.41	71.47	84.34	75.98	81.02	79.08	78.50	78.15	73.06	87.88	73.97	87.38	83.14	80.67	80.57	76.72	89.59	73.98	90.89	86.01	82.43	82.78
MPCLR	None	71.73	83.26	81.18	74.53	77.48	77.85	77.41	66.34	86.24	81.08	74.25	76.88	77.66	76.39	61.68	91.65	85.12	75.53	78.56	80.32	77.17	61.87	94.01	87.76	78.05	80.85	82.90	78.93
MPCLR	SP	76.37	80.56	75.29	81.46	78.73	78.38	78.42	71.93	86.14	79.28	80.64	80.12	79.96	79.36	72.06	90.67	80.99	85.47	84.05	83.23	82.13	69.48	94.65	88.27	84.27	85.42	86.27	83.45
MPCLR	FSASP	68.99	79.79	76.94	72.47	74.45	74.71	74.35	67.24	87.28	82.58	74.81	77.80	78.70	77.35	62.93	90.99	83.47	77.25	79.22	80.36	77.66	64.75	93.54	86.22	80.95	82.47	83.59	80.38
MPCLR	RFE	71.19	82.50	80.24	74.16	76.85	77.20	76.77	66.43	87.50	83.18	73.68	77.34	78.43	76.93	65.48	91.43	83.88	79.54	80.92	81.71	79.31	60.90	94.87	89.80	76.60	80.41	83.20	78.67
MPCLR	FSARFE	78.05	79.93	73.65	83.52	79.14	78.58	78.74	67.75	86.67	81.38	75.75	77.92	78.57	77.39	59.65	91.63	85.54	73.23	77.12	79.38	75.85	55.97	95.01	90.82	71.01	76.73	80.92	75.27

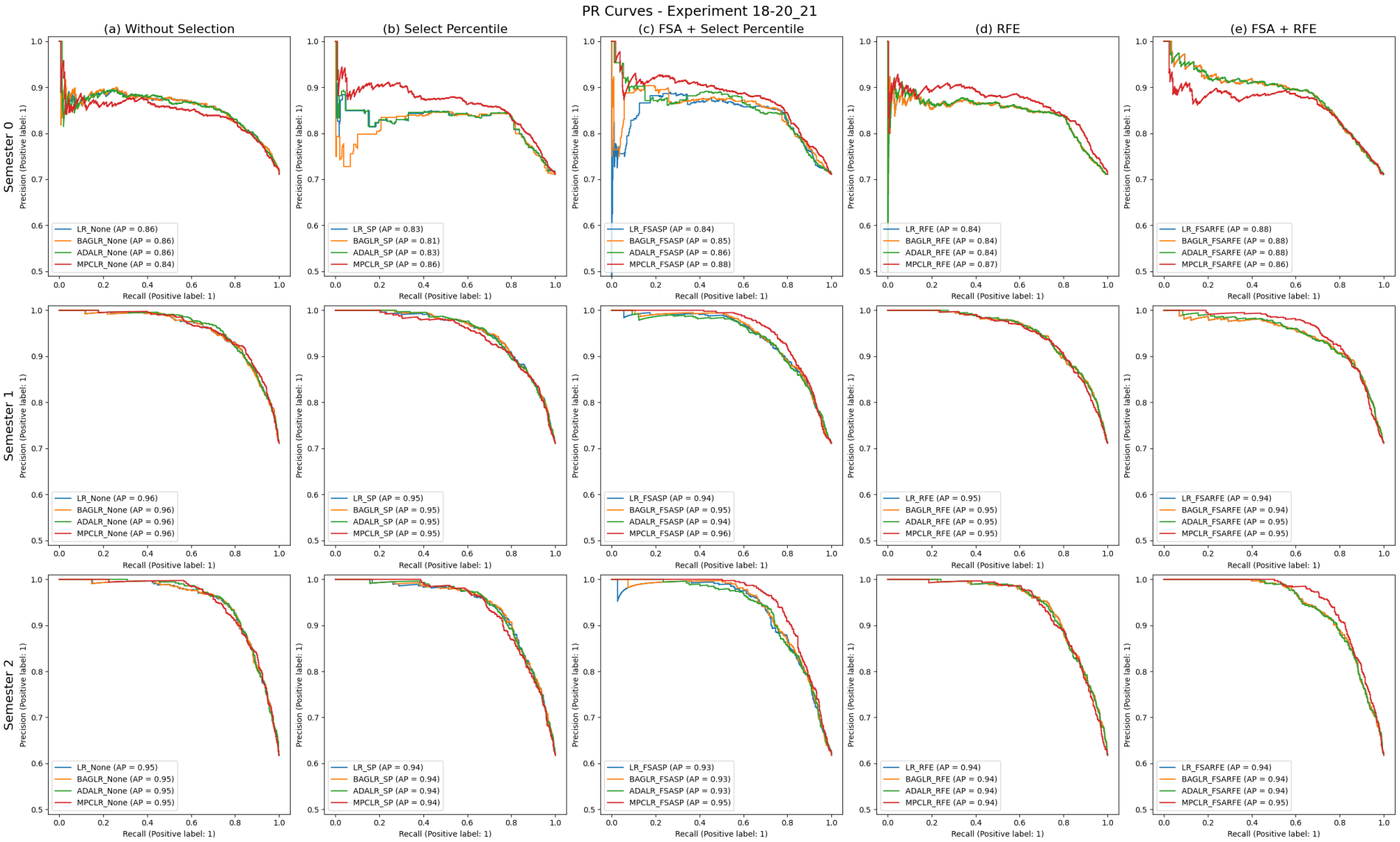
Fonte: Elaborada pela autora.

Figura 28 – Curvas PR dos modelos baseados em LR, para o experimento 17-19_20 e os semestres de 0 a 2



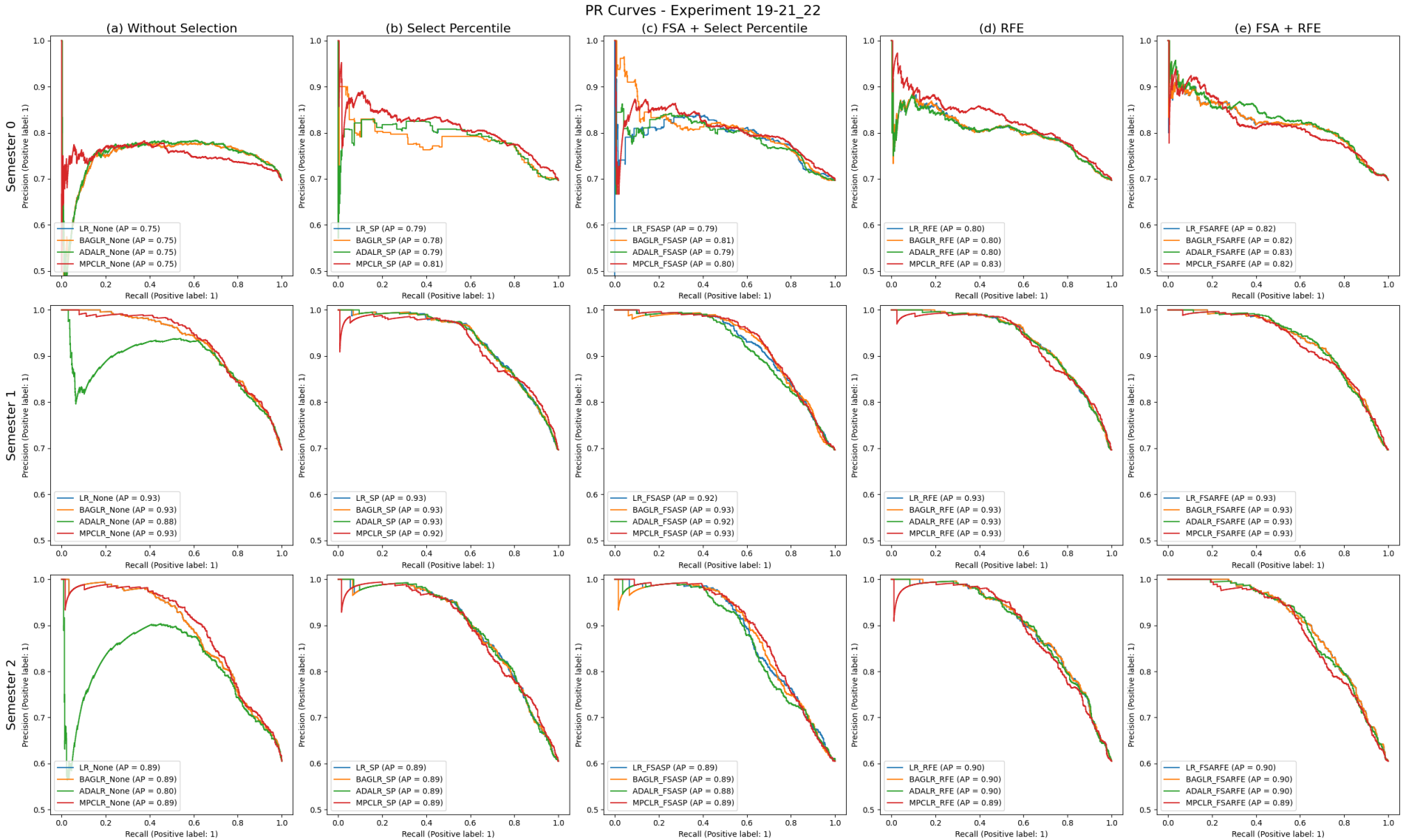
Fonte: Elaborada pela autora.

Figura 29 – Curvas PR dos modelos baseados em LR, para o experimento 18-20_21 e os semestres de 0 a 2



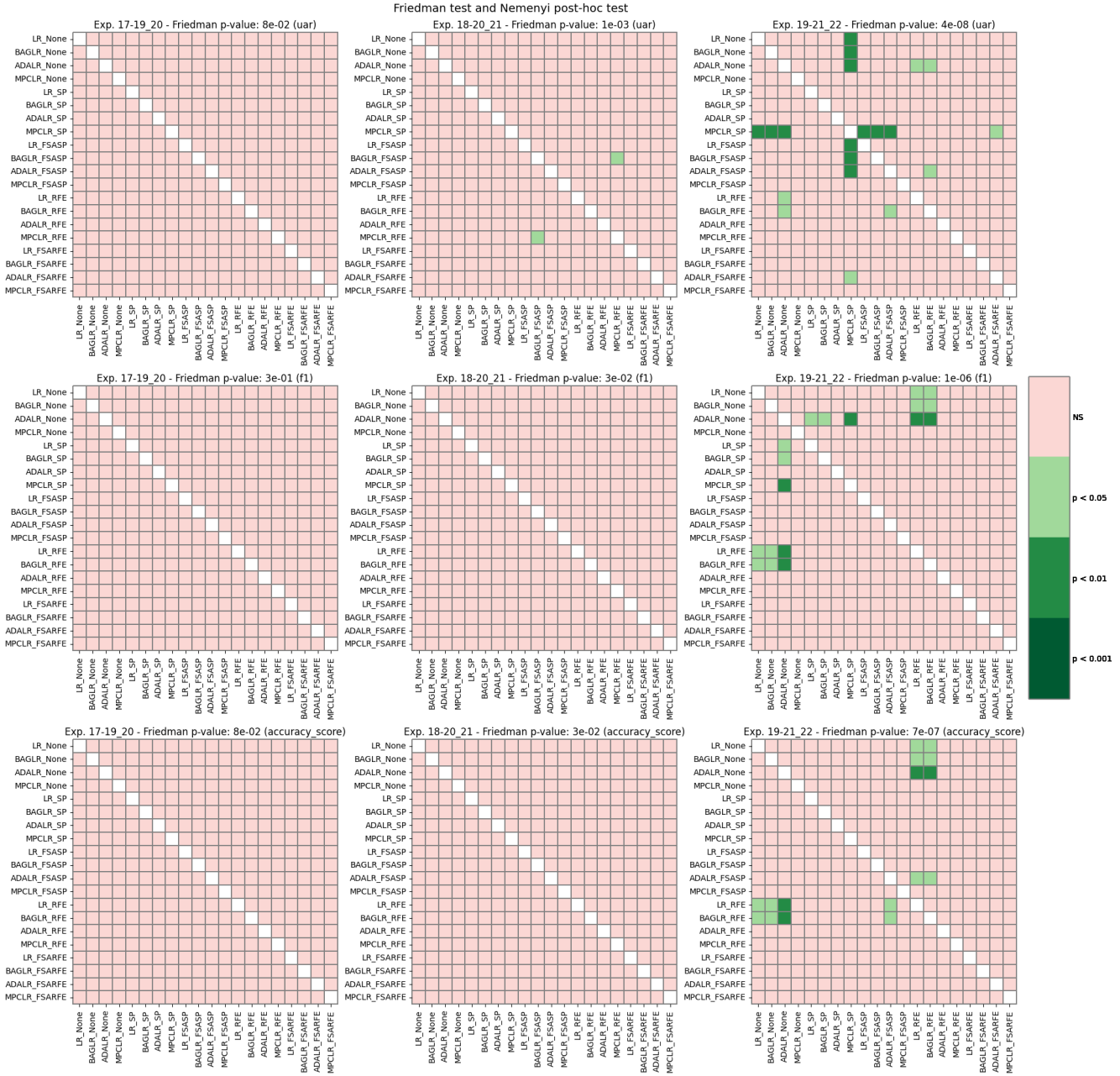
Fonte: Elaborada pela autora.

Figura 30 – Curvas PR dos modelos baseados em LR, para o experimento 19-21_22 e os semestres de 0 a 2



Fonte: Elaborada pela autora.

Figura 31 – Testes estatísticos sobre os desempenhos de UAR, F1, e acurácia dos modelos baseados em LR, para cada experimento



Fonte: Elaborada pela autora.

Resumidamente, os testes estatísticos indicaram superioridade dos modelos MPC-SDP com seleções RFE e SP sobre *baselines*, nos experimentos 18-20_21 e 19-21_22, respectivamente, além de terem evidenciado que, especialmente no experimento 19-21_22, o desempenho dos *baselines* foi beneficiado pela seleção RFE e prejudicado pela ausência de seleção de atributos e pelo uso da seleção FSA+SP. Considerando que não foi apontada diferença significativa entre os desempenhos dos modelos MPC-SDP com diferentes métodos de seleção de atributos, pode-se dizer que as seleções FSA não afetaram significativamente os modelos MPC-SDP baseados em LR, mostrando-se, então, boas opções de seleção, dado o possível ganho na variabilidade dos padrões preditivos. Inclusive, na Tabela 21, é possível observar que no experimento 19-21_22, que inclui padrões influenciados/modificados pela vivência do ERE e, portanto, considera um contexto mais próximo do atual, o MPC-SDP com seleção FSA+RFE apresentou, em geral, os melhores resultados de precisão positiva e UAR, durante os três primeiros semestres da trajetória acadêmica dos estudantes. Como os testes estatísticos se basearam nos desempenhos de primeiro a sétimo semestre e o MPC-SDP com seleção FSA+RFE não apresentou os melhores resultados a partir do quarto semestre, esse classificador não recebeu destaque. Porém, ainda assim, poderia ser considerado o melhor modelo preditivo, para uma aplicação prática.

7.2 Padrões

7.2.1 Modelos baseados em DT

Nas Tabelas 22, 23, 24, e 25 são apresentadas as três principais regras determinantes de evasão, em termos de abrangência, extraídas dos subclassificadores DT dos modelos MPC-SDP, com seleção de atributos e desenvolvidos nos experimentos 17-19_20 e 19-21_22[11]. Optou-se por apresentar as regras desses experimentos, pois eles envolvem dados de treinamento relacionados a estudantes evadidos e concluídos em contextos mais distantes/distintos, de antes e durante o ERE[12], respectivamente. As tabelas foram organizadas para facilitar a comparação entre as regras dos modelos vinculados aos métodos de seleção tradicionais e às respectivas seleções FSA. Dessa forma, as Tabelas 22 e 23 abrangem as principais regras dos modelos desenvolvidos com as seleções SP e FSA+SP, para os experimentos 17-19_20 e 19-21_22, respectivamente. Nas Tabelas 24 e 25 essa organização se mantém, porém, considerando dados dos modelos desenvolvidos a partir das seleções RFE e FSA+RFE.

[11]Quando são apresentadas menos de três regras, significa que o subclassificador não possui três regras determinantes de evasão que satisfaçam os requisitos de extração, descritos ao fim do Capítulo 6.

[12]Enquanto o experimento 17-19_20 só possui dados de treinamento relacionados a estudantes evadidos e concluídos antes da pandemia de COVID-19, o experimento 19-21_22 envolve o maior quantitativo de dados de estudantes evadidos e concluídos durante o contexto pandêmico (anos de 2020 e 2021).

Nas tabelas, estão assinaladas com o termo 'SS' (subárvores suprimidas) as regras simplificadas/unificadas a partir de um nó cujas subárvores só continham folhas indicadoras de evasão. Porém, caso as subárvores suprimidas contemplassem atributos de aspectos ainda não utilizados pela regra simplificada, esses atributos foram listados, entre colchetes, junto ao termo 'SS', indicando que eles fazem parte das regras unificadas e, dessa forma, contribuem para a variabilidade de aspectos entre os padrões preditivos. Para facilitar a visualização dessa variabilidade, também foram atribuídas cores distintas para as condições/atributos que compõem cada regra, segundo os aspectos representados (acadêmico, contextual, econômico, social, e interacional). Além disso, ao final de cada regra, é apresentada sua probabilidade e abrangência, que correspondem, respectivamente, à proporção de amostras/registros de estudantes evadidos que confirmam a regra, e à quantidade de amostras/registros (evadidos ou não) que satisfazem a regra.

Na Tabela 22, é possível observar que, no experimento 17-19_20, o modelo MPC-SDP com seleção FSA+SP apresentou maior variabilidade de aspectos nas principais regras de seus subclassificadores. Mais especificamente, no subclassificador especializado no semestre 1, a seleção FSA+SP propiciou a inclusão de atributos dos aspectos contextual (verde) e social/demográfico (laranja) entre as principais regras preditivas. Nos subclassificadores especializados em dados de segundo, quarto, e quinto semestre, foi adicionado o aspecto interacional (azul). Por fim, no subclassificador especializado em dados do terceiro semestre, foi contemplado, adicionalmente, o aspecto econômico (roxo). Embora menos frequente, na Tabela 23, relacionada ao experimento 19-20_21, também é possível observar um aumento na variabilidade dos aspectos dos atributos priorizados pelo modelo MPC-SDP com seleção FSA+SP.

De forma geral, as regras das Tabelas 24 e 25 também demonstram que a seleção FSA+RFE favoreceu a variabilidade de aspectos entre os principais padrões preditivos, tanto no experimento 17-19_20 quanto no 19-20_21. Porém, nos subclassificadores especializados no quinto semestre do experimento 17-19_20 (Tabela 24), e no primeiro e sétimo semestre do experimento 19-20_21 (Tabela 25), a situação foi oposta, sendo que um aspecto foi omitido/deixou de aparecer entre as principais regras preditivas, com o uso da seleção FSA+RFE. Ou seja, embora a FSA+RFE tenha forçado a seleção de mais atributos de diferente aspectos, esses subclassificadores não escolheram, para participar de seus principais padrões preditivos de evasão, atributos de aspectos mais variados do que os escolhidos pelos subclassificadores do modelo MPC-SDP com seleção RFE. Por implementar a técnica de *wrapper*, que testa a influência dos atributos em modelos de aprendizado de máquina, e por avaliar a totalidade dos atributos, sem segmentá-los por aspecto, a RFE pode antecipar a importância de atributos que melhor se complementam, na composição dos padrões preditivos. Ou seja, mesmo recuperando uma variabilidade menor de aspectos, em

alguns casos, os atributos de diferentes aspectos selecionados pela RFE podem demonstrar maior capacidade preditiva, ao serem combinados pelo subclassificador.

Por meio de uma análise geral, também se percebe que as principais regras dos subclassificadores especializados no momento de admissão (semestre 0) se concentram em atributos contextuais e sociais/demográficos, o que é natural, já que os atributos dos demais aspectos não possuem informação nesse estágio acadêmico. Nesse contexto, as regras mais abrangentes relacionam geralmente a evasão a estudantes que ingressaram por vagas de ampla concorrência[13]; que pertençam à área de Ciências Exatas e da Terra ou não pertençam à área de Ciências Sociais Aplicadas; e que não estudem nos *Campi* Frederico Westphalen, Santa Rosa, e São Vicente do Sul. De forma geral, esses padrões vão ao encontro das descrições estatísticas dos dados, vistas na Seção 5.2.5. Também há uma convergência com a literatura, já que estudos como Chen; Johri; Rangwala (2018); Casanova et al. (2023) apontam maiores dificuldades de adaptação e risco de evasão entre estudantes de cursos das áreas STEM (mais diretamente relacionados à área de Ciências Exatas e da Terra, no Brasil), que, geralmente, demandam um *background* educacional mais forte/rico.

Entre os principais padrões de evasão dos subclassificadores especializados em dados de primeiro e segundo semestre, há a predominância de atributos acadêmicos e interacionais. Porém, na primeira e mais representativa regra desses subclassificadores, os atributos interacionais não costumam ser considerados ou aparecem suprimidamente, compondo o caminho da classificação, mas sem interferir no resultado (abaixo de seus nós, todas as decisões são de evasão)[14]. Essas regras, em geral, associam a evasão a estudantes com menores médias de notas, médias de frequências e porcentagens de carga horária integralizada, no geral ou no semestre, ou a estudantes que não se sobressaem ou se sobressaem menos em relação à média de seus pares nesses atributos (desvios de média). Nas outras (segunda e terceira) principais regras dos subclassificadores de primeiro e segundo semestre, é possível observar que os atributos interacionais mais frequentes estão vinculados às médias diárias de ações e acessos, e que as condições desses atributos consideram limiares inferiores ou superiores, dependendo das condições acadêmicas associadas.

A segunda principal regra de evasão do subclassificador especializado em segundo

[13]Note que como um atributo binário/booleano só possui valores 0 e 1, "<= 0.5" indica que o atributo possui valor 0 (de negação/falso), enquanto "> 0.5" corresponde ao valor 1 (verdadeiro). Logo, a condição "s_reserva_vaga_ACG > 0.5" indica que a regra envolve estudantes que ingressaram por reserva de vaga de ampla concorrência.

[14]Apenas o primeiro e o segundo subclassificador dos modelos MPC-SDP com seleção FSA+RFE e RFE, respectivamente, apresentaram, no experimento 19-21_22, atributos interacionais na regra mais representativa. Porém, ao investigar a totalidade das regras desses subclassificadores, foi possível identificar que as subárvores associadas à negação das condições desses atributos interacionais também levavam a uma quantidade majoritária de registros evadidos. Ou seja, embora esses subclassificadores tenham decido utilizar condições/divisões relacionadas a atributos interacionais, essas condições/divisões tem pouca influência no principal padrão preditivo.

Tabela 22 – Principais regras de evasão de cada subclassificador DT dos modelos MPC-SDP, desenvolvidos no experimento 17-19_20, com as seleções SP e FSA+SP

Experimento 17-19_20 – MPCDT_SP – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) - SS - 87.77%/368
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) - SS - 87.77%/368
Experimento 17-19_20 – MPCDT_SP – Subclassificador 1
SE (a_media_notas_us <= 6.525) E (a_frequencia_media <= 95.17) - SS [i_media_diaria_acessos] - 97.0%/433
SE (a_media_notas_us > 6.525) E (i_media_diaria_acoes > 0.085) E (a_dvm_media_notas <= 2.5) - SS - 85.12%/121
SE (a_media_notas_us > 6.525) E (i_media_diaria_acoes <= 0.085) E (a_dvm_media_notas <= 2.48) E (i_dvm_media_diaria_acessos <= -0.082) E (a_frequencia_media <= 98.88) - 62.9%/62
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 1
SE (a_media_notas_us <= 6.525) - SS [s_anos_em_ingresso, s_reserva_vaga_ACG, i_media_diaria_acessos, c_forma_ingresso_Processo Seletivo, i_media_diaria_acoes] - 94.06%/488
SE (a_media_notas_us > 6.525) E (i_media_diaria_acoes > 0.085) E (a_media_notas <= 7.845) - SS - 85.12%/121
SE (a_media_notas_us > 6.525) E (i_media_diaria_acoes > 0.085) E (a_media_notas > 7.845) E (s_reserva_vaga_ACG > 0.5) - 70.59%/51
Experimento 17-19_20 – MPCDT_SP – Subclassificador 2
SE (a_media_notas_us <= 6.69) E (a_dvm_pct_ch_integralizada_us <= -4.678) E (a_dvm_pct_ch_integralizada <= -2.097) - 100.0%/354
SE (a_media_notas_us <= 6.69) E (a_dvm_pct_ch_integralizada_us > -4.678) E (a_pct_exames > 6.905) E (a_pct_reprovacoes > 17.16) - 100.0%/95
SE (a_media_notas_us <= 6.69) E (a_dvm_pct_ch_integralizada_us > -4.678) E (a_pct_exames > 6.905) E (a_pct_reprovacoes <= 17.16) E (a_media_freq_us <= 95.635) - 90.48%/42
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 2
SE (a_media_notas_us <= 6.92) E (a_dvm_pct_ch_integralizada_us <= 0.407) - SS - 97.41%/502
SE (a_media_notas_us <= 6.92) E (a_dvm_pct_ch_integralizada_us > 0.407) E (i_dif_media_diaria_acessos > 0.005) - 76.25%/80
SE (a_media_notas_us > 6.92) E (a_pct_ch_integralizada <= 19.375) E (a_pct_reprovacoes > 5.44) - 67.92%/53
Experimento 17-19_20 – MPCDT_SP – Subclassificador 3
SE (a_media_notas_us <= 6.685) E (a_pct_ch_integralizada_us <= 11.2) - SS - 97.38%/572
SE (a_media_notas_us > 6.685) E (a_pct_reprovacoes <= 5.13) E (a_media_freq_us <= 93.87) E (a_dvm_pct_ch_integralizada <= 3.189) - 62.07%/29
SE (a_media_notas_us <= 6.685) E (a_pct_ch_integralizada_us > 11.2) E (a_dif_freq <= -0.33) - 70.83%/24
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 3
SE (a_media_notas_us <= 6.685) E (a_pct_ch_integralizada_us <= 11.2) - SS [e_renda_familiar] - 97.38%/572
SE (a_media_notas_us > 6.685) E (a_pct_reprovacoes > 5.13) E (a_pct_ch_integralizada_us > 9.22) E (e_renda_familiar > 1150.0) - 64.71%/34
SE (a_media_notas_us > 6.685) E (a_pct_reprovacoes > 5.13) E (a_pct_ch_integralizada_us <= 9.22) - 83.33%/30
Experimento 17-19_20 – MPCDT_SP – Subclassificador 4
SE (a_pct_reprovacoes > 15.5) - SS - 98.44%/448
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dif_media_notas > -1.445) E (a_media_freq_us <= 94.135) - 63.53%/85
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dif_media_notas <= -1.445) - 93.75%/32
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 4
SE (a_pct_reprovacoes > 15.5) - SS [i_media_diaria_acessos_us, i_dif_media_diaria_acessos, i_media_diaria_acoes_us] - 98.44%/448
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dvm_pct_ch_integralizada_us > -4.064) E (a_media_notas_us <= 6.865) - 62.5%/56
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dvm_pct_ch_integralizada_us <= -4.064) - 88.89%/36
Experimento 17-19_20 – MPCDT_SP – Subclassificador 5
SE (a_dvm_pct_ch_integralizada_us <= -1.316) - SS - 95.53%/358
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 5
SE (a_pct_ch_integralizada_us <= 8.345) - SS [i_dif_media_diaria_acessos] - 95.53%/358
SE (a_pct_ch_integralizada_us > 8.345) E (a_pct_reprovacoes > 2.53) E (a_dvm_media_notas <= 0.918) E (a_iech <= 0.852) - 76.92%/26
SE (a_pct_ch_integralizada_us > 8.345) E (a_pct_reprovacoes > 2.53) E (a_dvm_media_notas <= 0.918) E (a_iech > 0.852) E (0.105 < i_dif_media_diaria_acoes <= 0.315) - 68.0%/25
Experimento 17-19_20 – MPCDT_SP – Subclassificador 6
SE (a_media_notas_us <= 7.315) E (a_dvm_pct_ch_integralizada <= -3.653) - SS - 97.6%/459
SE (a_media_notas_us <= 7.315) E (a_dvm_pct_ch_integralizada > -3.653) E (a_dvm_pct_ch_integralizada_us <= -5.653) - SS - 89.19%/37
SE (a_media_notas_us > 7.315) E (a_dvm_pct_ch_integralizada <= -2.176) E (a_pct_ch_integralizada_us <= 9.535) E (a_iech <= 0.933) - 100.0%/19
Experimento 17-19_20 – MPCDT_FSASP – Subclassificador 6
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 97.92) - SS - 93.46%/535
SE (a_media_notas_us > 7.315) E (a_iech <= 0.806) - 93.75%/16
SE (a_media_notas_us > 7.315) E (a_iech > 0.806) E (a_pct_ch_integralizada <= 65.14) E (a_pct_ch_integralizada_us <= 9.535) - 75.0%/12

Fonte: Elaborada pela autora.

Tabela 23 – Principais regras de evasão de cada subclassificador DT dos modelos MPC-SDP, desenvolvidos no experimento 19-21_22, com as seleções SP e FSA+SP

Experimento 19-21_22 – MPCDT_SP – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) E (c_area_Ciências Exatas e da Terra > 0.5) - SS - 83.94%/492
SE (s_reserva_vaga_ACG > 0.5) E (c_area_Ciências Exatas e da Terra <= 0.5) E (c_unidade_Alegrete <= 0.5) E (s_dvm_idade_ingresso > -5.5) - SS - 70.83%/432
SE (s_reserva_vaga_ACG > 0.5) E (c_area_Ciências Exatas e da Terra <= 0.5) E (c_unidade_Alegrete > 0.5) - SS - 95.5%/111
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) - SS [c_area_Ciências Exatas e da Terra] - 76.65%/1169
Experimento 19-21_22 – MPCDT_SP – Subclassificador 1
SE (a_media_notas_us <= 5.72) E (a_dvm_pct_ch_integralizada_us <= -3.757) E (s_reserva_vaga_ACG > 0.5) - 100.0%/467
SE (a_media_notas_us <= 6.69) E (-3.757 < a_dvm_pct_ch_integralizada_us <= 24.108) E (i_media_diaria_acoes > 0.085) E (a_pct_reprovacoes > 5.28) E (a_mcn <= 1169.645) E (i_media_diaria_acoes_us > 0.19) - 98.1%/158
SE (a_media_notas_us > 6.69) E (2.275 < i_pct_questionarios_respondidos <= 97.435) E (i_media_diaria_acoes > 0.505) - 100.0%/95
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 1
SE (a_media_notas_us <= 6.525) E (a_dvm_pct_ch_integralizada_us <= -6.317) - SS [s_reserva_vaga_ACG, i_pct_geral_respostas] - 99.05%/529
SE (a_media_notas_us <= 6.525) E (a_dvm_pct_ch_integralizada_us > -6.317) E (i_pct_questionarios_respondidos > -0.5) - SS - 95.14%/144
SE (a_media_notas_us > 6.525) E (s_reserva_vaga_ACG > 0.5) E (i_pct_geral_respostas > 1.925) - SS - 91.15%/113
Experimento 19-21_22 – MPCDT_SP – Subclassificador 2
SE (a_dvm_pct_ch_integralizada_us <= -1.491) E (a_media_notas_us <= 4.66) - SS [i_pct_geral_respostas, i_dif_media_diaria_acoes] - 97.83%/738
SE (a_dvm_pct_ch_integralizada_us > -1.491) E (a_pct_reprovacoes > 12.77) E (s_reserva_vaga_ACG > 0.5) - SS - 83.52%/91
SE (a_dvm_pct_ch_integralizada_us > -1.491) E (a_pct_reprovacoes > 12.77) E (s_reserva_vaga_ACG <= 0.5) E (-0.586 < a_dvm_frequencia_media <= 8.684) - SS - 72.06%/68
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 2
SE (a_pct_ch_integralizada_us <= 6.365) E (a_media_notas_us <= 4.66) - SS [i_pct_geral_respostas, i_dif_media_diaria_acoes, s_dvm_idade_ingresso] - 97.83%/738
SE (a_pct_ch_integralizada_us > 6.365) E (a_pct_reprovacoes > 12.77) E (s_reserva_vaga_ACG > 0.5) - SS [i_dif_media_diaria_acessos] - 83.52%/91
SE (a_pct_ch_integralizada_us <= 6.365) E (a_media_notas_us > 4.66) E (i_dif_media_diaria_acessos <= -0.005) - SS - 96.15%/52
Experimento 19-21_22 – MPCDT_SP – Subclassificador 3
SE (a_dvm_pct_ch_integralizada_us <= -2.364) - SS [i_dif_media_diaria_acessos] - 95.58%/860
SE (a_dvm_pct_ch_integralizada_us > -2.364) E (a_pct_reprovacoes > 3.51) E (a_stddev_media_notas > 1.905) - SS [i_dif_media_diaria_acoes] - 61.99%/321
SE (a_dvm_pct_ch_integralizada_us > -2.364) E (a_pct_reprovacoes > 3.51) E (a_stddev_media_notas <= 1.905) E (a_dif_media_notas <= -0.375) E (i_dif_media_diaria_acessos <= -0.005) - 75.0%/28
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 3
SE (a_pct_ch_integralizada_us <= 5.905) - SS [i_dif_media_diaria_acessos] - 95.58%/860
SE (a_pct_ch_integralizada_us > 5.905) E (a_pct_reprovacoes > 10.265) - SS [i_dif_media_diaria_acessos] - 62.26%/310
Experimento 19-21_22 – MPCDT_SP – Subclassificador 4
SE (a_dvm_pct_ch_integralizada_us <= -1.872) E (a_media_notas_us <= 5.665) - SS [i_dif_media_diaria_acessos] - 96.3%/568
SE (a_dvm_pct_ch_integralizada_us > -1.872) E (a_pct_reprovacoes > 22.83) - SS - 82.19%/73
SE (a_dvm_pct_ch_integralizada_us <= -1.872) E (a_media_notas_us > 5.665) E (a_pct_reprovacoes > 8.71) - SS [i_dif_media_diaria_acessos] - 81.54%/65
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 4
SE (a_pct_ch_integralizada_us <= 6.965) - SS [i_dif_media_diaria_acessos, i_dif_media_diaria_acoes] - 93.81%/630
SE (a_pct_ch_integralizada_us > 6.965) E (a_pct_reprovacoes > 22.83) - 81.61%/87
SE (a_pct_ch_integralizada_us > 6.965) E (3.075 < a_pct_reprovacoes <= 22.83) E (i_dif_media_diaria_acoes <= -0.025) - 65.96%/47
Experimento 19-21_22 – MPCDT_SP – Subclassificador 5
SE (a_media_notas_us <= 6.175) - SS - 92.66%/531
SE (a_media_notas_us > 6.175) E (a_pct_reprovacoes > 3.075) E (a_stddev_media_notas > 2.385) - 77.03%/74
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 5
SE (a_media_notas_us <= 6.175) - SS [s_dvm_idade_ingresso, i_dif_media_diaria_acessos] - 92.66%/531
SE (6.175 < a_media_notas_us <= 7.175) E (3.075 < a_pct_reprovacoes <= 22.795) E (s_pos_idhm_educacao <= 147.5) - 62.75%/51
SE (a_media_notas_us > 6.175) E (a_pct_reprovacoes > 22.795) - 82.98%/47
Experimento 19-21_22 – MPCDT_SP – Subclassificador 6
SE (a_pct_reprovacoes > 15.705) - SS - 93.91%/345
SE (a_pct_reprovacoes <= 15.705) E (a_pct_ch_integralizada_us <= 9.81) E (a_pct_ch_integralizada <= 74.555) - SS - 69.81%/106
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 6
SE (a_pct_reprovacoes > 16.445) - SS - 94.64%/336
SE (7.595 < a_pct_reprovacoes <= 16.445) E (a_pct_ch_integralizada_us <= 11.63) - 68.13%/91
SE (a_pct_reprovacoes <= 7.595) E (a_pct_ch_integralizada_us <= 0.97) - 84.21%/19
Experimento 19-21_22 – MPCDT_SP – Subclassificador 7
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes > 13.27) - SS - 96.09%/256
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes <= 13.27) E (a_dvm_pct_ch_integralizada_us <= -5.453) - SS - 77.27%/44
SE (a_media_notas_us > 6.99) E (a_pct_reprovacoes > 2.105) E (a_dvm_pct_ch_integralizada_us <= 0.402) E (a_ira <= 7.194) - 66.67%/30
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 7
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes > 13.27) - SS [i_media_diaria_acessos_us] - 96.09%/256
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes <= 13.27) E (a_pct_ch_integralizada_us <= 3.34) - SS - 77.27%/44
SE (a_media_notas_us <= 6.99) E (6.125 < a_pct_reprovacoes <= 13.27) E (a_pct_ch_integralizada_us > 3.34) - 65.22%/23
Experimento 19-21_22 – MPCDT_SP – Subclassificador 8
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_media_notas_us <= 4.98) E (a_pct_ch_integralizada <= 65.8) - 100.0%/269
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_media_notas_us <= 4.98) E (a_pct_ch_integralizada > 65.8) E (a_dvm_pct_aprovacoes > -33.444) - 98.82%/85
SE (a_dvm_pct_ch_integralizada > -4.985) E (a_pct_ch_integralizada <= 92.405) E (a_pct_ch_integralizada_us <= 1.855) E (a_media_notas <= 7.985) - 79.49%/78
Experimento 19-21_22 – MPCDT_FSASP – Subclassificador 8
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_dvm_pct_ch_integralizada_us <= -1.924) - SS - 96.58%/409
SE (a_dvm_pct_ch_integralizada > -4.985) E (a_pct_ch_integralizada_us <= 1.855) E (a_pct_ch_integralizada <= 92.005) - SS [i_pct_questionarios_respondidos] - 77.5%/80
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_dvm_pct_ch_integralizada_us > -1.924) E (a_media_notas <= 6.235) - SS [i_media_diaria_acessos] - 73.33%/30

Fonte: Elaborada pela autora.

Tabela 24 – Principais regras de evasão de cada subclassificador DT dos modelos MPC-SDP, desenvolvidos no experimento 17-19_20, com as seleções RFE e FSA+RFE

Experimento 17-19_20 – MPCDT_RFE – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) E (c_unidade_Frederico Westphalen <= 0.5) - SS - 93.87%/310
SE (s_reserva_vaga_ACG <= 0.5) E (s_reserva_vaga_EP <= 1.5 > 0.5) - 74.7%/83
SE (s_reserva_vaga_ACG <= 0.5) E (s_reserva_vaga_EP > 1.5 > 0.5) E (c_unidade_Santa Rosa <= 0.5) - 82.43%/74
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) - SS [c_tipo_curso_Tecnologia, c_area_Ciências Exatas e da Terra] - 87.77%/368
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 1
SE (a_media_notas_us <= 6.525) E (a_media_freq_us <= 95.17) - SS [i_media_diaria_acessos, i_media_diaria_acoes] - 97.0%/433
SE (a_media_notas_us > 6.525)E (0.085 < i_media_diaria_acoes <= 1.55) E (a_dvm_media_notas <= 2.5) E (i_media_diaria_acessos_us <= 0.555) - 88.24%/68
SE (a_media_notas_us > 6.525) E (i_media_diaria_acoes <= 0.085) E (a_dvm_media_notas <= 2.48) E (i_dvm_media_diaria_acessos <= -0.082) E (a_frequencia_media <= 98.88) - 62.9%/62
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 1
SE (a_dvm_media_notas <= 1.18) E (a_media_freq_us <= 95.17) - SS [i_media_diaria_acessos, i_media_diaria_acoes_us, s_pos_idhm_renda] - 96.99%/432
SE (1.18 < a_dvm_media_notas <= 2.5) E (i_media_diaria_acoes > 0.085) E (i_media_diaria_acoes_us <= 1.55) E (i_media_diaria_acessos_us <= 0.555) - 88.06%/67
SE (1.18 < a_dvm_media_notas <= 2.48) E (i_media_diaria_acoes <= 0.085) E (i_media_diaria_acessos > 0.005) E (a_media_freq_us <= 97.11) - 66.67%/45
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 2
SE (a_media_notas_us <= 6.92) E (a_dvm_pct_ch_integralizada_us <= 0.407) - SS - 97.41%/502
SE (a_media_notas_us <= 6.92) E (a_dvm_pct_ch_integralizada_us > 0.407) E (a_pct_exames > 6.905) - 77.78%/90
SE (a_media_notas_us > 6.92) E (a_pct_ch_integralizada <= 19.375) E (i_media_diaria_acessos_us > 0.045) - 69.64%/56
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 2
SE (a_media_notas_us <= 6.92) E (a_pct_ch_integralizada_us <= 8.265) - SS - 97.41%/502
SE (a_media_notas_us <= 6.92) E (a_pct_ch_integralizada_us > 8.265) E (i_media_diaria_acessos_us > 0.005) E (a_pct_ch_integralizada <= 30.425) - 83.53%/85
SE (a_media_notas_us > 6.92) E (a_pct_ch_integralizada > 19.375) E (i_media_diaria_acoes > 0.215) E (a_pct_ch_integralizada_us <= 11.525) - 70.0%/40
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 3
SE (a_media_notas_us <= 6.235) - SS [i_media_diaria_acoes] - 96.79%/561
SE (a_media_notas_us > 6.235) E (a_pct_reprovacoes > 3.135) E (a_media_freq_us <= 92.51) - 71.76%/85
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 3
SE (a_media_notas_us <= 6.235) E (a_pct_ch_integralizada_us <= 11.2) - SS [i_media_diaria_acessos, s_idhm_educacao] - 98.17%/545
SE (6.235 < a_media_notas_us <= 7.21) E (a_pct_reprovacoes > 3.135) E (a_media_freq_us <= 92.51) - SS - 85.11%/47
SE (a_media_notas_us > 6.235) E (a_pct_reprovacoes > 3.135) E (a_media_freq_us > 92.51) E (a_pct_ch_integralizada_us > 9.22) E (i_media_diaria_acoes_us > 0.54) - 61.11%/36
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 4
SE (a_pct_reprovacoes > 15.5) - SS - 98.44%/448
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dif_media_notas > -1.445) E (a_media_freq_us <= 94.135) - 63.53%/85
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_dif_media_notas <= -1.445) - 93.75%/32
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 4
SE (a_pct_reprovacoes > 15.5) - SS [s_idhm_renda, s_dvm_anos_em_ingresso] - 98.44%/448
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_media_notas_us > 4.945) E (a_media_freq_us <= 94.135) - 64.89%/94
SE (2.39 < a_pct_reprovacoes <= 15.5) E (a_media_notas_us <= 4.945) - 100.0%/23
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 5
SE (a_dvm_pct_ch_integralizada_us <= -1.316) E (a_media_notas_us <= 5.77) - SS - 99.04%/312
SE (-1.316 < a_dvm_pct_ch_integralizada_us <= 2.319) E (2.53 < a_pct_reprovacoes <= 12.31) E (s_idhm_educacao > 0.595) - 68.42%/38
SE (a_dvm_pct_ch_integralizada_us > -1.316) E (a_pct_reprovacoes > 12.31) - 76.92%/26
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 5
SE (a_dvm_pct_ch_integralizada_us <= -1.316) E (a_media_notas_us <= 5.77) - SS - 99.04%/312
SE (a_dvm_pct_ch_integralizada_us > -1.316) E (a_pct_reprovacoes > 12.31) - 76.92%/26
SE (a_dvm_pct_ch_integralizada_us <= -1.316) E (a_media_notas_us > 5.77) E (a_pct_reprovacoes > 11.56) - 91.67%/24
Experimento 17-19_20 – MPCDT_RFE – Subclassificador 6
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 97.92) E (a_pct_reprovacoes > 5.505) - SS - 97.66%/470
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 97.92) E (a_pct_reprovacoes <= 5.505) E (a_dvm_pct_ch_integralizada_us <= -6.28) - 95.65%/23
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 71.53) E (a_pct_reprovacoes <= 5.505) E (a_dvm_pct_ch_integralizada_us > -6.28) - 63.16%/19
Experimento 17-19_20 – MPCDT_FSARFE – Subclassificador 6
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 82.725) E (a_dvm_pct_ch_integralizada_us <= -4.109) - SS [i_media_diaria_acessos_us] - 99.5%/403
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 97.92) E (a_dvm_pct_ch_integralizada_us > -4.109) E (a_iech <= 0.813) - 100.0%/40
SE (a_media_notas_us <= 7.315) E (a_pct_ch_integralizada <= 97.92) E (a_dvm_pct_ch_integralizada_us > -4.109) E (a_iech > 0.813) E (a_media_freq_us <= 89.78) - 75.86%/29

Fonte: Elaborada pela autora.

Tabela 25 – Principais regras de evasão de cada subclassificador DT dos modelos MPC-SDP, desenvolvidos no experimento 19-21_22, com as seleções RFE e FSA+RFE

Experimento 19-21_22 – MPCDT_RFE – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) E (c_area_Ciências Sociais Aplicadas <= 0.5) E (c_unidade_Santa Rosa <= 0.5) E (s_idade_ingresso <= 28.5) - SS - 81.01%/595
SE (s_reserva_vaga_ACG > 0.5) E (c_area_Ciências Sociais Aplicadas <= 0.5) E (c_unidade_Santa Rosa <= 0.5) E (s_idade_ingresso > 28.5) E (s_idhm_educacao <= 0.732) - 94.03%/201
SE (s_reserva_vaga_EP <= 1.5 > 0.5) E (c_area_Ciências Sociais Aplicadas <= 0.5) E (s_dvm_idade_ingresso <= 0.5) E (s_dvm_anos_em_ingresso <= 1.5) - 69.29%/127
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 0
SE (s_reserva_vaga_ACG > 0.5) E (c_unidade_São Vicente do Sul <= 0.5) E (c_unidade_Frederico Westphalen <= 0.5) - SS - 81.53%/926
SE (s_reserva_vaga_ACG > 0.5) E (c_unidade_Frederico Westphalen > 0.5) E (c_tipo_curso_Tecnologia <= 0.5) - 73.53%/68
SE (s_reserva_vaga_ACG > 0.5) E (c_unidade_São Vicente do Sul > 0.5) E (c_area_Ciências Exatas e da Terra > 0.5) - SS - 80.0%/45
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 1
SE (a_media_notas_us <= 6.525) E (a_pct_ch_integralizada_us <= 1.98) E (s_reserva_vaga_ACG > 0.5) - 100.0%/411
SE (a_media_notas_us <= 6.525) E (a_pct_ch_integralizada_us > 1.98) E (i_dvm_media_diaria_acoes > -0.847) E (a_pct_reprovacoes > 5.28) - SS [s_reserva_vaga_ACG] - 94.59%/259
SE (a_media_notas_us > 6.525) E (s_reserva_vaga_ACG > 0.5) E (i_pct_questionarios_respondidos > -0.5) E (a_dvm_pct_ch_integralizada <= 10.553) E (i_media_diaria_acoes_us > 0.365) - 99.01%/101
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 1
SE (a_media_notas_us <= 5.72) E (a_dvm_pct_ch_integralizada_us <= -3.757) E (i_dvm_media_diaria_acessos > -0.932) - SS - 98.95%/573
SE (a_media_notas_us <= 6.69) E (-3.757 < a_dvm_pct_ch_integralizada_us <= 13.143) E (i_dvm_media_diaria_acessos > -0.699) E (i_media_diaria_acoes_us > 0.015) - 90.67%/225
SE (a_media_notas_us > 6.69) E (2.275 < i_pct_questionarios_respondidos <= 97.435) E (i_media_diaria_acoes > 0.505) - 100.0%/95
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 2
SE (a_media_notas_us <= 5.905) E (a_dvm_pct_ch_integralizada_us <= -2.931) E (a_iean <= 1979.262) E (i_media_diaria_acoes > 0.005) E (a_media_notas <= 9.01) E (i_dif_media_diaria_acoes <= 1.26) - SS - 99.41%/675
SE (a_media_notas_us <= 5.905) E (a_dvm_pct_ch_integralizada_us > -2.931) E (a_pct_exames > 3.175) E (a_dvm_frequencia_media <= 8.069) E (a_dvm_pct_ch_integralizada <= 26.292) E (i_media_diaria_acessos > 0.01) - SS - 92.78%/97
SE (a_media_notas_us <= 5.905) E (a_dvm_pct_ch_integralizada_us <= -2.931) E (a_iean <= 1979.262) E (i_media_diaria_acoes <= 0.005) E (a_media_notas <= 5.8) E (a_dvm_pct_ch_integralizada <= -3.583) E (a_pct_exames <= 56.35) - 92.86%/56
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 2
SE (a_media_notas_us <= 5.905) E (a_pct_ch_integralizada_us <= 4.925) - SS [s_reserva_vaga_ACG, i_dif_media_diaria_acoes, i_dif_media_diaria_acessos] - 97.33%/750
SE (a_media_notas_us <= 5.905) E (4.925 < a_pct_ch_integralizada_us <= 10.405) E (i_media_diaria_acoes_us > 0.025) - 86.21%/116
SE (a_media_notas_us > 5.905) E (s_reserva_vaga_ACG > 0.5) E (a_pct_ch_integralizada_us <= 9.955) - SS [i_dif_media_diaria_acessos] - 73.33%/90
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 3
SE (a_dvm_pct_ch_integralizada_us <= -2.364) - SS [i_media_diaria_acoes, i_media_diaria_acessos] - 95.58%/860
SE (a_dvm_pct_ch_integralizada_us > -2.364) E (a_pct_reprovacoes > 3.51) E (a_stddev_media_notas > 1.905) E (i_media_diaria_acoes > 0.175) - SS - 72.15%/219
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 3
SE (a_media_notas_us <= 5.815) - SS [i_dif_media_diaria_acoes, i_media_diaria_acoes_us] - 92.51%/948
SE (a_media_notas_us > 5.815) E (a_pct_reprovacoes > 2.78) E (s_reserva_vaga_ACG > 0.5) - SS [i_dif_media_diaria_acessos] - 66.04%/159
SE (a_media_notas_us > 5.815) E (s_reserva_vaga_ACG <= 0.5) E (a_pct_reprovacoes > 10.82) E (i_media_diaria_acoes_us > 0.255) - 60.44%/91
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 4
SE (a_dvm_pct_ch_integralizada_us <= -2.282) - SS [i_dif_media_diaria_acessos, i_media_diaria_acoes] - 93.81%/630
SE (a_dvm_pct_ch_integralizada_us > -2.282) E (a_pct_reprovacoes > 22.83) - 81.61%/87
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 4
SE (a_dvm_pct_ch_integralizada_us <= -2.282) - SS [i_dif_media_diaria_acessos, s_dvm_anos_em_ingresso, e_renda_per_capta] - 93.81%/630
SE (a_dvm_pct_ch_integralizada_us > -2.282) E (a_pct_reprovacoes > 22.83) - 81.61%/87
SE (-2.282 < a_dvm_pct_ch_integralizada_us <= 1.953) E (3.075 < a_pct_reprovacoes <= 22.83) E (i_dif_media_diaria_acoes > -0.025) E (a_dif_freq <= -1.9) - 62.3%/61
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 5
SE (a_media_notas_us <= 6.175) - SS - 92.66%/531
SE (a_media_notas_us > 6.175) E (a_pct_reprovacoes > 3.075) E (a_iech <= 0.882) - 60.67%/150
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 5
SE (a_media_notas_us <= 6.175) - SS - 92.66%/531
SE (a_media_notas_us > 6.175) E (a_pct_reprovacoes > 18.14) - 75.0%/68
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 6
SE (a_pct_reprovacoes > 15.705) - SS [i_media_diaria_acessos_us] - 93.91%/345
SE (a_pct_reprovacoes <= 15.705) E (a_pct_ch_integralizada_us <= 9.81) E (a_pct_mat_interrompidas > 2.86) - 86.05%/43
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 6
SE (a_pct_reprovacoes > 15.705) E (a_media_notas_us <= 5.245) - SS [i_dif_media_diaria_acoes] - 98.46%/259
SE (a_pct_reprovacoes <= 15.705) E (a_dvm_pct_ch_integralizada_us <= -0.4) E (a_dvm_pct_ch_integralizada <= 9.665) - SS [i_dif_media_diaria_acoes] - 69.81%/106
SE (a_pct_reprovacoes > 15.705) E (a_media_notas_us > 5.245) E (a_iean <= 756.214) - SS [s_anos_em_ingresso] - 88.24%/68
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 7
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes > 6.125) - SS - 93.05%/302
SE (a_media_notas_us > 6.99) E (a_pct_reprovacoes > 22.825) - 83.33%/12
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes <= 6.125) E (i_pct_contatos_docentes > 41.665) - 90.91%/11
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 7
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes > 6.125) - SS - 93.05%/302
SE (a_media_notas_us <= 6.99) E (a_pct_reprovacoes <= 6.125) E (a_dvm_pct_ch_integralizada_us <= -4.208) - 65.22%/23
SE (a_media_notas_us > 6.99) E (a_pct_reprovacoes > 22.825) - 83.33%/12
Experimento 19-21_22 – MPCDT_RFE – Subclassificador 8
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_media_notas_us <= 4.98) E (a_pct_ch_integralizada <= 65.8) - 100.0%/269
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_media_notas_us <= 4.98) E (a_pct_ch_integralizada > 65.8) E (a_pct_aprovacoes > 47.655) E (i_dif_media_diaria_acoes <= 0.21) - 100.0%/73
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_media_notas_us > 4.98) E (a_pct_aprovacoes <= 71.2) E (i_dvm_media_diaria_acessos <= 0.032) E (a_dvm_pct_ch_integralizada_us <= 0.591) - 100.0%/35
Experimento 19-21_22 – MPCDT_FSARFE – Subclassificador 8
SE (a_dvm_pct_ch_integralizada <= -4.985) E (a_dvm_pct_ch_integralizada_us <= -1.924) E (a_pct_ch_integralizada_us <= 3.135) E (a_pct_ch_integralizada <= 79.12) - SS [i_dvm_media_diaria_acessos] - 99.7%/332
SE (a_dvm_pct_ch_integralizada > -4.985) E (a_pct_ch_integralizada_us <= 1.855) E (a_pct_ch_integralizada <= 92.005) E (s_idade_ingresso > 17.5) E (a_dif_freq > -98.535) - 84.93%/73
SE (a_dvm_pct_ch_integralizada <= -25.77) E (a_dvm_pct_ch_integralizada_us <= -1.924) E (a_pct_ch_integralizada_us > 3.135) - 100.0%/24

Fonte: Elaborada pela autora.

semestre do modelo MPC-SDP com seleção FSA+SP, no experimento 17-19_20 (Tabela 22), por exemplo, relaciona a evasão a estudantes que, mesmo tendo aumentado seu nível de interação (i_dif_media_diaria_acessos $>$ 0.005) e integralizado uma carga horária acima da média de seus pares (a_dvm_pct_ch_integralizada_us $>$ 0.407), apresentaram uma média de notas baixa (a_media_notas_us $<=$ 6.92)[15], no semestre. Já a terceira principal regra de evasão do subclassificador especializado em segundo semestre do modelo MPC-SDP com seleção FSA+SP, no experimento 19-21_22 (Tabela 23), relaciona à evasão estudantes que, no segundo semestre, apresentaram uma redução na média diária de acessos (i_dif_media_diaria_acessos $<=$ -0.005), e que, tendo uma média de notas superior a 4.66 (a_media_notas_us $>$ 4.66), integralizaram uma baixa porcentagem da carga horária do curso, no semestre[16]. Ou seja, essas regras sinalizam que tanto níveis de interação mais altos quanto níveis de interação mais baixos levam à evasão, quando associados a desempenhos acadêmicos inferiores. Enquanto, no segundo caso, as evasões podem ser resultantes de desengajamento; no primeiro, podem decorrer de desmotivação e desengajamento, após o emprego de um maior esforço (acessos ao AVA) não se mostrar suficiente para suprir as dificuldades de aprendizagem e/ou uma expectativa de desempenho.

De forma geral, estes padrões se alinham a estudos da literatura, como Casanova et al. (2018); Barroso et al. (2022); Casanova et al. (2023), que indicam que os estudantes podem apresentar dificuldades de adaptação ao ensino superior por, dentre outros motivos, não apresentarem *background* educacional, habilidades, e níveis de autonomia que supram as demandas desse nível de ensino. Sem o suporte adequado, esses fatores podem desencadear baixo rendimento acadêmico, desmotivação, desengajamento, e, finalmente, evasão.

A partir do terceiro semestre, é possível verificar que os atributos interacionais perdem poder preditivo entre as principais regras dos subclassificadores especializados[17], que passam a ser dominadas, majoritariamente, apenas por atributos acadêmicos. Dentre as principais regras, além das já citadas relações envolvendo notas e porcentagens de carga horária integralizada, são consideradas, bastantemente recorrente, as porcentagens de reprovação dos estudantes, com valores acima de determinados limiares sendo associados aos padrões de evasão. Essa predominância de

[15]Considerando que, institucionalmente, a aprovação, sem exame, se dá a partir de médias maiores ou iguais a 7.0, médias de notas inferiores a esse limiar indicam que os estudantes tiveram reprovação ou necessidade de exame em alguma(s) disciplina(s).

[16]Observe que essa baixa integralização associada a uma média de notas superior a 4.66 inclui tanto alunos com baixo aproveitamento nas disciplinas (médias de notas inferior a 5, que, no IFFar, é o limite de aprovação em exame), quanto alunos com bom aproveitamento, mas que tenham se matriculado em poucas disciplinas no semestre.

[17]Embora os atributos interacionais continuem aparecendo entre as regras de alguns subclassificadores, eles passam a aparecer com menor frequência entre os padrões mais abrangentes ou a fazer parte das subárvores suprimidas. Nesse caso, eles compõem o caminho da classificação, mas não interferem em seu resultado (abaixo de seus nós, todas as decisões são de evasão).

atributos acadêmicos entre os padrões preditivos vai ao encontro de diversos estudos, como Costa et al. (2021); Berka; Marek (2021), que apontam o maior poder preditivo dos dados de performance acadêmica, no contexto da evasão estudantil.

Também se faz importante mencionar que, dentre as três principais regras de evasão listadas para os subclassificadores, algumas são menos/pouco abrangentes, porém, possuem uma confiabilidade/probabilidade, perante os registros de treinamento, bastante alta. Por exemplo:

- No subclassificador especializado em quarto semestre do MPC-SDP com FSA+RFE do experimento 17-19_20 (Tabela 24), é apresentada uma regra que abrange apenas 23 estudantes, mas com confiabilidade de 100%: Se o estudante apresentar uma porcentagem de reprovações entre 2.39% e 15.5%[18], e uma média de notas menor ou igual a 4.945 no semestre, ele evade;

- No subclassificador especializado em dados de primeiro semestre do modelo MPC-SDP com seleção FSA+RFE do experimento 19-21_22 (Tabela 25), consta um padrão que abrange 95 estudantes, também com confiabilidade/probabilidade de 100%: Se o estudante obter uma média de notas no semestre superior a 6.69, responder, mas não todos, os questionários disponibilizados em suas turmas virtuais, e apresentar uma média diária de ações[19] superior a 0.505, ele evade;

- No subclassificador especializado em dados de sétimo semestre do MPC-SDP com seleção RFE do experimento 19-21_22 (Tabela 25), a terceira regra abrange apenas 11 estudantes, mas com probabilidade de 91%: Se o estudante apresentar uma média de notas no semestre menor ou igual a 6.99, uma porcentagem de reprovações menor ou igual a 6.125, e uma porcentagem de contatos docentes[20] maior que 41.665, ele evade.

Na primeira dessas regras, são incluídos estudantes que possuam um histórico de reprovações, seja ele pequeno ou considerável, e que apresentem uma média de notas muito baixa no semestre, indicando diversos exames e reprovações. Na segunda,

[18]Observe que, por simplificação, a regra foi apresentada integrando duas condições de um mesmo atributo: (2.39 < a_pct_reprovacoes <= 15.5). Porém, como DTs são árvores binárias, cujos nós representam uma única condição, isso significa que dois nós, em níveis distintos de um mesmo ramo (caminho de decisão) da DT, apresentavam, separadamente, cada uma das condições descritas.

[19]Note que, embora os modelos com seleção RFE e FSA+RFE tenham considerado, respectivamente, os atributos de média diária de ações no semestre (i_media_diaria_acoes_us) e média diária geral de ações (i_media_diaria_acoes), no primeiro semestre eles se equivalem. Além disso, diferentemente de acessos, ações se referem a todas as atividades (acessar, listar, alterar, responder, submeter, remover, etc.) realizadas pelos estudantes no AVA, em entidades como turmas, tópicos de aula, notas, frequências, conteúdos, arquivos, enquetes, questionários, fóruns, tarefas, etc.

[20]Porcentagem de docentes entre os usuários/contatos com quem o estudante trocou mensagens, no AVA.

são considerados evadidos estudantes com um regular/bom desempenho acadêmico no semestre, mas que, apesar de terem uma alta média diária de ações, não responderam a todos os questionários disponibilizados nas turmas virtuais. Por sua vez, a terceira regra considera, como um fator diferencial de evasão, altas porcentagens de docentes entre os contatos de estudantes com desempenho acadêmico abaixo da média, o que pode indicar uma maior dificuldade e uma procura mais ativa desses estudantes por auxílios/esclarecimentos dos docentes.

Por fim, é importante observar que atributos interacionais mais específicos (que não sejam indicadores gerais de ações e acessos) só passaram a compor a estrutura das principais regras de evasão nos modelos do experimento 19-21_22, que agregam dados de treinamento do contexto pandêmico. Isso demonstra que atributos interacionais tiveram um ganho de potencial preditivo durante o ERE. Tal observação, já esperada devido à maior utilização do AVA nesse cenário, também foi apontada em Colpo; Primo; Aguiar (2024), mediante a comparação de modelos DT baseados em dados de evasão de antes e durante a pandemia de COVID-19.

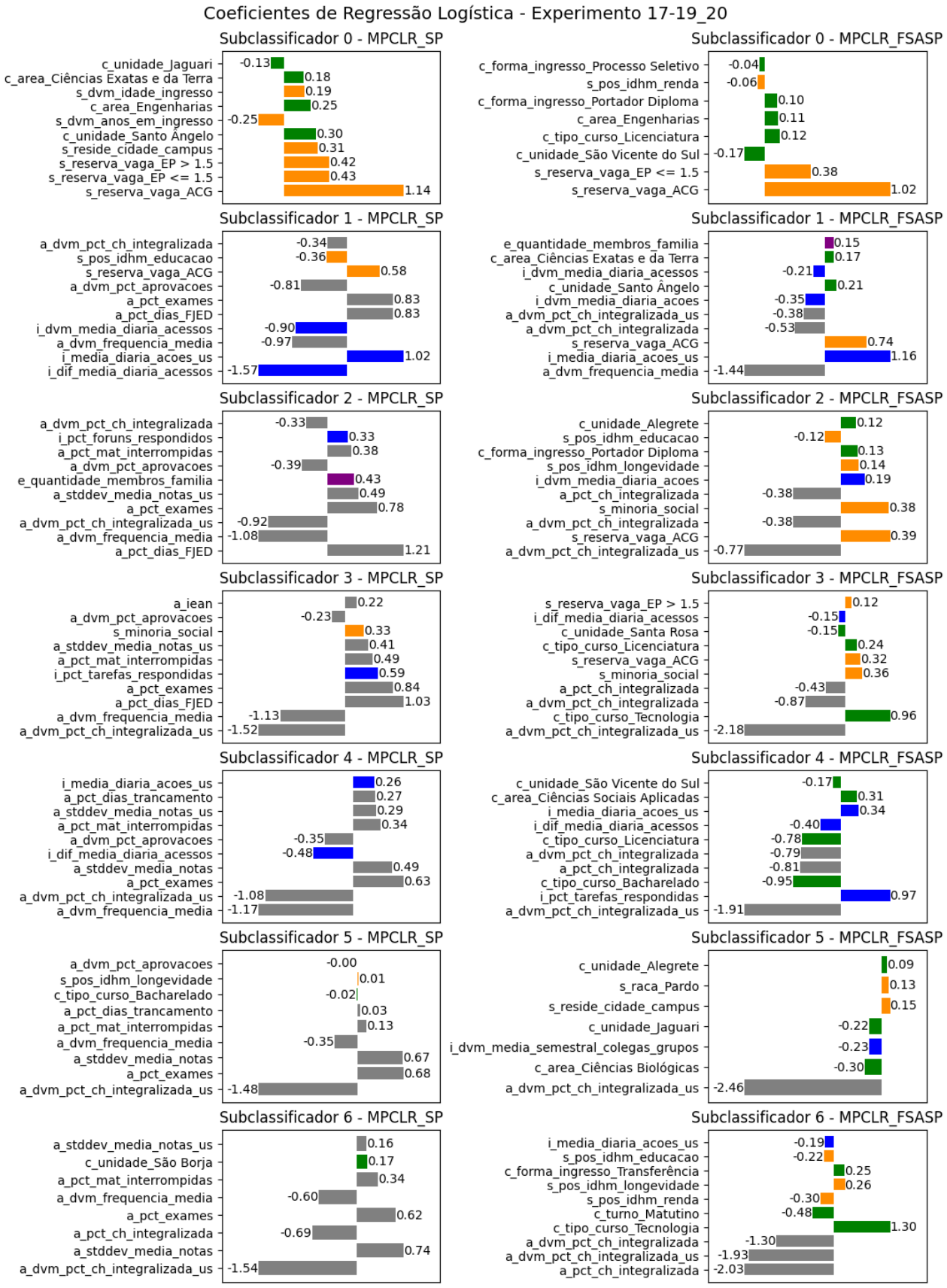
7.2.2 Modelos baseados em LR

Seguindo a mesma lógica de análise da Seção anterior, nas Figuras 32 e 33 são apresentados os dez principais atributos, considerando a magnitude de seus coeficientes, para cada subclassificador LR dos modelos MPC-SDP com seleção SP (primeira coluna) e FSA+SP (segunda coluna), dos experimentos 17-19_20 e 19-21_22, respectivamente. Já nas Figuras 34 e 35 são apresentados os coeficientes dos modelos com seleção RFE e FSA+RFE, também considerando os experimentos 17-19_20 e 19-21_22, respectivamente. Note que os coeficientes de cada subclassificador estão apresentados em gráficos de barra, ordenados por suas magnitudes (crescentemente), e acompanhados por seus respectivos atributos. Para facilitar a identificação dos diferentes aspectos dos atributos, as barras foram categorizadas por cores, seguindo o mesmo padrão de cores utilizado na Seção 7.2.1[21].

Como foram realizadas a padronização/normalização dos dados e a remoção de atributos multicolineares, antes do treinamento dos modelos baseados em LR, nas Figuras 32, 33, 34, e 35, de forma geral, a magnitude dos coeficientes pode ser interpretada como a força da relação entre o atributo e a classe/evasão, enquanto o sinal indica o tipo da relação (positivo = direta; negativo = inversa). Porém, essa interpretação deve ser feita com cuidado, já que não há garantia de que todos os atributos multicolineares tenham sido identificados no processo de remoção. Ou seja, alguns coeficientes de atributos correlatos podem aparecer com significado contraditório ou de forma superestimada, parecendo importantes quando, na realidade, possuem in-

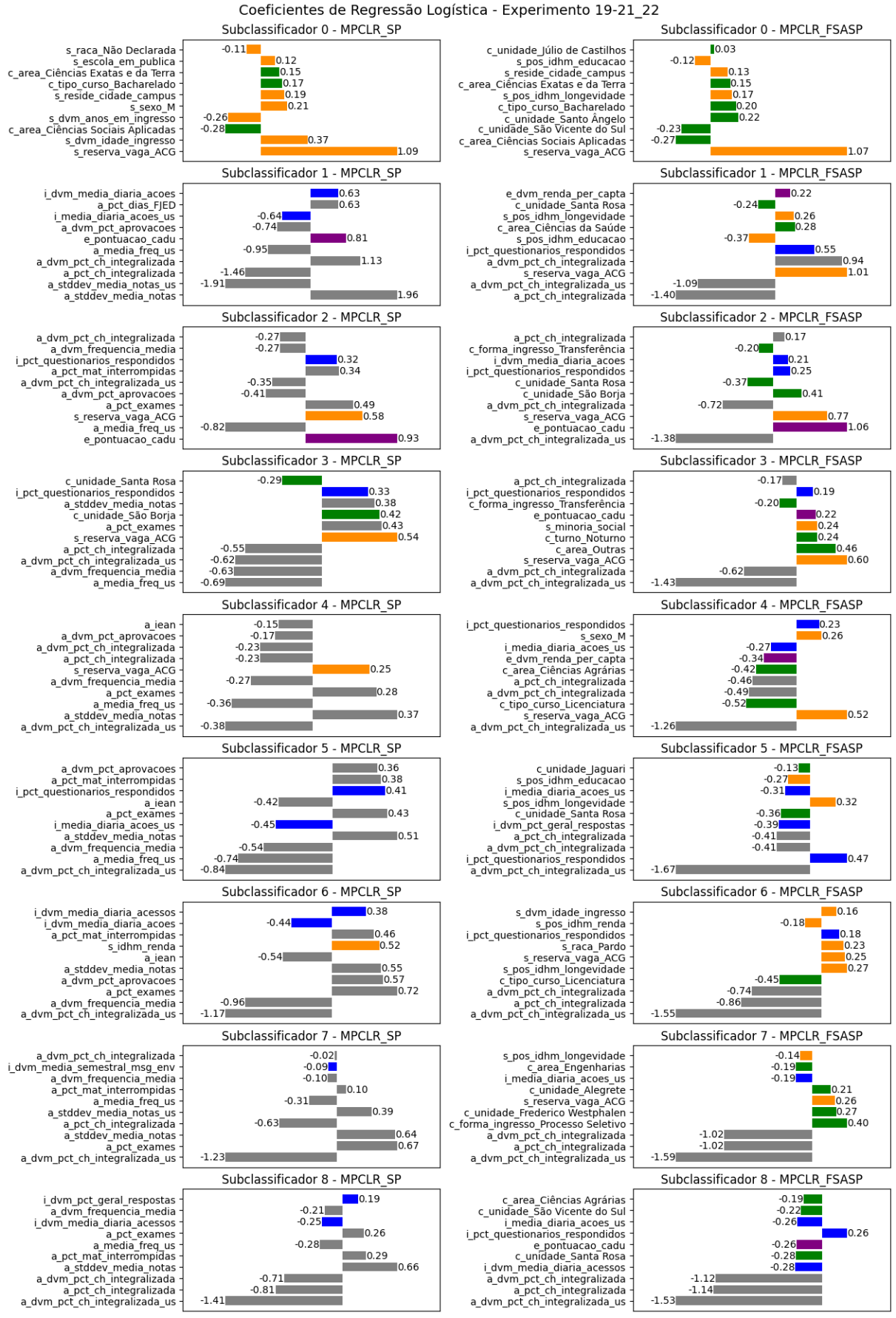
[21] Padrão de cores/aspectos: verde - contextual; laranja - social/demográfico; roxo - econômico; azul - interacional; cinza - acadêmico.

Figura 32 – Principais coeficientes de cada subclassificador LR dos modelos MPC-SDP, desenvolvidos no experimento 17-19_20, com as seleções SP e FSA+SP



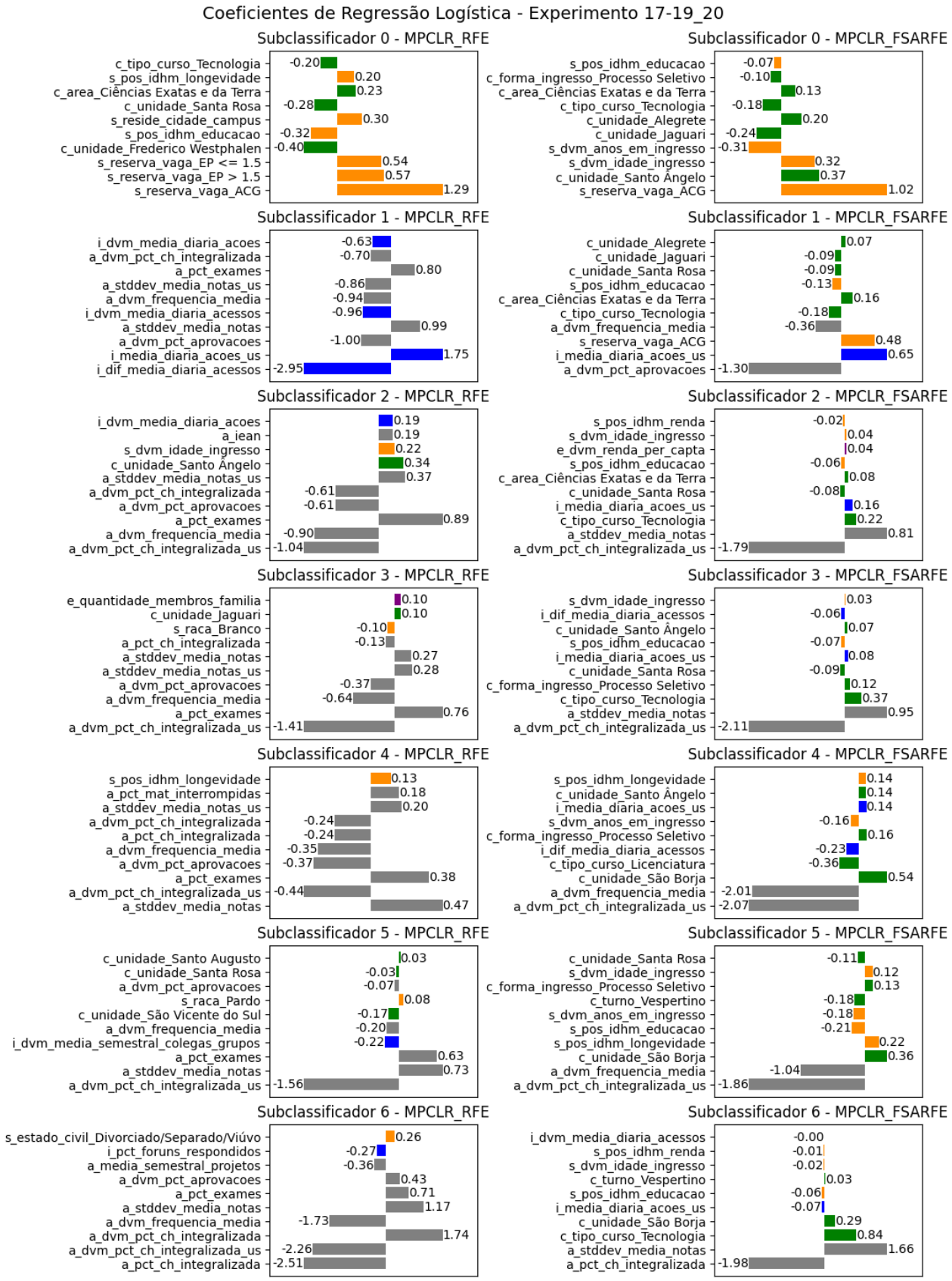
Fonte: Elaborada pela autora.

Figura 33 – Principais coeficientes de cada subclassificador LR dos modelos MPC-SDP, desenvolvidos no experimento 19-21_22, com as seleções SP e FSA+SP



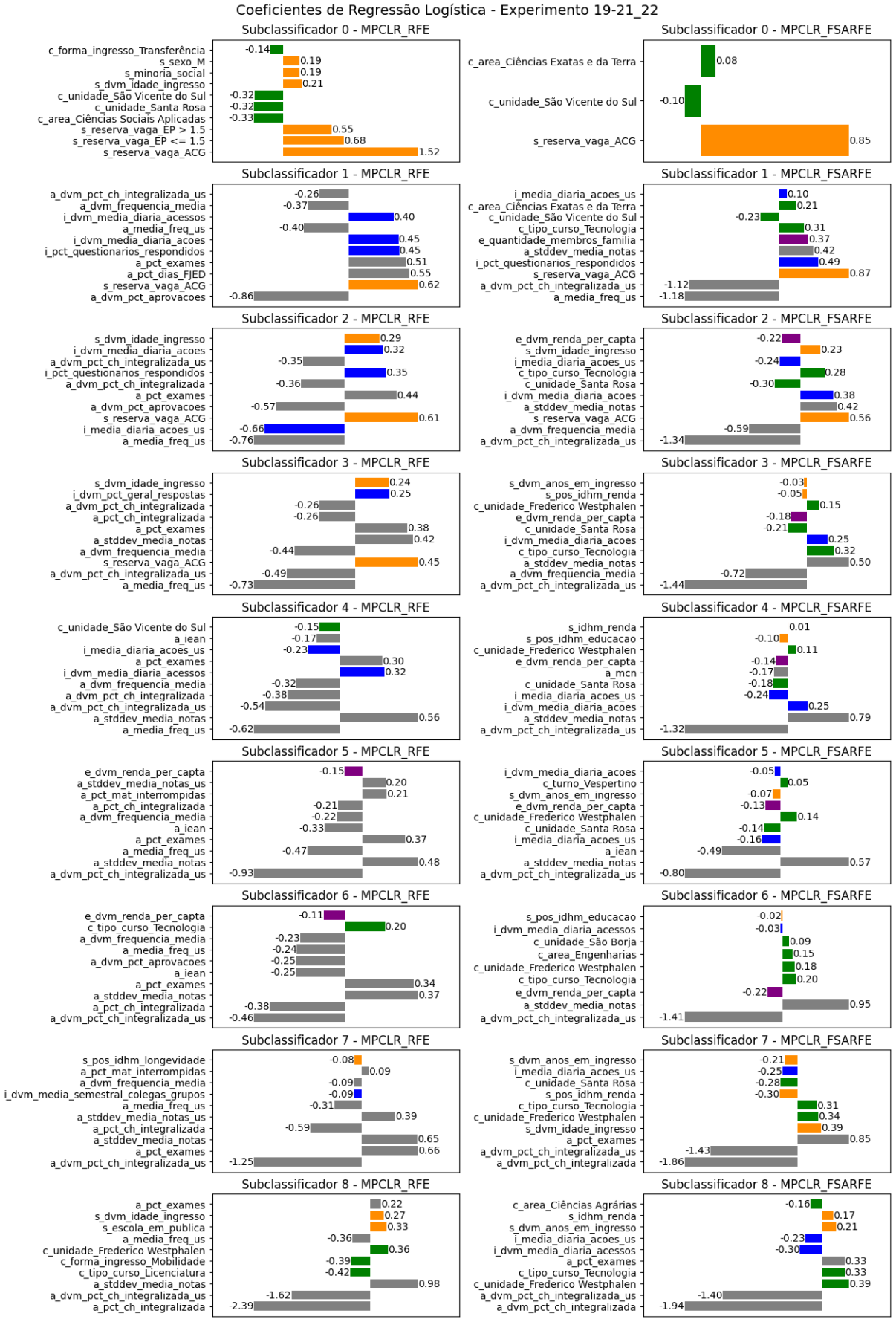
Fonte: Elaborada pela autora.

Figura 34 – Principais coeficientes de cada subclassificador LR dos modelos MPC-SDP, desenvolvidos no experimento 17-19_20, com as seleções RFE e FSA+RFE



Fonte: Elaborada pela autora.

Figura 35 – Principais coeficientes de cada subclassificador LR dos modelos MPC-SDP, desenvolvidos no experimento 19-21_22, com as seleções RFE e FSA+RFE



Fonte: Elaborada pela autora.

fluência pouco significativa.

De forma geral, percebe-se que o uso da seleção FSA também propiciou aos subclassificadores LR maior variabilidade de aspectos entre seus principais atributos/coeficientes, a partir do primeiro semestre. Porém, como esperado, assim como na Seção 7.2.1, os principais coeficientes dos subclassificadores especializados no momento de admissão (semestre 0) se concentraram em atributos contextuais e sociais/demográficos, já que os atributos dos demais aspectos não possuem informação nesse estágio acadêmico. Pode-se observar, entre os subclassificadores de semestre 0, que os coeficientes atrelados aos atributos sociais de reservas de vaga de ACG e de escola pública frequentemente apresentaram relação direta com a evasão, sendo as maiores magnitudes associadas às reservas de ACG. Os coeficientes dos principais atributos sociais sugerem que um contexto de origem mais favorável (ingresso por ampla concorrência, residência local, e municípios de naturalidade com melhores/menores posições no IDHM de educação e renda) contribui para a evasão, assim como um contexto social desfavorável (origem de escola pública, minoria social, e municípios de naturalidade com piores/maiores posições no IDHM de longevidade). Enquanto, no primeiro caso, os estudantes podem ter um maior suporte familiar e evadir em busca de oportunidades educacionais mais alinhadas ao seu desejo (dedicar-se a seleções de outras instituições e/ou cursos, por exemplo); no segundo, é possível que as evasões ocorram por dificuldades de adaptação/aprendizagem ou necessidade de ingresso ou conciliação com o mercado de trabalho.

Além disso, chama atenção o fato de o desvio da média de anos entre a conclusão do ensino médio e o ingresso no ensino superior geralmente apresentar relação inversa com a evasão, nos subclassificadores especializados no momento de admissão (semestre 0), ao contrário da idade ou do desvio da idade de ingresso. Essa dualidade poderia se justificar no fato de que, apesar de enfrentarem dificuldades relacionadas à perda de hábitos de aprendizagem, estudantes mais velhos também são mais experientes e tendem a ser mais comprometidos/motivados (Berka; Marek, 2021). Porém, é preciso considerar que, em muitos casos, esses atributos podem ser multicolineares: estudantes que concluíram o ensino médio há mais tempo naturalmente ingressam no ensino superior com mais idade, embora estudantes que concluíram o ensino médio com maior idade também ingressem no ensino superior mais velhos, sem, necessariamente, um grande intervalo temporal entre essas etapas. Logo, nem a importância/magnitude e nem o significado/sinal desses coeficientes podem ser interpretados como relevantes.

Em relação aos atributos contextuais, os coeficientes dos subclassificadores LR baseados em dados de admissão, semelhantemente aos padrões dos subclassificadores DT (Seção 7.2.1), associaram à evasão a área de Ciências Exatas e da Terra e o *campus* Santo Ângelo, enquanto a área de Ciências Sociais Aplicadas e as unidades

São Vicente do Sul, Santa Rosa, e Jaguari apresentaram recorrente relação inversa. Embora muitas dessas relações entre a evasão e os atributos sociais e contextuais continuem sendo referenciadas pelos subclassificadores especializados nos demais semestres, faz-se importante pontuar que os subclassificadores treinados a partir de dados do segundo semestre também relacionam frequentemente à evasão os *campi* São Borja e Frederico Westphalen, o que sugere que estudantes do *campus* Santo Ângelo possuem uma maior tendência à evasão precoce (no primeiro semestre).

Nos subclassificadores especializados em dados de primeiro e segundo semestre, ganharam importância os coeficientes relacionados a atributos econômicos, interacionais e, principalmente, acadêmicos. Neles, a evasão foi diretamente associada aos atributos de porcentagem de exames, desvio padrão de notas, e porcentagem de dias de faltas justificadas e estudos domiciliares (a_pct_dias_FJED); e inversamente relacionada aos atributos de médias (ou desvios de médias) de notas e frequências, e de porcentagens de aprovação e carga horária integralizada[22]; gerais ou do semestre. Tais relações, novamente, confirmam que quanto melhor o desempenho/aproveitamento acadêmico e quanto mais avançado o estudante estiver no curso, menor sua tendência à evasão. Quanto ao aspecto interacional, os subclassificadores frequentemente apresentaram relações divergentes, perante à evasão, para os atributos relacionados às médias diárias de ações (e de porcentagens de respostas a questionários ou tarefas), e para os atrelados às médias diárias de acesso. Além da possível multicolinearidade entre os atributos interacionais (o acesso mais frequente ao AVA pode refletir diretamente no quantitativo de ações realizadas nas turmas virtuais), é válido destacar que essas divergências também podem ter relação com a complexidade dos padrões e com as observações feitas na Seção 7.2.1: os atributos interacionais podem ter impacto distinto na evasão, dependendo das condições acadêmicas associadas. Por fim, os atributos econômicos de quantidade de membros do grupo familiar e de pontuação no Cadastro Único também apresentaram relação direta com a evasão, ao contrário do desvio da média de renda per capita, o que sinaliza que estudantes em situação socioeconômica mais vulnerável possuem maior propensão à evasão.

A partir do terceiro semestre, embora ainda considerem atributos de diferentes aspectos, os subclassificadores demonstram dar ainda mais importância aos atributos acadêmicos, o que, novamente, vai ao encontro de estudos como Costa et al. (2021); Berka; Marek (2021), que apontam o maior poder preditivo dos dados de performance acadêmica, no contexto da evasão estudantil. Porém, apesar de menos relevantes, é

[22]Embora seja evidente a permanência de atributos multicolineares relacionados às cargas horárias integralizadas (especialmente no primeiro semestre, quando os totalizadores gerais e do semestre possuem a mesma informação), pode-se observar, a exemplo do subclassificador de primeiro semestre do MPC-SDP com seleção FSA+SP do experimento 17-19_20 (segunda coluna da Figura 32), que geralmente os coeficientes desses atributos apresentam o mesmo sinal e, consequentemente, ao invés de superestimados, dividem importância no modelo logístico.

perceptível que atributos econômicos e interacionais apresentaram maior destaque e recorrência entre os subclassificadores dos modelos associados ao experimento 19-21_22, que agrega o contexto da pandemia de COVID-19 nos dados de treinamento. Além de sugerir a existência de padrões mais complexos no cenário pandêmico, isso confirma que o aumento da utilização do AVA, durante o ERE, proporcionou maior riqueza aos dados do aspecto interacional, incrementando seu potencial preditivo, conforme previamente observado em Colpo; Primo; Aguiar (2024).

8 CONCLUSÃO

Esta Tese teve como objetivo geral a proposição e a construção, a partir de técnicas de EDM e aprendizado de máquina, de um modelo de predição genérico de evasão, mas guiado por perspectivas especializadas ao domínio educacional. Para isso, como primeiro objetivo de pesquisa, foi desenvolvida uma RSL acerca da aplicação de EDM no contexto da predição de evasão estudantil, considerando os escopos institucional e de curso. Por meio dessa RSL, constatou-se um maior foco de pesquisa no nível de graduação, na rede pública de ensino, e na modalidade presencial. Sob a perspectiva dos dados utilizados, além de geralmente considerarem combinações de um a três aspectos, entre acadêmicos, sociais, e econômicos, na representação dos estudantes, os artigos analisados frequentemente apontaram atributos de desempenho acadêmico dentre os de maior potencial preditivo. Consequentemente, trabalhos que construíram modelos relacionados a diferentes estágios da trajetória acadêmica dos alunos observaram que, quanto mais avançado o momento de predição e maior o volume de dados acadêmicos disponível, melhores os resultados preditivos.

Em relação às técnicas empregadas, a maioria dos trabalhos abordou o problema da evasão por meio da tarefa de classificação, com destaque ao uso de modelos de DT, considerados na forma individual ou combinada (*ensembles*). Também mostrou-se comum o desenvolvimento de modelos especializados em cursos e/ou momentos de predição. Embora tais modelos possam ter sua capacidade preditiva beneficiada pela exploração de fatores mais específicos, como o desempenho de estudantes em determinadas disciplinas, essa abordagem provê resultados menos abrangentes, institucionalmente. Adicionalmente, junto à identificação de uma baixa adoção de técnicas de seleção de atributos e de balanceamento, além de falhas de validação em algumas pesquisas, observou-se pouca adaptação das técnicas utilizadas para o domínio de predição de evasão. Embora considerem questões/objetivos de pesquisa direcionados ao problema da evasão estudantil, em geral, os trabalhos adotam técnicas de mineração de dados e aprendizado de máquina em suas formas tradicionais, sem especializações/adaptações voltadas às particularidades desse domínio.

A partir da revisão da literatura, deu-se início às atividades de identificação, obten-

ção, e pré-processamento dos dados de interesse, segundo objetivo de pesquisa e pré-requisito de qualquer trabalho baseado em EDM e aprendizado de máquina. Considerando o contexto de cursos de graduação presenciais do IFFar como estudo de caso, foram extraídos dados semestrais de diferentes aspectos dos estudantes (acadêmicos, contextuais, econômicos, interacionais, e sociais/demográficos), disponíveis na base de dados do sistema de gestão acadêmica da instituição. Também foram considerados dados externos, mais especificamente do Atlas do Desenvolvimento Humano no Brasil, para agregar informações relacionadas aos IDHMs das origens dos estudantes. No pré-processamento, além de atividades de integração, limpeza, e transformação dos dados, foi executada a derivação de atributos. Em especial, atributos absolutos foram utilizados na derivação de atributos relativos, oportunizando a equiparação e a generalização de dados de estudantes que, em suas formas brutas/absolutas, não seriam diretamente comparáveis entre si, devido à variabilidade relacionada aos diferentes cursos e períodos envolvidos na predição genérica.

Essas particularidades relacionadas à preparação dos dados constituíram o ponto de partida para o atendimento do terceiro objetivo de pesquisa: o estabelecimento de uma estrutura de aplicação de EDM focada na predição genérica de evasão, e apoiada em conhecimentos já revelados na literatura. A estrutura proposta compreende etapas relacionadas ao desenvolvimento de um modelo de *ensemble*, que, embora voltado à predição genérica, considera perspectivas especializadas e importantes para o domínio da evasão estudantil, buscando: (i) dar maior atenção à variabilidade dos padrões de evasão no decorrer da trajetória acadêmica, a partir da construção de um comitê de decisão formado por subclassificadores especializados por períodos/semestres letivos; e (ii) oportunizar uma participação mais abrangente de diferentes aspectos dos estudantes (acadêmicos, contextuais, econômicos, interacionais, e sociais/demográficos) na representação do comportamento de evasão, por meio de um processo de seleção de atributos orientado/separado por aspecto, ao qual se atribuiu o nome de FSA, acrônimo para *Feature Selection by Aspect*. Por consistir em um comitê de classificação genérico, mas guiado por perspectivas especializadas e importantes para o domínio de predição da evasão estudantil, o modelo proposto recebeu o nome de MPC-SDP, acrônimo, em Inglês, para *Multi-Perspective Classifier for Student Dropout Prediction*.

Conforme sua proposta, o MPC-SDP foi implementado para conduzir os processos de balanceamento de dados, seleção de atributos, e otimização de hiperparâmetros sobre os subconjuntos de dados de cada subclassificador especializado por semestre. Atendendo ao quarto objetivo de pesquisa, após implementado, o MPC-SDP foi treinado e avaliado sob dados de estudantes concluídos e evadidos em três diferentes recortes temporais/experimentos, e sob a variação do algoritmo de classificação base (DT e LR) e do método de seleção de atributos (nenhum, SP, RFE,

FSA+SP, FSA+RFE), adotados na construção de cada subclassificador. Além da análise comparativa dos desempenhos preditivos, considerando como *baselines* modelos de predição genérica construídos tradicionalmente, foram analisados os atributos e padrões de maior influência/importância preditiva para os subclassificadores dos modelos MPC-SDP.

Como resultado, todos os modelos (MPC-SDP e *baselines*) apresentaram desempenhos preditivos insuficientes perante registros de ingresso (semestre 0), o que demonstrou, no contexto desse estudo, a incapacidade dos modelos genéricos em predições de evasão muito precoces, realizadas antes dos resultados acadêmicos de primeiro semestre e, portanto, utilizando apenas dados socioeconômicos e contextuais. Porém, considerando que diversos estudos da literatura apontam para um maior poder preditivo de dados acadêmicos, é provável que a performance dos modelos sobre dados de admissão poderia ser melhorada por meio da inclusão de dados do desempenho acadêmico pregresso dos estudantes (relativos ao histórico escolar ou ao processo de seleção para o ensino superior). Infelizmente, isso não foi possível neste estudo, considerando que, atualmente, essas informações não são alimentadas/disponibilizadas no sistema de gestão acadêmica do IFFar.

Por outro lado, as avaliações sobre dados dos demais semestres letivos demonstraram uma importante vantagem preditiva dos modelos MPC-SDP: independentemente dos experimentos e das variações técnicas consideradas, ao contrário dos *baselines*, os modelos MPC-SDP demonstraram priorizar o desempenho de revocação da classe negativa (de estudantes concluídos/diplomados) nos semestres iniciais, proporcionando-lhes, consequentemente, melhor precisão na predição das evasões (classe positiva). Considerando que os semestres iniciais concentram os maiores quantitativos de evasões e de estudantes matriculados, na prática, isso significaria oportunizar um melhor direcionamento institucional das medidas preventivas, justamente nas etapas mais críticas/importantes do acompanhamento dos estudantes. Ou seja, apesar do inegável *trade-off* entre precisão e revocação, como a limitação institucional de recursos humanos e financeiros inviabilizaria provavelmente o acompanhamento/auxílio de todos os estudantes em risco, é mais vantajoso que os modelos recuperem uma parcela menor (menor revocação), mas mais precisa, de alunos em risco de evasão (classe positiva), nos semestres iniciais.

Ainda sob a perspectiva de desempenho preditivo, mas considerando especificamente o uso da seleção FSA, proposta como parte integrante do MPC-SDP mas também avaliada sobre os modelos de predição genérica tradicionais, foi verificado um reflexo negativo sobre a performance preditiva dos *baselines*, especialmente nas avaliações sobre dados de semestres iniciais e considerando a seleção FSA+SP. Porém, os modelos MPC-SDP, em geral, demonstraram maior estabilidade preditiva frente aos diferentes métodos de seleção de atributos, sem perdas significativas associadas às

seleções FSA. Inclusive, os testes estatísticos sinalizaram, em alguns casos, uma significativa superioridade de modelos MPC-SDP com seleção FSA sobre *baselines*. E, apesar de não terem sido apontadas diferenças significativas entre os classificadores MPC-SDP com seleções FSA+SP e FSA+RFE, os modelos associados à seleção FSA+RFE se sobressaíram mais vezes sobre os *baselines*, demonstrando-se uma melhor combinação. Além disso, embora não tenham sido comparados diretamente, de forma geral, os modelos MPC-SDP com subclassificadores DT e LR apresentaram comportamentos preditivos semelhantes. Porém, é inegável que os subclassificadores DT, por serem baseados em regras de decisão, possibilitam uma interpretação mais clara e sem prejuízo de multicolinearidade.

Em relação aos padrões preditivos, confirmou-se que, de forma geral, o uso da seleção FSA contribuiu para um aumento na variabilidade de aspectos dos principais atributos/padrões preditivos, entre os subclassificadores especializados por semestre dos modelos MPC-SDP. Embora isso possa contribuir para um entendimento mais amplo/multifacetado dos padrões de evasão, assim como na literatura, também se percebeu a priorização e a maior importância preditiva de atributos acadêmicos. De forma geral, os padrões indicaram maior tendência à evasão de estudantes com menor desempenho/aproveitamento acadêmico, da área de Ciências Exatas e da Terra, e em maior vulnerabilidade socioeconômica. Além disso, considerando a maior clareza das regras de decisão dos modelos MPC-SDP baseados em DT, verificou-se que níveis de interação mais altos também podem levar à evasão, quando associados a desempenhos acadêmicos inferiores. Isso indica que um elevado nível interacional no AVA, para além de um maior interesse/engajamento, pode sinalizar que o estudante está enfrentando dificuldades no processo de ensino-aprendizagem.

A análise dos padrões preditivos também evidenciou que atributos econômicos e interacionais ganharam maior importância nos modelos desenvolvidos sobre dados relacionados ao período pandêmico. Isso confirma que, institucionalmente, o ERE impulsionou a exploração/uso das turmas virtuais/AVA, enriquecendo os registros de interação e aumentando o potencial preditivo desses dados. Com o reconhecimento das facilidades e funcionalidades disponibilizadas pelo AVA, além da naturalidade alcançada em seu uso, durante o ERE, é provável que esses recursos continuem sendo ativamente explorados por docentes e discentes, no ensino presencial. Logo, torna-se importante manter/considerar o uso de dados interacionais no desenvolvimento de modelos preditivos de evasão.

Retomando a primeira questão de pesquisa desta Tese, o uso de atributos relativos, no pré-processamento, mostrou-se uma solução viável para a representação de estudantes de diferentes tipos de cursos e períodos letivos, tornando-os comparáveis entre si e oportunizando, assim, a generalização dos padrões preditivos. Considerando a segunda questão de pesquisa, dados os resultados de desempenho preditivo,

também se mostrou satisfatória a estruturação proposta para a aplicação das técnicas de EDM e, consequentemente, o desenvolvimento do MPC-SDP. A construção de um comitê de decisão composto por subclassificadores especializados por semestres/-períodos letivos, para os quais são conduzidas, individualizadamente, as etapas de balanceamento, seleção de atributos, e otimização de hiperparâmetros, proporcionou uma maior atenção à variabilidade de padrões preditivos, ao longo dos semestres. Essa estrutura possibilitou que não apenas os padrões mais abrangentes (modelos genéricos tradicionais) ou os relacionados a um único momento de predição (modelos especializados) fossem considerados pelo MPC-SDP, na tarefa preditiva. Já em relação à terceira questão de pesquisa, a condução do processo de seleção de atributos orientado por aspecto (FSA), embora não tenha proporcionado ganho significativo de desempenho ao MPC-SDP, demonstrou incrementar a variabilidade/representatividade de aspectos entre seus principais padrões/atributos preditivos, o que pode oportunizar uma percepção mais abrangente acerca dos fatores relacionados à evasão estudantil.

8.1 Trabalhos Futuros

Diversos trabalhos podem ser conduzidos para estender as investigações realizadas nesta Tese ou suas aplicações. Visando uma ainda melhor interpretabilidade dos padrões dos subclassificadores DT, pode ser avaliada a execução de pós-poda ou o desenvolvimento de um *parser* sobre as regras preditivas, para identificar e desconsiderar condições que não são, efetivamente, distintivas (ou seja, condições/nós cujas subárvores, embora apresentem folhas com classes/decisões distintas, representam, na grande maioria, registros de uma única classe). Além disso, técnicas de XAI também podem ser avaliadas para aprofundar a explicabilidade dos resultados preditivos do comitê MPC-SDP.

Em relação ao desempenho do MPC-SDP, pode-se buscar a integração de dados oriundos do processo seletivo dos estudantes, com o intuito de enriquecer as informações pós-matrícula/admissão e, assim, melhorar os resultados de predições precoces. Além disso, outros métodos de balanceamento, seleção de atributos e classificação podem ser considerados base no treinamento do MPC-SDP, assim como incluídos no processo de otimização dos subclassificadores especializados por semestre, possibilitando ao MPC-SDP a constituição de um comitê heterogêneo. Nesse caso, aumentando a complexidade dos algoritmos e da otimização, também cabe a avaliação de outras técnicas de ajuste automatizado de hiperparâmetros, a fim de reduzir o custo da busca por uma solução ótima.

Em relação ao contexto de aplicação, a construção do MPC-SDP também pode ser considerada e avaliada sobre dados de estudantes de outros níveis, modalidades

ou instituições de ensino. Técnicas de *transfer learning* também podem ser utilizadas com o intuito de melhorar a generalização do aprendizado em diferentes contextos. Por exemplo, como no IFFar existem poucos cursos de graduação EAD, o volume de dados históricos de seus estudantes pode ser insuficiente para uma aplicação de aprendizado supervisionado tradicional. Nesse caso, poderia ser avaliada uma abordagem de *transfer learning*, para tentar aproveitar, no contexto da graduação EAD, o aprendizado alcançado por modelos voltados à graduação presencial.

Do ponto de vista prático e institucional, os padrões preditivos evidenciados nesta Tese podem auxiliar em revisões de políticas preventivas, pelo IFFar. Considerando que a evasão possui grande concentração em semestres iniciais e foi mais frequentemente associada a baixos desempenhos acadêmicos, podem ser reforçadas medidas que visem melhorar a adaptação e a integração dos novos estudantes à realidade do ensino superior. A intensificação de esforços para a promoção de programas de reforço acadêmico e de monitoria, de auxílios estudantis, e de atividades de integração social e institucional podem ser citadas como exemplo. Para melhorar o planejamento dessas estratégias preventivas, também podem ser conduzidos estudos voltados especialmente para os alunos de primeiro ano de graduação, em busca de uma análise mais atenta de seus desafios e dificuldades. Investigar, em cursos críticos (especialmente da área de Ciências Exatas e da Terra), as disciplinas com maiores índices de reprovações, por meio de análise estatística, ou comumente associadas à evasão, a partir da mineração de padrões sequenciais sobre o histórico de disciplinas cursadas pelos alunos, pode evidenciar quais disciplinas devem ser priorizadas em monitorias, por exemplo.

Além disso, o modelo proposto pode ser utilizado para compor o desenvolvimento de um sistema de predição, integrado, em tempo real, às bases de dados institucionais, para acompanhamento do risco de evasão de estudantes ativos/matriculados, pelas equipes pedagógicas e de assistência estudantil. Como o MPC-SDP é construído/treinado a partir de atributos relativos, ele pode ser utilizado para realizar predições tanto sobre registros consolidados, referentes aos dados obtidos pelo estudante até o último semestre integralizado, quanto sobre registros parciais/dinâmicos, que incluam uma simulação dos resultados/dados do estudante no semestre em andamento, com base apenas nas informações lançadas/disponíveis até o momento da coleta/predição. Além disso, junto a cada indicação de risco, o sistema pode ser implementado para apresentar indícios dos atributos que contribuíram para a predição de evasão, por meio dos valores SHAP associados às predições individuais, por exemplo. Nesse caso, além de ser uma abordagem de explicabilidade disponível para modelos baseados em qualquer algoritmo de classificação, seus resultados são de fácil interpretação e auxiliariam as equipes de acompanhamento no direcionamento de medidas preventivas individualizadas.

REFERÊNCIAS

AGRUSTI, F.; MEZZINI, M.; BONAVOLONTA, G. Deep learning approach for predicting university dropout: a case study at Roma Tre University. **Journal of e-Learning and Knowledge Society**, Rome, Italy, v.16, n.1, SI, p.44–54, Apr 2020. DOI: https://doi.org/10.20368/1971-8829/1135192.

AGUIRRE, C. E.; PéREZ, J. C. Predictive data analysis techniques applied to dropping out of university studies. In: XLVI LATIN AMERICAN COMPUTING CONFERENCE (CLEI), 2020., 2020. **Anais...** [S.l.: s.n.], 2020. p.512–521. DOI: https://doi.org/10.1109/CLEI52000.2020.00066.

ALTURKI, S.; COHAUSZ, L.; STUCKENSCHMIDT, H. Predicting Master's students' academic performance: an empirical study in Germany. **Smart Learning Environments**, TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY, v.9, n.1, 12 2022. DOI: https://doi.org/10.1186/s40561-022-00220-y.

ANDIFES; ABRUEM; SESu/MEC. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. Brasília, DF: Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras, 1996. Disponível em: <http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf/>. Acesso em: 2022-01-11.

ASSIS, B. d. S. de; OGASAWARA, E.; BARBASTEFANO, R.; CARVALHO, D. Frequent pattern mining augmented by social network parameters for measuring graduation and dropout time factors: A case study on a production engineering course. **Spcio-Economic Planning Sciences**, STE 800, 230 PARK AVE, NEW YORK, NY 10169 USA, v.81, 6 2022. DOI: https://doi.org/10.1016/j.seps.2021.101200.

ATLAS BRASIL. Atlas do Desenvolvimento Humano no Brasil: Programa das Nações Unidas para o Desenvolvimento, Instituto de Pesquisa Econômica Aplicada, Fundação João Pinheiro. **Ranking dos Munícipios do Brasil em 2010**. Disponível em: <http://www.atlasbrasil.org.br/ranking>. Acesso em: 2022-04-01.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, [S.l.], v.19, n.02, p.03–13, 2011. DOI: https://doi.org/10.5753/rbie.2011.19.02.03.

BARANYI, M.; NAGY, M.; MOLONTAY, R. Interpretable Deep Learning for University Dropout Prediction. In: ANNUAL CONFERENCE ON INFORMATION TECHNOLOGY EDUCATION, 21., 2020, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2020. p.13–19. (SIGITE '20). DOI: https://doi.org/10.1145/3368308.3415382.

BARROS, T. M.; SOUZA NETO, P. A.; SILVA, I.; GUEDES, L. A. Predictive Models for Imbalanced Data: A School Dropout Perspective. **Education Sciences**, Basel, Switzerland, v.9, n.4, Dec 2019. DOI: https://doi.org/10.3390/educsci9040275.

BARROSO, P. C. F.; OLIVEIRA, I. M.; NORONHA-SOUSA, D.; NORONHA, A.; MATEUS, C. C.; VáZQUEZ-JUSTO, E.; COSTA-LOBO, C. FATORES DE EVASÃO NO ENSINO SUPERIOR: UMA REVISÃO DE LITERATURA. **Psicologia Escolar e Educacional**, [S.l.], v.26, n.Psicol. Esc. Educ., 2022 26, p.e228736, 2022. DOI: https://doi.org/10.1590/2175-35392022228736.

BASSETTI, E.; CONTI, A.; PANIZZI, E.; TOLOMEI, G. ISIDE: Proactively Assist University Students at Risk of Dropout. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2022., 2022. **Anais...** [S.l.: s.n.], 2022. p.1776–1783. DOI: https://doi.org/10.1109/BigData55660.2022.10020920.

BEAULAC, C.; ROSENTHAL, J. S. Predicting University Students' Academic Success and Major Using Random Forests. **Research in Higher Education**, NY, USA, v.60, n.7, p.1048–1064, Nov 2019. DOI: https://doi.org/10.1007/s11162-019-09546-y.

BERKA, P.; MAREK, L. Bachelor's degree student dropouts: Who tend to stay and who tend to leave? **Studies in Educational Evaluation**, [S.l.], v.70, 2021. DOI: https://doi.org/10.1016/j.stueduc.2021.100999.

BITENCOURT, W. A.; SILVA, D. M.; CARMO XAVIER, G. do. Pode a inteligência artificial apoiar ações contra evasão escolar universitária. **Ensaio**, [S.l.], v.30, n.116, p.669 – 694, 2022. DOI: https://doi.org/10.1590/S0104-403620220003002854.

BÖTTCHER, A.; THURNER, V.; HÄFNER, T.; HERTLE, J. A Data Science-based Approach for Identifying Counseling Needs in first-year Students. In: IEEE GLOBAL ENGINEERING EDUCATION CONFERENCE (EDUCON), 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p.420–429. DOI: https://doi.org/10.1109/EDUCON46332.2021.9454042.

BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional**. Diário Oficial da República Federativa do Brasil. Brasília, DF, 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 2022-01-05.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. **Plataforma Nilo Peçanha – PNP**, 2023. Disponível em: <https://www.gov.br/mec/pt-br/pnp>. Acesso em: 2023-05-05.

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD WORKSHOP: LANGUAGES FOR DATA MINING AND MACHINE LEARNING, 2013. **Anais...** [S.l.: s.n.], 2013. p.108–122.

CASANOVA, J. R.; CASTRO-LÓPEZ, A.; BERNARDO, A. B.; ALMEIDA, L. S. The Dropout of First-Year STEM Students: Is It Worth Looking beyond Academic Achievement? **Sustainability**, [S.l.], v.15, n.2, 2023. DOI: https://doi.org/10.3390/su15021253.

CASANOVA, J. R.; CERVERO, A.; NúñEZ, J. C.; ALMEIDA, L. S.; BERNARDO, A. Factors that determine the persistence and dropout of university students. **Psicothema**, [S.l.], v.30, n.4, p.408–414, 2018. DOI: https://doi.org/10.7334/psicothema2018.155.

CECHINEL, C.; CAMARGO, S. Mineração de dados educacionais: avaliação e interpretação de modelos de classificação. In: JAQUES, P. A.; SIQUEIRA, S.; BITTENCOURT, I.; PIMENTEL, M. (Ed.). **Metodologia de Pesquisa Científica em Informática na Educação**: Abordagem Quantitativa. Porto Alegre, RS, Brasil: SBC, 2020. (Série Metodologia de Pesquisa em Informática na Educação, v.2). Disponível em: https://metodologia.ceie-br.org/livro-2.

CHAN, J. Y.-L.; LEOW, S. M. H.; BEA, K. T.; CHENG, W. K.; PHOONG, S. W.; HONG, Z.-W.; CHEN, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. **Mathematics**, [S.l.], v.10, n.8, p.1283, Apr 2022. DOI: https://doi.org/10.3390/math10081283.

CHEN, Y.; JOHRI, A.; RANGWALA, H. Running out of STEM: A Comparative Study across STEM Majors of College Students at-Risk of Dropping out Early. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 8., 2018, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2018. p.270–279. (LAK '18). DOI: https://doi.org/10.1145/3170358.3170410.

CHUNG, J. Y.; LEE, S. Dropout early warning systems for high school students using machine learning. **Children and Youth Services Review**, Oxford, England, v.96, p.346–353, Jan 2019. DOI: https://doi.org/10.1016/j.childyouth.2018.11.030.

COIMBRA, C. L.; SILVA, L. B. e.; COSTA, N. C. D. A evasão na educação superior: definições e trajetórias. **Educação e Pesquisa**, [S.l.], v.47, n.Educ. Pesqui., 2021 47, p.e228764, 2021. DOI: https://doi.org/10.1590/S1678-4634202147228764.

COLPO, M. P.; ALVES, B. C.; PEREIRA, K. S.; BRANDÃO, A. F. Z.; AGUIAR, M. S. d.; PRIMO, T. T. Attribute Selection Based on Genetic and Classification Algorithms in the Prediction of Hospitalization Need of COVID-19 Patients. In: XVII BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 2021, New York, NY, USA. **Anais...** Association for Computing Machinery, 2021. p.2:1–2:8. (SBSI 2021). DOI: https://doi.org/10.1145/3466933.3466935.

COLPO, M. P.; ALVES, B. C.; PEREIRA, K. S.; BRANDÃO, A. F. Z.; AGUIAR, M. S. d.; PRIMO, T. T. Predicting COVID-19 hospitalizations with attribute selection based on genetic and classification algorithms. **iSys - Brazilian Journal of Information Systems**, Porto Alegre, RS, Brazil, v.15, n.1, p.4:1–4:30, Apr. 2022. DOI: https://doi.org/10.5753/isys.2022.2187.

COLPO, M. P.; PRIMO, T. T.; AGUIAR, M. S. d. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: XXXII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2021, Porto Alegre, RS, Brasil. **Anais...** SBC, 2021. p.873–884. DOI: https://doi.org/10.5753/sbie.2021.218517.

COLPO, M. P.; PRIMO, T. T.; AGUIAR, M. S. de. Lessons learned from the student dropout patterns on COVID-19 pandemic: An analysis supported by machine learning. **British Journal of Educational Technology**, [S.l.], v.55, n.2, p.560–585, 2024. DOI: https://doi.org/10.1111/bjet.13380.

COLPO, M. P.; PRIMO, T. T.; AGUIAR, M. S. de; CECHINEL, C. Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges. **Revista Brasileira de Informática na Educação**, [S.l.], v.32, 2024. No prelo.

COLPO, M. P.; PRIMO, T. T.; PERNAS, A. M.; CECHINEL, C. Mineração de Dados Educacionais na Previsão de Evasão: uma RSL sob a Perspectiva do Congresso Brasileiro de Informática na Educação. In: XXXI SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2020, Porto Alegre, RS, Brasil. **Anais...** SBC, 2020. p.1102–1111. DOI: https://doi.org/10.5753/cbie.sbie.2020.1102.

COSTA, A. G.; MATTOS, J. C. B.; PRIMO, T. T.; CECHINEL, C.; MUñOZ, R. Model for Prediction of Student Dropout in a Computer Science Course. In: XVI LATIN AMERICAN CONFERENCE ON LEARNING TECHNOLOGIES (LACLO), 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p.137–143. DOI: https://doi.org/10.1109/LACLO54177.2021.00020.

COSTA, E.; BAKER, R.; AMORIM, L.; MAGALHãES, J.; MARINHO, T. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. In: ISOTANI, S.; CAMPOS, F. C. A. (Ed.). **Jornada de Atualização em Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2013. v.1, p.1–29.

COSTA, J.; SILVEIRA, F. G.; COSTA, R.; WALTENBERG, F. Expansão da educação superior e progressividade do investimento público. In: **Texto para Discussão**. Brasília, DF: Instituto de Pesquisa Econômica Aplicada – Ipea, 2021. n.2631. DOI: http://dx.doi.org/10.38116/td2631.

CRESPO, C. Two Become One: Improving the Targeting of Conditional Cash Transfers with a Predictive Model of School Dropout. **Economia-Journal of the Latin American and Caribbean Economic Association**, LSE Press, LSE Library, 10 Portugal Street, London, ENGLAND, v.21, n.1, p.1–45, 2020. DOI: https://doi.org/10.1353/eco.2020.0011.

da SILVA, P. M.; LIMA, M. N. C. A.; SOARES, W. L.; SILVA, I. R. R.; FAGUNDES, R. A.; de SOUZA, F. F. Ensemble Regression Models Applied to Dropout in Higher Education. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p.120–125. DOI: https://doi.org/10.1109/BRACIS.2019.00030.

DEHO, O. B.; ZHAN, C.; LI, J.; LIU, J.; LIU, L.; LE, T. D. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? **British Journal of Education Technology**, 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, v.53, n.4, p.822–843, 7 2022. DOI: https://doi.org/10.1111/bjet.13217.

DEL BONIFRO, F.; GABBRIELLI, M.; LISANTI, G.; ZINGARO, S. Student dropout prediction. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, [S.l.], v.12163 LNAI, p.129–140, 2020. DOI: https://doi.org/10.1007/978-3-030-52237-7_11.

DEMETER, E.; DORODCHI, M.; AL-HOSSAMI, E.; BENEDICT, A.; WALKER, L. S.; SMAIL, J. Predicting first-time-in-college students' degree completion outcomes. **Higher Education**, [S.l.], v.84, n.3, p.589–609, 9 2022. DOI: https://doi.org/10.1007/s10734-021-00790-9.

DTI/IFFAR. Diretoria de Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha. **Manual SIGAA: Portal do Discente – Matrícula On-Line**. Disponível em: <https://sig.iffarroupilha.edu.br/sigrh/downloadArquivo?idArquivo=59088&key=d4772368508818834352dfda46e667b>. Acesso em: 2022-03-29.

FERNáNDEZ-GARCíA, A. J.; PRECIADO, J. C.; MELCHOR, F.; RODRIGUEZ-ECHEVERRIA, R.; CONEJERO, J. M.; SáNCHEZ-FIGUEROA, F. A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. **IEEE Access**, [S.l.], v.9, p.133076–133090, 2021. DOI: https://doi.org/10.1109/ACCESS.2021.3115851.

FLORES, V.; HERAS, S.; JULIAN, V. Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. **Electronics**, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, v.11, n.3, 2 2022. DOI: https://doi.org/10.3390/electronics11030457.

FONTANA, L.; MASCI, C.; IEVA, F.; PAGANONI, A. M. Performing Learning Analytics via Generalised Mixed-Effects Trees. **Data**, [S.l.], v.6, n.7, 7 2021. DOI: https://doi.org/10.3390/data6070074.

FREITAS, F. A. d. S.; VASCONCELOS, F. F. X.; PEIXOTO, S. A.; HASSAN, M. M.; DEWAN, M. A. A.; de ALBUQUERQUE, V. H. C.; REBOUCAS FILHO, P. P. IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data. **Electronics**, Basel, Switzerland, v.9, n.10, Oct 2020. DOI: https://doi.org/10.3390/electronics9101613.

GAMAO, A. O.; GERARDO, B. D. Prediction-based model for student dropouts using modified mutated firefly algorithm. **International Journal of Advanced Trends in Computer Science and Engineering**, [S.l.], v.8, n.6, p.3461–3469, 2019. DOI: https://doi.org/10.30534/ijatcse/2019/122862019.

GOMES, I.; FERREIRA, I. PNAD Contínua – Em 2022, analfabetismo cai, mas continua mais alto entre idosos, pretos e pardos e no Nordeste. **Agência de Notícias - IBGE**, [S.l.], jun 2023. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/37089-em-2022-analfabetismo-cai-mas-continua-mais-alto-entre-idosos-pretos-e-pardos-e-no-nordeste/>. Acesso em: 2023-08-14.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd.ed. Waltham, MA: Morgan Kaufmann Publishers, 2012. 560p.

HANNAFORD, L.; CHENG, X.; KUNES-CONNELL, M. Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study*. **Nurse Education Today**, Midlothian, Scotland, v.99, Apr 2021. DOI: https://doi.org/10.1016/j.nedt.2021.104784.

HOFFAIT, A.-S.; SCHYNS, M. Early detection of university students with potential difficulties. **Decision Support Systems**, Amsterdam, Netherlands, v.101, p.1–11, Sep 2017. DOI: https://doi.org/10.1016/j.dss.2017.05.003.

HUTAGAOL, N.; SUHARJITO. Predictive modelling of student dropout using ensemble classifier method in higher education. **Advances in Science, Technology and Engineering Systems**, [S.l.], v.4, n.4, p.206–211, 2019. DOI: https://doi.org/10.25046/aj040425.

IAM-ON, N.; BOONGOEN, T. Generating descriptive model for student dropout: a review of clustering approach. **Human-Centric Computing and Information Sciences**, London, England, v.7, Jan 2 2017. DOI: https://doi.org/10.1186/s13673-016-0083-0.

IAM-ON, N.; BOONGOEN, T. Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. **International Journal of Machine Learning and Cybernetics**, Heidelberg, Germany, v.8, n.2, p.497–510, Apr 2017. DOI: https://doi.org/10.1007/s13042-015-0341-x.

IFFAR. Instituto Federal de Educação, Ciência e Tecnologia Farroupilha. **Resolução CONSUP Nº 178/2014, de 28 de novembro de 2014. Aprova o Projeto do Programa Permanência e Êxito dos Estudantes do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha**. Santa Maria, RS, 2014. Disponível em: <https://www.iffarroupilha.edu.br/component/k2/attachments/download/20928/678063b3d55f50113928e95f6ce93fe6>. Acesso em: 2022-01-07.

IFFAR. Instituto Federal de Educação, Ciência e Tecnologia Farroupilha. **Plano de Desenvolvimento Institucional 2019-2026**, 2019. Disponível em: <https://www.iffarroupilha.edu.br/component/k2/attachments/download/19776/7400a07627ff8bd98a8aa0ca7b06e2ab>. Acesso em: 2022-01-07.

IFFAR. Instituto Federal de Educação, Ciência e Tecnologia Farroupilha. **Relatório de Gestão 2021**, 2022. Disponível em: <https://www.iffarroupilha.edu.br/component/k2/attachments/download/31139/047ad136551548577cb349f400676685>. Acesso em: 2023-12-07.

IFFAR. Instituto Federal de Educação, Ciência e Tecnologia Farroupilha. **Relatório de**

Gestão 2022, 2023. Disponível em: <https://sig.iffarroupilha.edu.br/public/jsp/processos/info_documento.jsf?idDoc=512445>. Acesso em: 2023-12-07.

INEP. **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira/Inep, Ministério da Educação/MEC, 2017. Disponível em: <https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf/>. Acesso em: 2022-01-11.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior 2022 - Divulgação dos resultados**, 2023. Disponível em: <https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2022/apresentacao_censo_da_educacao_superior_2022.pdf>. Acesso em: 2024-01-06.

IPEA. Instituto de Pesquisa Econômica Aplicada. **Ipeadata - Série histórica do salário mínimo vigente**, 2023. Disponível em: <http://www.ipeadata.gov.br/exibeserie.aspx?stub=1&serid1739471028=1739471028>. Acesso em: 2023-04-06.

JARDIM, C. C. Apresentação. In: GARCEZ, C. L.; HAIGERT, C. G.; BATALHA, D. V.; UBERTI, H. G. (Ed.). **IFFar 10 anos**: ensaios dessa trajetória. Santa Maria, RS: Instituto Federal Farroupilha - IFFar, 2018. p.11–14.

KANG, K.; WANG, S. Analyze and Predict Student Dropout from Online Programs. In: INTERNATIONAL CONFERENCE ON COMPUTE AND DATA ANALYSIS, 2., 2018, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2018. p.6–12. (ICCDA 2018). DOI: https://doi.org/10.1145/3193077.3193090.

KARIMI-HAGHIGHI, M.; CASTILLO, C.; HERNANDEZ-LEO, D. A Causal Inference Study on the Effects of First Year Workload on the Dropout Rate of Undergraduates. In: ARTIFICIAL INTELLIGENCE IN EDUCATION, PT I, 2022. **Proceedings Paper...** [S.l.: s.n.], 2022. p.15–27. (Lecture Notes in Computer Science, v.13355). DOI: https://doi.org/10.1007/978-3-031-11644-5_2.

KISS, B.; NAGY, M.; MOLONTAY, R.; CSABAY, B. Predicting Dropout Using High School and First-semester Academic Achievement Measures. In: INTERNATIONAL CONFERENCE ON EMERGING ELEARNING TECHNOLOGIES AND APPLICATIONS (ICETA), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p.383–389. DOI: https://doi.org/10.1109/ICETA48886.2019.9040158.

KITCHENHAM, B. A.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Keele, UK: School of Computer Science

and Mathematics, Keele University, 2007. (EBSE-2007-01). Disponível em: https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf.

KURNIAWATI, G.; MAULIDEVI, N. U. Multivariate Sequential Modelling for Student Performance and Graduation Prediction. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY, COMPUTER, AND ELECTRICAL ENGINEERING (ICITACEE), 2022., 2022. **Anais...** [S.l.: s.n.], 2022. p.293–298. DOI: https://doi.org/10.1109/ICITACEE55701.2022.9923971.

KUZILEK, J.; ZDRAHAL, Z.; FUGLIK, V. Student success prediction using student exam behaviour. **Future Generation Computer Systems - The International Journal of Escience**, [S.l.], v.125, p.661–671, 12 2021. DOI: https://doi.org/10.1016/j.future.2021.07.009.

LEE, S.; CHUNG, J. Y. The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. **Applied Sciences-Basel**, Basel, Switzerland, v.9, n.15, 2019. DOI: https://doi.org/10.3390/app9153093.

LEMAîTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. **Journal of Machine Learning Research**, [S.l.], v.18, n.17, p.1–5, 2017.

LOTTERING, R.; HANS, R.; LALL, M. A model for the identification of students at risk of dropout at a university of technology. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, BIG DATA, COMPUTING AND DATA COMMUNICATION SYSTEMS (ICABCD), 2020., 2020. **Anais...** [S.l.: s.n.], 2020. p.1–8. DOI: https://doi.org/10.1109/icABCD49160.2020.9183874.

MASOOD, S. W.; BEGUM, S. A. Comparison of Resampling Techniques for Imbalanced Datasets in Student Dropout Prediction. In: IEEE SILCHAR SUBSECTION CONFERENCE (SILCON), 2022., 2022. **Anais...** [S.l.: s.n.], 2022. p.1–7. DOI: https://doi.org/10.1109/SILCON55242.2022.10028915.

MDUMA, N.; KALEGELE, K.; MACHUVE, D. An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools. **Journal of Information Systems Engineering and Management**, [S.l.], v.4, n.3, 2019. DOI: https://doi.org/10.29333/jisem/5893.

MDUMA, N.; MACHUVE, D. Machine Learning Model for Predicting Student Dropout: A Case of Tanzania, Kenya and Uganda. In: IEEE AFRICON, 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p.1–6. DOI: https://doi.org/10.1109/AFRICON51333.2021.9570956.

MORAES, G. H.; ALMEIDA, S. C. L. d.; ALVES, T. E.; LIMA, M. V. S. F.; FILHO, R. P. R.; JULIATTO, M. A.; BOTELHO, A. d. F.; GODOY, D. F. d.; GALLINDO, E. d. L.; RIBEIRO, F. P.; KENCHIAN, G.; SOUZA, I. R. M. d. S.; SILVA, J. C.; LEãO, P. H. A.; BERMEJO, P. H. d. S.; SILVA, S. S. d.; JUNIOR, W. T. d. S. **Plataforma Nilo Peçanha**: Guia de Referência Metodológica. DDR/SETEC/MEC. Ministério da Educação. Brasília, DF: Editora Evobiz, 2018. 101p.

NAGY, M.; MOLONTAY, R. Predicting Dropout in Higher Education Based on Secondary School Performance. In: IEEE 22ND INTERNATIONAL CONFERENCE ON INTELLIGENT ENGINEERING SYSTEMS (INES), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p.389–394. DOI: https://doi.org/10.1109/INES.2018.8523888.

NASEEM, M.; CHAUDHARY, K.; SHARMA, B. Predicting Freshmen Attrition in Computing Science using Data Mining. **Education and Information Technologies**, [S.l.], v.27, n.7, p.9587 – 9617, 2022. DOI: https://doi.org/10.1007/s10639-022-11018-3.

NUANMEESRI, S.; POOMHIRAN, L.; CHOPVITAYAKUN, S.; KADMATEEKARUN, P. Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network. **International Journal of Information and Education Technology**, [S.l.], v.12, n.9, p.851 – 857, 2022. DOI: https://doi.org/10.18178/ijiet.2022.12.9.1693.

OPAZO, D.; MORENO, S.; ALVAREZ-MIRANDA, E.; PEREIRA, J. Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities. **Mathematics**, [S.l.], v.9, n.20, 10 2021. DOI: https://doi.org/10.3390/math9202599.

ORESHIN, S.; FILCHENKOV, A.; PETRUSHA, P.; KRASHENINNIKOV, E.; PANFILOV, A.; GLUKHOV, I.; KALIBERDA, Y.; MASALSKIY, D.; SERDYUKOV, A.; KAZAKOVTSEV, V.; KHLOPOTOV, M.; PODOLENCHUK, T.; SMETANNIKOV, I.; KOZLOVA, D. Implementing a Machine Learning Approach to Predicting Students Academic Outcomes. In: 2020. **Anais...** [S.l.: s.n.], 2020. p.78 – 83. (ACM International Conference Proceeding Series). DOI: https://doi.org/10.1145/3437802.3437816.

OROOJI, M.; CHEN, J. Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques. In: IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS (ICMLA), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p.456–461. DOI: https://doi.org/10.1109/ICMLA.2019.00085.

ORTIGOSA, A.; CARRO, R. M.; BRAVO-AGAPITO, J.; LIZCANO, D.; ALCOLEA, J. J.; BLANCO, O. From Lab to Production Lessons Learnt and Real-Life Challenges of an

Early Student-Dropout Prevention System. **IEEE Transactions on Learning Technologies**, [S.l.], v.12, n.2, p.264–277, April 2019. DOI: https://doi.org/10.1109/TLT.2019.2911608.

OSÓRIO, A. C. d. N. "Educação/Education. In: **Brasil em Números/Brazil in Figures**. Rio de Janeiro, RJ: Instituto Brasileiro de Geografia e Estatístico - IBGE, 2017. v.25, p.149–171.

PACHAS, D. A. G.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, A. J.; CAMARA-CHAVEZ, G.; POCO, J.; GOMEZ-NIETO, E. A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk. In: XLVII LATIN AMERICAN COMPUTING CONFERENCE (CLEI), 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p.1–10. DOI: https://doi.org/10.1109/CLEI53233.2021.9640119.

PALACIOS, C. A.; REYES-SUAREZ, J. A.; BEARZOTTI, L. A.; LEIVA, V.; MARCHANT, C. Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. **Entropy**, Basel, Switzerland, v.23, n.4, Apr 2021. DOI: https://doi.org/10.3390/e23040485.

PARK, H. S.; YOO, S. J. Early Dropout Prediction in Online Learning of University using Machine Learning. **International Journal on Informatics Visualization**, [S.l.], v.5, n.4, p.347 – 353, 2021. DOI: https://doi.org/10.30630/JOIV.5.4.732.

PERCHINUNNO, P.; BILANCIA, M.; VITALE, D. A Statistical Analysis of Factors Affecting Higher Education Dropouts. **Social Indicators Research**, Dordrecht, Netherlands, Dec 2019. DOI: https://doi.org/10.1007/s11205-019-02249-y.

PEREZ, B.; CASTELLANOS, C.; CORREAL, D. Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study. In: APPLICATIONS OF COMPUTATIONAL INTELLIGENCE, ColCACI 2018, 2018. **Anais...** [S.l.: s.n.], 2018. p.111–125. (Communications in Computer and Information Science, v.833). DOI: https://doi.org/10.1007/978-3-030-03023-0_10.

PNUD Brasil. Programa das Nações Unidas para o Desenvolvimento. **O que é o IDHM**. Disponível em: <https://www.br.undp.org/content/brazil/pt/home/idh0/conceitos/o-que-e-o-idhm.html>. Acesso em: 2022-04-01.

PONTILI, R.; STADUTO, J.; HENRIQUE, J. Abandono e atraso escolar e sua relação com indicadores socioeconômicos: uma análise para a região Sul do Brasil. **Gestão & Regionalidade**, [S.l.], v.34, n.101, p.4–22, 2018. DOI: https://doi.org/10.13037/gr.vol34n101.4173.

PRADA, M. A.; DOMÍNGUEZ, M.; VICARIO, J. L.; ALVES, P. A. V.; BARBU, M.; PODPORA, M.; SPAGNOLINI, U.; PEREIRA, M. J. V.; VILANOVA, R. Educational Data Mining for Tutoring Support in Higher Education: A Web-Based Tool Case Study in Engineering Degrees. **IEEE Access**, [S.l.], v.8, p.212818–212836, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3040858.

PYTHON. **The multiprocessing package – Process-based parallelism**. Python Software Foundation. Disponível em: <https://docs.python.org/3/library/multiprocessing.html?highlight=multiprocessing#module-multiprocessing>. Acesso em: 2022-02-16.

QUEIROGA, E. M.; BATISTA MACHADO, M. F.; PARAGARINO, V. R.; PRIMO, T. T.; CECHINEL, C. Early Prediction of At-Risk Students in Secondary Education: A Countrywide K-12 Learning Analytics Initiative in Uruguay. **Information**, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, v.13, n.9, 9 2022. DOI: https://doi.org/10.3390/info13090401.

QUEIROGA, E. M.; LOPES, J. L.; KAPPEL, K.; AGUIAR, M.; ARAUJO, R. M.; MUNOZ, R.; VILLARROEL, R.; CECHINEL, C. A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. **Applied Sciences-Basel**, Basel, Switzerland, v.10, n.11, Jun 2020. DOI: https://doi.org/10.3390/app10113998.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow**. 2nd.ed. Birmingham, UK: Packt Publishing, 2017. 622p.

RIGO, S.; CAMBRUZZI, W.; BARBOSA, J.; CAZELLA, S. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, [S.l.], v.22, n.01, p.132, 2014. DOI: https://doi.org/10.5753/rbie.2014.22.01.132.

ROMERO, C.; ROMERO, J. R.; VENTURA, S. A Survey on Pre-Processing Educational Data. In: PEÑA-AYALA, A. (Ed.). **Educational Data Mining**: Applications and Trends. Cham, Switzerland: Springer International Publishing, 2014. p.29–64. (Studies in Computational Intelligence). DOI: https://doi.org/10.1007/978-3-319-02738-8_2.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **WIREs Data Mining and Knowledge Discovery**, [S.l.], v.10, n.3, p.e1355, 2020. DOI: https://doi.org/10.1002/widm.1355.

RONDADO de SOUSA, L.; OLIVEIRA de CARVALHO, V.; PENTEADO, B. E.; AFFONSO, F. J. A Systematic Mapping on the Use of Data Mining for the Face-to-Face School Dropout Problem. In: INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED EDUCATION - VOLUME 1: CSEDU, 13., 2021. **Proceedings...** SciTePress, 2021. p.36–47. GS SEARCH.

ROVIRA, S.; PUERTAS, E.; IGUAL, L. Data-driven system to predict academic grades and dropout. **PLOS ONE**, CA, USA, v.12, n.2, Feb 14 2017. DOI: https://doi.org/10.1371/journal.pone.0171207.

SANTOS, G. A. S.; BELLOZE, K. T.; TARRATACA, L.; HADDAD, D. B.; BORDIGNON, A. L.; BRANDAO, D. N. EvolveDTree Analyzing Student Dropout in Universities. In: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING (IWSSIP), 2020., 2020. **Anais...** [S.l.: s.n.], 2020. p.173–178. DOI: https://doi.org/10.1109/IWSSIP48289.2020.9145203.

SANTOS, J. R.; ZABOROSKI, E. Ensino Remoto e Pandemia de CoViD-19: Desafios e oportunidades de alunos e professores. **Revista Interacções**, [S.l.], v.16, n.55, p.41–57, Dez. 2020.

SCIKIT-LEARN. **Precision-Recall**. Disponível em: <https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html>. Acesso em: 2022-01-25.

SCIKIT-LEARN. **API Reference**. Disponível em: <https://scikit-learn.org/stable/modules/classes.html>. Acesso em: 2023-04-18.

SEGURA, M.; MELLO, J.; HERNANDEZ, A. Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? **Mathematics**, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, v.10, n.18, 9 2022. DOI: https://doi.org/10.3390/math10183359.

SEMESP Instituto. **Mapa do Ensino Superior no Brasil, 13a edição/2023**. Disponível em: <https://www.semesp.org.br/wp-content/uploads/2023/06/mapa-do-ensino-superior-no-brasil-2023.pdf>. Acesso em: 2023-06-30.

SETEC/MEC. **Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica**. Brasília, DF: Ministério da Educação, Secretaria de Educação Profissional e Tecnológica, 2014. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=110401-documento-orientador-evasao-retencao-vfinal&category_slug=abril-2019-pdf&Itemid=30192/>. Acesso em: 2022-01-11.

SHIAU, Y. University Dropout Prevention through the Application of Big Data. In: INTERNATIONAL CONFERENCE ON INFORMATION MANAGEMENT AND MANAGEMENT SCIENCE, 2020., 2020, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2020. p.1–7. (IMMS 2020). DOI: https://doi.org/10.1145/3416028.3416029.

SHILBAYEH, S.; ABONAMAH, A. Predicting Student Enrolments and Attrition Patterns in Higher Educational Institutions using Machine Learning. **International Arab Journal of Information Technology**, [S.l.], v.18, n.4, p.562–567, 7 2021. DOI: https://doi.org/10.34028/18/4/8.

SILVA FILHO, R. L. L.; MOTEJUNAS, P. R.; HIPOLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, São Paulo, v.37, n.132, p.641–659, 2007. DOI: https://doi.org/10.1590/S0100-15742007000300007.

SILVA, G. P. d. Análise de evasão no ensino superior: uma proposta de diagnóstico de seus determinantes. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, [S.l.], v.18, n.2, p.311–333, 2013. DOI: https://doi.org/10.1590/S1414-40772013000200005.

SILVA, L. B. e.; MARIANO, A. S. A DEFINIÇÃO DE EVASÃO E SUAS IMPLICAÇÕES (LIMITES) PARA AS POLÍTICAS DE EDUCAÇÃO SUPERIOR. **Educação em Revista**, [S.l.], v.37, n.Educ. rev., 2021 37, p.e26524, 2021. DOI: https://doi.org/10.1590/0102-469826524.

SINFO/UFRN. Superintendência de Informática da Universidade Federal do Rio Grande do Norte. **Manuais SIGAA: Relatório dos Índices Acadêmicos do Aluno**. Disponível em: <https://docs.info.ufrn.br/doku.php?id=suporte:manuais:sigaa:graduacao:relatorios_daca:alunos:listagens:lista_de_alunos:relatorio_dos_indices_academicos_do_aluno>. Acesso em: 2022-03-28.

SINFO/UFRN. Superintendência de Informática da Universidade Federal do Rio Grande do Norte. **Manuais SIGAA: Portal do Discente – Meus Dados Pessoais**. Disponível em: <https://docs.info.ufrn.br/doku.php?id=suporte:manuais:sigaa:portal_do_discente:meus_dados_pessoais>. Acesso em: 2022-03-29.

SOLIS, M.; MOREIRA, T.; GONZALEZ, R.; FERNANDEZ, T.; HERNANDEZ, M. Perspectives to Predict Dropout in University Students with Machine Learning. In: IEEE INTERNATIONAL WORK CONFERENCE ON BIOINSPIRED INTELLIGENCE (IWOBI), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p.1–6. DOI: https://doi.org/10.1109/IWOBI.2018.8464191.

SORENSEN, L. C. “Big Data” in Educational Administration: An Application for Predicting School Dropout Risk. **Educational Administration Quarterly**, CA, USA, v.55, n.3, p.404–446, Aug 2019. DOI: https://doi.org/10.1177/0013161X18799439.

STATSMODELS. **statsmodels 0.14.0 - User Guide - Variance Inflation Factor**. Disponível em: <https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html>. Acesso em: 2023-09-18.

STI/UFRN. Superintendência de Tecnologia da Informação da Universidade Federal do Rio Grande do Norte. **Portal de Cooperação – Instituições Parceiras**. Disponível em: <https://portalcooperacao.info.ufrn.br/pagina.php?a=parceiros>. Acesso em: 2022-03-28.

SUNDUS, K. I.; HAMMO, B. H.; AL-ZOUBI, M. B.; AL-OMARI, A. Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. **Informatics in Medicine Unlocked**, [S.l.], v.33, p.101088, 2022. DOI: https://doi.org/10.1016/j.imu.2022.101088.

TINTO, V. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. **Review of Educational Research**, [S.l.], v.45, n.1, p.89–125, 1975. DOI: https://doi.org/10.1590/2175-35392022228736.

TSAI, S.-C.; CHEN, C.-H.; SHIAO, Y.-T.; CIOU, J.-S.; WU, T.-N. Precision education with statistical learning and deep learning: a case study in Taiwan. **International Journal of Educational Technology in Higher Education**, NY, USA, v.17, n.1, Apr 8 2020. DOI: https://doi.org/10.1186/s41239-020-00186-2.

UNDP United Nations Development Programme. Human Development Reports. **Human Development Index (HDI)**. Disponível em: <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>. Acesso em: 2023-12-27.

URBINA-NAJERA, A. B.; MENDEZ-ORTEGA, L. A. Predictive Model for Taking Decision to Prevent University Dropout. **International Journal of Interactive Multimedia and Artificial Intelligence**, [S.l.], v.7, n.4, p.205–213, 6 2022. DOI: https://doi.org/10.9781/ijimai.2022.01.006.

VEGA, H.; SANEZ, E.; CRUZ, P. De La; MOQUILLAZA, S.; PRETELL, J. Intelligent System to Predict University Students Dropout. **International journal of online and biomedical engineering**, [S.l.], v.18, n.7, p.27 – 43, 2022. DOI: https://doi.org/10.3991/ijoe.v18i07.30195.

VERDUGO, J. V.; GITIAUX, X.; ORTEGA, C.; RANGWALA, H. FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context. In:

LAK22: 12TH INTERNATIONAL LEARNING ANALYTICS AND KNOWLEDGE CONFERENCE, 2022, New York, NY, USA. **Anais...** Association for Computing Machinery, 2022. p.271–281. (LAK22). DOI: https://doi.org/10.1145/3506860.3506902.

VILLEGAS-CH, W.; PALACIOS-PACHECO, X.; LUJAN-MORA, S. A Business Intelligence Framework for Analyzing Educational Data. **Sustainability**, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, v.12, n.14, 7 2020. DOI: https://doi.org/10.3390/su12145745.

VILORIA, A.; GARCIA PADILLA, J.; VARGAS-MERCADO, C.; HERNANDEZ-PALMA, H.; ORELLANO LLINAS, N.; ARROZOLA DAVID, M. Integration of Data Technology for Analyzing University Dropout. In: INTERNATIONAL CONFERENCE ON MOBILE SYSTEMS AND PERVASIVE COMPUTING (MOBISPC 2019), 14TH INTERNATIONAL CONFERENCE ON FUTURE NETWORKS AND COMMUNICATIONS (FNC-2019), 9TH INTERNATIONAL CONFERENCE ON SUSTAINABLE ENERGY INFORMATION TECHNOLOGY, 16., 2019. **Anais...** [S.l.: s.n.], 2019. p.569–574. (Procedia Computer Science, v.155). DOI: https://doi.org/10.1016/j.procs.2019.08.079.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: Practical Machine Learning Tools and Techniques. 3rd.ed. Burlington, MA, USA: Morgan Kaufmann Publishers Inc., 2011.

XU, Y.; WILSON, K. Early Alert Systems During a Pandemic: A Simulation Study on the Impact of Concept Drift. In: LAK21: 11TH INTERNATIONAL LEARNING ANALYTICS AND KNOWLEDGE CONFERENCE, 2021, New York, NY, USA. **Anais...** Association for Computing Machinery, 2021. p.504–510. (LAK21). DOI: https://doi.org/10.1145/3448139.3448190.

YANG, H.; OLSON, T. W.; PUDER, A. Analyzing Computer Science Students' Performance Data to Identify Impactful Curricular Changes. In: IEEE FRONTIERS IN EDUCATION CONFERENCE (FIE), 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p.1–9. DOI: https://doi.org/10.1109/FIE49875.2021.9637474.

YOO, J. S.; WOO, Y.-S.; PARK, S. J. Mining Course Trajectories of Successful and Failure Students: A Case Study. In: IEEE INTERNATIONAL CONFERENCE ON BIG KNOWLEDGE (ICBK), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p.270–275. DOI: https://doi.org/10.1109/ICBK.2017.55.

YU, R.; LEE, H.; KIZILCEC, R. F. Should College Dropout Prediction Models Include Protected Attributes? In: EIGHTH ACM CONFERENCE ON LEARNING @ SCALE, 2021, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2021. p.91–100. (L@S '21). DOI: https://doi.org/10.1145/3430895.3460139.

Apêndices

APÊNDICE A – Artigos selecionados na RSL

Tabela 26 – Lista dos artigos selecionados na RSL

#	Referência	País(es) de Afiliação	Base
1	Baranyi; Nagy; Molontay (2020)	Hungria	ACM
2	Chen; Johri; Rangwala (2018)	EUA	ACM
3	Kang; Wang (2018)	EUA	ACM
4	Shiau (2020)	China	ACM
5	Verdugo et al. (2022)	Chile, EUA	ACM
6	Xu; Wilson (2021)	EUA	ACM
7	Yu; Lee; Kizilcec (2021)	EUA	ACM
8	Aguirre; Pérez (2020)	Equador, Espanha	IEEE
9	Bassetti et al. (2022)	Itália	IEEE
10	Böttcher et al. (2021)	Alemanha	IEEE
11	Costa et al. (2021)	Brasil, Chile	IEEE
12	da Silva et al. (2019)	Brasil	IEEE
13	Fernández-García et al. (2021)	Espanha	IEEE
14	Kiss et al. (2019)	Hungria	IEEE
15	Kurniawati; Maulidevi (2022)	Indonésia	IEEE
16	Lottering; Hans; Lall (2020)	África do Sul	IEEE
17	Masood; Begum (2022)	Índia	IEEE
18	Mduma; Machuve (2021)	Tanzânia	IEEE
19	Nagy; Molontay (2018)	Hungria	IEEE
20	Orooji; Chen (2019)	EUA	IEEE
21	Ortigosa et al. (2019)	Espanha	IEEE
22	Pachas et al. (2021)	Brasil, Peru	IEEE
23	Prada et al. (2020)	Itália, Polônia, Portugal, Romênia, Espanha	IEEE
24	Santos et al. (2020)	Brasil	IEEE
25	Solis et al. (2018)	Costa Rica	IEEE
26	Yang; Olson; Puder (2021)	EUA	IEEE
27	Yoo; Woo; Park (2017)	EUA	IEEE
28	Alturki; Cohausz; Stuckenschmidt (2022)	Alemanha	WoS
29	Barros et al. (2019)	Brasil	WoS
30	Beaulac; Rosenthal (2019)	Canadá	WoS
31	Berka; Marek (2021)	República Checa	WoS
32	Chung; Lee (2019)	Coreia do Sul, EUA	WoS
33	Crespo (2020)	Reino Unido	WoS
34	Assis et al. (2022)	Brasil	WoS
35	Deho et al. (2022)	Austrália	WoS

Fonte: Traduzida de Colpo et al. (2024).

Tabela 27 – Lista dos artigos selecionados na RSL (Continuação da Tabela 26)

#	Referência	País(es) de Afiliação	Base
36	Del Bonifro et al. (2020)	Itália, França	WoS
37	Demeter et al. (2022)	EUA	WoS
38	Flores; Heras; Julian (2022)	Peru, Espanha	WoS
39	Fontana et al. (2021)	Itália	WoS
40	Freitas et al. (2020)	Brasil, Arábia Saudita, Canadá	WoS
41	Hannaford; Cheng; Kunes-Connell (2021)	EUA	WoS
42	Hoffait; Schyns (2017)	Bélgica	WoS
43	Iam-On; Boongoen (2017a)	Tailândia	WoS
44	Iam-On; Boongoen (2017b)	Tailândia	WoS
45	Karimi-Haghighi; Castillo; Hernandez-Leo (2022)	Espanha	WoS
46	Kuzilek; Zdrahal; Fuglik (2021)	República Checa, Alemanha	WoS
47	Lee; Chung (2019)	EUA, Coreia do Sul	WoS
48	Opazo et al. (2021)	Chile	WoS
49	Palacios et al. (2021)	Chile	WoS
50	Perchinunno; Bilancia; Vitale (2019)	Itália	WoS
51	Perez; Castellanos; Correal (2018)	Colômbia	WoS
52	Queiroga et al. (2022)	Brasil, Uruguai	WoS
53	Queiroga et al. (2020)	Brasil, Chile	WoS
54	Segura; Mello; Hernandez (2022)	Espanha, Paraguai	WoS
55	Shilbayeh; Abonamah (2021)	Emirados Árabes Unidos	WoS
56	Sorensen (2019)	EUA	WoS
57	Tsai et al. (2020)	Taiwan	WoS
58	Urbina-Najera; Mendez-Ortega (2022)	México	WoS
59	Villegas-Ch; Palacios-Pacheco; Lujan-Mora (2020)	Equador, Espanha	WoS
60	Viloria et al. (2019)	Colômbia	WoS
61	Agrusti; Mezzini; Bonavolonta (2020)	Itália	Scopus
62	Bitencourt; Silva; Carmo Xavier (2022)	Brasil	Scopus
63	Gamao; Gerardo (2019)	Filipinas	Scopus
64	Hutagaol; Suharjito (2019)	Indonésia	Scopus
65	Mduma; Kalegele; Machuve (2019)	Tanzânia	Scopus
66	Naseem; Chaudhary; Sharma (2022)	Fiji	Scopus
67	Nuanmeesri et al. (2022)	Tailândia	Scopus
68	Oreshin et al. (2020)	Rússia	Scopus
69	Park; Yoo (2021)	Coreia do Sul	Scopus
70	Rovira; Puertas; Igual (2017)	Espanha	Scopus
71	Vega et al. (2022)	Peru	Scopus

Fonte: Traduzida de Colpo et al. (2024).

APÊNDICE B – Relação dos cursos envolvidos na extração

		AL	FW	JA	JC	PB	SR	SAN	SA	SB	SVS
LICENCIATURA	*Educação do Campo - Ciências Agrárias			✓							
	*Educação do Campo - Ciências Biológicas			✓							
	Ciências Biológicas	✓			✓	✓	✓		✓		✓
	Computação							✓	✓		
	Física									✓	
	Matemática	✓	✓		✓		✓			✓	
	Química	✓				✓					✓
TECNOLOGIA	Agroindústria	✓									
	Agronegócio								✓		
	Alimentos						✓		✓		
	Produção de Grãos	✓			✓	✓					
	Estética e Cosmética							✓			
	Análise e Desenvolvimento de Sistemas	✓									✓
	Automação Industrial					✓					
	Processos Químicos					✓					
	Sistemas Elétricos			✓							
	Sistemas para Internet		✓			✓		✓			
	Gestão de Turismo									✓	
	Gestão do Agronegócio				✓			✓	✓		
	Gestão Pública										✓
	Gastronomia									✓	
BACHARELADO	Agronomia				✓				✓		✓
	Medicina Veterinária		✓								
	Zootecnia	✓									
	Ciência da Computação		✓								
	Sistemas de Informação									✓	
	Arquitetura e Urbanismo						✓				
	Administração		✓		✓		✓		✓		✓
	Engenharia Agrícola	✓									

Área de Conhecimento

- Ciências Agrárias
- Ciências Biológicas
- Ciências da Saúde
- Ciências Exatas e da Terra
- Ciências Sociais Aplicadas
- Engenharias
- Outras

Campus

AL: Alegrete
FW: Frederico Westphalen
JA: Jaguari
JC: Júlio de Castilhos
PB: Panambi
SR: Santa Rosa
SAN: Santo Ângelo
SA: Santo Augusto
SB: São Borja
SVS: São Vicente do Sul

*** Cursos sem novas ofertas**