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


Tese

# Learning Analytics e Mineração de Dados Educacionais da Teoria à Prática: Aspectos Envolvidos na Implementação em Diferentes Contextos e Níveis Educacionais

**Emanuel Marques Queiroga**

Pelotas, 2022

**Emanuel Marques Queiroga**

# Learning Analytics e Mineração de Dados Educacionais da Teoria à Prática: Aspectos Envolvidos na Implementação em Diferentes Contextos e Níveis Educacionais

Tese Doutorado em Ciência da Computação– Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico

Orientador: Prof. Dr. Cristian Cechinel
Coorientador: Prof. Dr. Ricardo Matsumura de Araujo

Pelotas, 2022

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

Q3l Queiroga, Emanuel Marques

Learning analytics e mineração de dados educacionais da teoria à prática : aspectos envolvidos na implementação em diferentes contextos e níveis educacionais / Emanuel Marques Queiroga ; Cristian Cechinel, orientador ; Ricardo Matsumura Araujo, coorientador. — Pelotas, 2022.

151 f.

Tese (Doutorado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2022.

1. Learning analytics. 2. Algoritmo genético. 3. Mineração de dados educacionais. 4. Evasão e retenção de estudantes. 5. Predição de risco acadêmico. I. Cechinel, Cristian, orient. II. Araujo, Ricardo Matsumura, coorient. III. Título.

CDD : 005

Elaborada por Aline Herbstrith Batista CRB: 10/1737

**Emanuel Marques Queiroga**

# Learning Analytics e Mineração de Dados Educacionais da Teoria à Prática: Aspectos Envolvidos na Implementação em Diferentes Contextos e Níveis Educacionais

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

**Data da Defesa:** 30 de Junho de 2022

**Banca Examinadora:**

Prof. Dr. Cristian Cechinel (orientador)
Doutor em Engenharia da Informação e do Conhecimento pela Universidade de Alcalá (Espanha)

Prof. Dr. Ricardo Matsumura Araujo (co-orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul (Brasil)

Prof. Dr. Xavier Antonio Ochoa Chebab
Doutor em Engenharia pela Katholieke Universiteit Leuve (Bélgica)

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

Prof. Dr. Ulisses Brisolara Corrêa
Doutor em Ciência da Computação pela Universidade Federal de Pelotas (Brasil)

# RESUMO

QUEIROGA, Emanuel Marques. **Learning Analytics e Mineração de Dados Educacionais da Teoria à Prática: Aspectos Envolvidos na Implementação em Diferentes Contextos e Níveis Educacionais**. Orientador: Cristian Cechinel. 2022. 152 f. Tese (Doutorado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2022.

Esta tese busca apresentar as semelhanças e as diferenças da aplicação prática de Learning Analitycs (LA) e a Mineração de Dados Educacionais (EDM) em diferentes contextos e níveis educacionais. Desta forma, são apresentados os conceitos teóricos e práticos envolvidos no processo de transformação de dados educacionais em informação e conhecimento em diferentes contextos e níveis de ensino. Com esse objetivo são apresentadas as teorias envolvidas no processo, o estado da arte, as metodologias e métodos criados para o processo e o contexto das aplicações. Nesse sentido, são relatados três diferentes casos de usos com aplicações práticas desenvolvidas nesta tese: a educação secundária presencial no Uruguai, a educação universitária no Uruguai e a educação de nível médio técnico na modalidade a distância no Brasil. Nesse contexto, buscou-se gerar metodologias práticas para exploração de dados oriundos de diferentes bases e contextos educacionais, com foco na geração de modelos de alerta antecipado para evasão e retenção escolar. A primeira metodologia foi criada para identificação de estudantes do ensino secundário em risco de desvinculação em nível nacional no Uruguai. Para isso, foram coletados dados de 258.440 estudantes em nove diferentes sistemas. Esses dados foram transformados em life times temporais e depois utilizados para treinamento dos classificadores. Essa aplicação apresentou resultados interessantes e atualmente está em fase de implantação. A segunda aplicação descreve a utilização de técnicas de Data Science e EDM em dados de diversas fontes de 4.529 estudantes presenciais da Universidade da República do Uruguai. A principal contribuição dessa abordagem foi a combinação de diferentes fontes de dados, que demonstrou alto poder preditivo, atingindo taxas de predição com excelente discriminação já na quarta semana de um curso. Além disso, a análise demonstrou que os alunos com mais interações dentro do Ambiente Virtual de Aprendizagem (AVA) tendem a ter mais sucesso em suas disciplinas. Os resultados revelaram alguns atributos relevantes que influenciaram o sucesso dos alunos, como o número de disciplinas em que o aluno está matriculado e a escolaridade da mãe. Desse resultados emergiram algumas políticas institucionais, como a alocação de recursos para a infraestrutura do AVA e o desenvolvimento de ferramentas para acompanhamento dos alunos. A terceira abordagem é um algoritmo genético (AG), que busca melhorar a seleção de

hiperparâmetros de classificadores. Esse algoritmo foi proposto buscando aumentar as taxas de precisão obtidas em dados da educação técnica de cursos híbridos no Brasil. Tem-se como principal contribuição científica o desempenho da abordagem em comparação com as técnicas tradicionais, onde o AG se mostrou uma alternativa viável, produzindo resultados melhores que as técnicas tradicionais, tanto na precisão quanto no custo computacional. Dessa forma, esta tese apresenta os resultados obtidos nas aplicações, bem como as semelhanças e diferentes em 19 aspectos técnicos, como a existência de temporalidade, os tamanhos das bases de dados, a existência de múltiplas fontes de dados, a técnica de aplicação, entre outros. Ao fim, ainda são demonstradas a contribuição científica e os trabalhos futuros relacionados ao tema.

Palavras-chave: Learning Analytics. Mineração de dados educacionais. Predição de risco acadêmico. Algoritmo Genético . Evasão e retenção de estudantes.

# ABSTRACT

QUEIROGA, Emanuel Marques. **Learning Analytics and Educational Data Mining from Theory to Practice: Aspects Involved during the Implementation in Different Educational Contexts and Levels**. Advisor: Cristian Cechinel. 2022. 152 f. Thesis (Doctorate in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2022.

This thesis seeks to present the similarities and differences of the practical application of Learning Analitycs (LA) and Educational Data Mining (EDM) in different educational contexts and levels. In this way, the theoretical and practical concepts involved in the process of transforming educational data into information and knowledge in different contexts and levels of education are presented. To accomplish such objective this thesis explains the theories involved, as well as the state of art, methodologies, processes, methods and the context of the applications. Therefore, three distinct use cases and practical applications developed are presented for the following scenarios: face-to-face secondary education in Uruguay, university education in Uruguay and technical secondary distance education in Brazil. In this context, we seek to generate practical methodologies for exploring data from different educational databases and contexts, focusing on generating early warning models for school dropout. The first methodology aims to identify secondary school students at risk of dropout and failure. For this purpose, the models used data collected from 258,440 students in nine different databases. These data were transformed into temporal data and then used to train the classifiers. This application presented good results and is currently under implementation in Uruguay. The second application describes the use of Data Science and EDM techniques with data from 4,529 onsite students at the University of the Republic of Uruguay (Udelar). The main contribution of this approach was the combination of different data sources that demonstrated high predictive power, achieving predictive rates with excellent discrimination in the fourth week of a course. In addition, the analysis showed that students with more interactions within the Virtual Learning Environment (VLE) tend to be more successful in their subjects. Results also revealed some relevant attributes that influenced students' success, such as: the number of subjects in which the student is enrolled and the mother's education. From these results, some institutional policies emerged, such as allocating resources for the VLE infrastructure and developing tools for monitoring students. The third approach is a genetic algorithm (GA) that seeks to improve the selection of classifier hyperparameters. The proposal of this algorithm seeks to increase the accuracy rates of automated models to detect at-risk students enrolled in a technical distance education course in Brazil. The main scientific contribution is the performance of the approach compared

to traditional techniques, where GA proved to be a viable alternative, producing better results than conventional techniques in terms of precision and computational cost. This thesis presents the results obtained in these applications, the similarities and differences among them according 19 technical aspects (existence of temporality in the data, size of the databases, existence of multiple data sources, techniques used, among others). Lastly, the thesis establishes the scientific contribution and future work related to the topic.

Keywords: Learning Analytics. Educational Data Mining. Genetic Algorithm. Dropout and Persistence Prediction.

# LISTA DE ABREVIATURAS E SIGLAS

AA - Academic Analytics

AG - Algoritmo Genético

ANEP - Administración Nacional de Educación Pública

AVA - Ambiente Virtual de Aprendizagem

BI - Business Intelligence

BID- Banco Interamericano de Desenvolvimento

EAD - Educação a Distância

EDA - Análise Exploratória de Dados

EDM - Mineração de Dados Educacionais

ETEC - Rede de Educação Tecnológica do Brasil

FP - Falso Positivos

FN - Falso Negativo

HP - Hiperparâmetrização

BGGP - Grammar-Based Genetic Programming

LA - Learning Analytics

LSTM - Redes de Memória de Longo Prazo

MDS - Escalonamento Mutidimensional

MEC - Ministério da Educação

MOODLE - Modular Object-Oriented Dynamic Learning Environment

MOOC'S - Massive OpenOn-line Courses

MLP - Multi Layer Perceptron

ONU - Organização das Nações Unidas

PLA - Predictive Learning Analytics

T-SNE - T-Distributed Stochastic Neighbor Embedding

TIC's - Tecnologias da Informação e Comunicação Digitais na educação

RNN - Redes neurais recorrentes

SVM - Máquinas de Suporte Vetorial

SMOOTE - Sobre Amostragem de Minoria Sintética

TP - Verdadeiros Positivos

TN - Verdadeiros Negativos

VLP - Virtual Programming Laboratory Plugin

XGBOOST - Gradient Boosting Machine

UFPel - Universidade Federal de Pelotas

UDELAR - Universidad de la República del Uruguay

# SUMÁRIO

# 1 INTRODUÇÃO

A educação desempenha um papel fundamental na formação das sociedades, auxiliando na formação dos indivíduos que a compõem e na geração de conhecimento em uma ampla gama de aspectos (GREEN, 1990). Assim sendo, tem um papel transformador na vida dos seres humanos e nos projetos das sociedades modernas, permitindo que os indivíduos não só adquiram conhecimento, como também sejam capazes de desenvolver suas capacidades individuais necessárias para o convívio e sejam capazes de galgar um futuro melhor para si e suas famílias (BLOSSFELD; KIERNAN, 2019; GREEN, 1990; YOUNGMAN, 2018; DAHRENDORF, 2022).

O processo educacional é altamente mutável, evoluindo a todo instante acompanhando as tendências do mundo ao seu redor. Nesse sentido, os diferentes níveis educacionais são metodologicamente planejados para abordar e consolidar diferentes aspectos importantes aos indivíduos e as sociedades atuais (GARFIELD; AHLGREN, 1988). Como alguns exemplos, podemos citar a exploração dos sentidos básicos, a socialização e as emoções básicas na pré-escola, a alfabetização, a introdução e a consolidação dos conceitos básicos das disciplinas iniciais na educação fundamental, bem como o aprimoramento dos conhecimentos e a preparação para a vida adulta no ensino médio (CURY, 2008).

Nesse sentido, a importância da educação é reforçada pelo relatório da Organização para a Cooperação e Desenvolvimento Econômico Oecd (2019), que aponta que o acesso à educação digna, de qualidade e gratuita é um direito humano garantido na Constituição de, pelo menos, 5 a cada 10 países, inclusive no Brasil. No entanto, as preocupações com os sistemas educacionais no geral são vastas e os problemas ligados ao desempenho e a evasão ocupam significativo espaço nas pesquisas sobre a área.

Corroborando com esses aspectos, o relatório do Monitoramento Global da Educação de 2020 da UNESCO aponta que na média mundial, desconsiderando a América do Norte e os países de alta renda per capita da Europa, somente 18 indivíduos pobres concluem seus estudos em nível secundário para cada 100 jovens ricos (UNITED NATIONS EDUCATIONAL; ORGANIZATION, 2020). Esse fato se acentua conforme

o declínio da renda per capita e já é sentido em países com renda média, onde 25% dos jovens com 15 anos já não estão mais frequentando a escola e pelo menos 50% apresentam algum grau de defasagem escolar (UNITED NATIONS EDUCATIONAL; ORGANIZATION, 2020).

Essas taxas se convertem em inúmeros problemas aos países, que são potencializados no atual contexto social mundial. A literatura sobre a situação define como alguns dos principais problemas relacionados ao abandono escolar precoce as questões como a marginalização, os baixos salários, a exclusão social, a dificuldade na interação com as mudanças tecnológicas e a desindustrialização devido a falta de mão de obra qualificada para exercer determinadas funções laborais(HOSOKAWA; KATSURA, 2018; AFZAL, 2019; RAITASALO; ØSTERGAARD; ANDRADE, 2021; PARK, 2018).

Os reflexos desses problemas também são sentidos dentro dos diferentes níveis educacionais, como nas universidades, onde, de acordo com (OECD, 2019), apesar do vasto escopo do ensino superior atual, as altas taxas de evasão e retenção de estudantes são uma preocupação generalizada na área. Essas taxas são comumente associadas às ineficiências do ensino superior, embora também dependam de outros fatores, como o perfil dos estudantes, sua trajetória acadêmica e a impossibilidade de permanecer no curso por questões socioeconômicas (OECD, 2019; GRALKA, 2018; SALAZAR-FERNANDEZ et al., 2021; OCHOA et al., 2011; CECHINEL et al., 2020; BARROSO; FALCÃO, 2004).

Dessa forma, nos últimos anos, governos, organismos internacionais de desenvolvimento e entidades privadas têm buscado formas de otimizar os processos de ensino e aprendizagem em seus diferentes níveis e aspectos. Essa situação foi amplificada e beneficiada pelo crescente uso das Tecnologias da Informação e Comunicação Digitais na educação (TICs), que têm o poder de gerar a todo instante uma quantidade expressiva de dados.

Esses dados podem ser adquiridos a partir de múltiplas fontes, como os sistemas de gerenciamento acadêmico, os censos escolares, os Ambientes Virtuais de Aprendizagem (AVA) e diversos tipos de sensores. Ainda, esses dados podem ser originários de diferentes contextos educacionais, como a educação presencial, a educação híbrida, a educação a distância e os Massive Open On-line Courses (MOOCs) (SIEMENS, 2013; SIEMENS; LONG, 2011; ROMERO; VENTURA, 2020).

Nesse contexto, emergiram diversos campos de pesquisa que utilizam técnicas de Data Science, Big Data e Machine Learning, mas não restritos a elas, para a geração de informação a partir desses dados, dentre os quais podemos destacar como os principais a Learning Analytics (LA), a Academic Analytics (AA) e a Educational Data Mining (EDM).

Esses campos, apesar de estarem interligados, apresentam diversas diferenças

em seus propósitos e aplicações. No entanto, existe uma convergência em seus principais objetivos, em que podemos citar como objetivo geral a geração e o aperfeiçoamento de técnicas que possam impactar positivamente na experiência dos usuários, bem como possam fornecer feedback sobre os mesmos para as instituições (BAKER; INVENTADO, 2014).

Segundo Ochoa (2019), dentro das diversas abordagens de pesquisa que surgiram para estudar o tema, a Learning Analytics, em português Análise de Aprendizagem, apresenta-se como uma forma de compreender, analisar e propor melhorias aos processos de ensino. Com esse objetivo, ela busca explorar a ampla utilização de sistemas de apoio ao ensino baseados em tecnologias, principalmente em sistemas on-line, que geram grandes quantidades de dados. Atualmente, esses sistemas de apoio são amplamente utilizados na educação a distância e seu uso vem crescendo de forma exponencial na educação presencial.

Em um contexto regional, a América Latina se apresenta como uma das regiões de destaque em pesquisas e iniciativas que busquem utilizar a LA e a EDM como ferramentas de desenvolvimento regional (CECHINEL et al., 2020). Esse fato é auxiliado pelas semelhanças culturais, históricas e linguísticas dos 21 países que compõem a região e pelos seus problemas, como as altas taxas de concentração de renda, a baixa qualidade dos ensinos primário e secundário, os baixos índices de investimento em ciência e pesquisa e os consequentes números de pesquisadores por habitantes em suas sociedades (CECHINEL et al., 2020).

Contudo, apesar das claras contribuições que esses campos de pesquisa podem entregar, eles ainda enfrentam uma série de dificuldades, como a falta de metodologias próprias para a extração, pré-processamento e análise dos dados, e a de aplicações práticas de Learning analytics em diferentes contextos educacionais (GREGORI et al., 2018; HERNÁNDEZ-LEAL; DUQUE-MÉNDEZ; CECHINEL, 2021). Outro fator importante é que as estratégias atuais costumam trabalhar com dados de número limitado de cursos, baixos quantitativos de estudantes e são voltadas, principalmente, para educação de nível superior (HERNÁNDEZ-LEAL; DUQUE-MÉNDEZ; CECHINEL, 2021).

Dessa forma, a falta de metodologias para tratar grandes quantidades de dados de estudantes de diferentes níveis e contextos educacionais é latente. Ademais, é necessário traduzir as pesquisas desenvolvidas na área de LA e EDM em ferramentas que possam auxiliar no desenvolvimento educacional, principalmente na América Latina (CECHINEL et al., 2020; FERGUSON, 2012).

Em suma, esta tese é baseada em três diferentes metodologias de aplicação de Learning Analytics e Mineração de Dados Educacionais em diferentes contextos educacionais. Assim, busca-se demonstrar os panoramas práticos e teóricos da aplicação das técnicas baseadas em LA e EDM nos diferentes contextos educacionais.

Nesse sentido, esta tese tem como sua principal pergunta de pesquisa (PP):

- Considerando as particularidades metodológicas e as diferentes origens dos dados utilizados, quais as semelhanças e diferenças na aplicação de Learning Analytics e Mineração de Dados Educacionais em diferentes níveis e contextos educacionais?

## 1.1 Motivação

Os problemas educacionais relacionados à retenção e evasão se perpetuam em diferentes níveis da educação, causando efeitos sociais e econômicos significativos a longo prazo. Esses problemas são sentidos de forma mais intensa em populações com maiores índices de vulnerabilidade social, sendo apontados como um tipo de desigualdade educacional (WELLS; CRAIN, 1994; ZEICHNER, 2010). Ainda, é demonstrado que quanto menor a renda familiar maiores são os índices de evasão escolar e, consequentemente, menores são os índices de formação educacional (TORRACO, 2018).

A literatura na área costuma apontar a desigualdade social como um fator que se perpetua através das diferentes gerações e afeta as famílias tanto nos aspectos financeiros quanto culturais (TORRACO, 2018). Além disso, a desigualdade social é um fator desencadeador importante na contextualização dos problemas relacionados à evasão e retenção (SILVA FILHO; LIMA ARAÚJO, 2017; TORRACO, 2018). Ademais, a discussão sobre esses temas é significativamente mais ampla e envolve questões sociais, econômicas, culturais, entre outras (SILVA FILHO; LIMA ARAÚJO, 2017; TORRACO, 2018), questões que não fazem parte do escopo desta tese.

No contexto regional, a população da América Latina, que é de 680 milhões de habitantes, apresenta índices de escolaridade com média em torno de 2,5 anos abaixo da média geral dos países membros da OCDE (OECD, 2019). Isso acontece mesmo com o significativo aumento no número de matriculas na educação primária e secundária nos últimos 30 anos (OECD, 2019; DURYEA et al., 2007; AVVISATI et al., 2018).

Quando analisados, esses valores são ainda mais pessimistas. Isso se dá porque os países de outras regiões, como os asiáticos, apresentavam valores próximos aos da América Latina até os anos 1970 e hoje estão acima ou próximos à média mundial (DURYEA et al., 2007). Esse mesmo fato ocorre com as taxas de evasão e retenção, onde, apesar de os índices estarem diminuindo desde os anos 1990, os países da América Latina ainda apresentam índices alarmantes (BUSSO; BASSI; MUÑOZ, 2013; PEÑA; PÉREZ, 2013; BASSI; BUSSO; MUÑOZ, 2015).

Assim, entende-se que tecnologias que busquem aplicações práticas de Learning Analytics e Mineração de Dados Educacionais como forma de auxiliar os sistemas educacionais em diferentes estágios de formação trariam resultados, melhorando

tanto os índices educacionais pessoais quanto os familiares e, consequentemente, a renda. Dessa forma, as metodologias práticas apresentadas nesta tese buscam auxiliar na qualificação dos processos educacionais em diferentes níveis, como a educação formal presencial de nível secundário, a educação de nível superior e a educação a distância/híbrida.

Portanto, as aplicações práticas e teóricas desenvolvidas nesta tese buscam abordar os três principais modelos educacionais atuais, a Educação Primária e Secundária tradicional, a Educação Secundária Híbrida de Nível Técnico e a Educação Superior. Ao final, as três abordagens propostas se interligam no auxílio à resolução dos problemas nas três frentes de ensino, utilizando a LA e a EDM como forma de auxílio aos processos educacionais.

## 1.2 Objetivos e metas

Esta tese tem como objetivo central a geração de diferentes métodos de aplicação de Learning Analytics e Mineração de Dados Educacionais para diferentes contextos educacionais. Assim, busca-se demonstrar os panoramas práticos e teóricos da aplicação das técnicas baseadas em LA e EDM nos diferentes contextos educacionais, bem como as semelhanças e diferenças existentes nessas aplicações.

O objetivo geral se desdobra em 6 metas específicas a serem contempladas nesta tese. Essas metas são descritas abaixo:

- Meta 1 - Investigar e documentar a fundamentação teórica e o estado da arte na utilização de Machine Learning, Learning Analytics e Educational Data Mining, principalmente como ferramentas de apoio e geração de conhecimento em diferentes contextos educacionais;

- Meta 2 - Gerar uma metodologia baseada em LA que englobe a aquisição de dados, a transformação de dados, a geração de modelos e a criação de conhecimento sobre as populações estudantis, auxiliando na identificação de estudantes com risco de desligamento precoce para os principais planos de ensino médio atuais do Uruguai;

- Meta 3 - Identificar como os dados gerados pela utilização de ambientes virtuais de aprendizagem, enquanto método de apoio na educação presencial de nível superior, principalmente como repositórios de conteúdo, pode auxiliar a encontrar padrões ocultos que revelem possíveis problemas de formação (reprovação e/ou evasão). Ainda, se a combinação de dados de diferentes fontes com dados dos AVAs é uma opção para geração de conhecimento e auxilio na formação de políticas públicas baseadas em dados;

- Meta 4 - Gerar e analisar modelos de predição para desvinculação educacionais no ensino secundário, bem como se a inserção de dados de desempenho anteriores podem trazer um ganho significativo para a predição de estudantes em risco de desvinculação;
- Meta 5 - Desenvolver mecanismos que possam auxiliar no aumento das taxas de precisão dos classificadores de risco utilizados atualmente para a educação a distância;
- Meta 6 - Identificar quais padrões a mineração de dados educacionais pode ajudar a desvendar nos diferentes níveis educacionais estudados.

## 1.3 Estrutura do texto

Esta tese está estruturada no formato de compêndio de artigos, assim contendo as publicações de maior relevância realizadas pelo autor no período do doutorado. Esses artigos estão disponíveis nos apêndices A, B e C. Além disso, esta tese apresenta 5 capítulos, conforme estrutura descrita a seguir.

No capítulo 1 é apresentada a introdução e uma breve contextualização do problema e das soluções propostas nesta tese.

No capítulo 2 é apresentada a fundamentação teórica e o estado da arte sobre o tema amplamente analisado e estudado para o desenvolvimento das metodologias e teorias geradas.

O capítulo 3 apresenta os trabalhos desenvolvidos, dividido em 4 seções onde são apresentados os artigos e um comparativo entre as diferenças e semelhanças entre eles.

O capítulo 4 apresenta uma discussão dos resultados gerais das metodologias propostas, com foco principal na resposta ao objetivo geral e nas metas estabelecidas para esta tese.

O capítulo 5 apresenta de forma breve as considerações finais sobre o trabalho desenvolvido ao longo do doutorado.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 Learning Analytics

A Learning Analytics é uma área de pesquisa recente. Segundo (FERGUSON, 2012), surgiu ao longo da década inicial dos anos 2000 e começou a se estabelecer como um campo de pesquisa propriamente dito a partir de sua primeira conferência em 2011(BAKER; INVENTADO, 2014)(LEARNING ANALYTICS; KNOWLEDGE, 2011). A LA foi impulsionada não somente pelo aumento da oferta de qualificação on-line, mas também pelo incentivo político, com iniciativas como a Open University no Reino Unido e a Rede ETEC e UAB no Brasil, que tinham como objetivo levar educação de qualidade a localidades afastadas dos grandes centros universitários(QUEIROGA et al., 2016).

A definição de LA amplamente utilizada entre os autores da área é dada, de acordo com a primeira conferência internacional (LAK) realizada em 2011 (SIEMENS, 2013), por "É a medição, coleta, análise e descrição dos dados sobre os alunos e seus contextos, com o objetivo de compreender e otimizar o aprendizado e os ambientes em que ele ocorre."(LEARNING ANALYTICS; KNOWLEDGE, 2011)

A LA é considerada uma área de pesquisa multidisciplinar, a qual, segundo autores como (CHATTI et al., 2013) e (BAKER; INVENTADO, 2014), tem relação direta com campos de pesquisa como Aprendizagem de Máquina, Inteligência Artificial, Estatística, DataViz, entre outras. A LA faz uso de técnicas dessas linhas de pesquisas para desenvolver métodos que possam auxiliar na melhora da experiência no processo de aprendizagem e em todo o seu contexto.

Dessa forma, podemos afirmar que a LA busca o entendimento integral dos fatores da aprendizagem, analisando de uma forma mais humana as diferentes variáveis que podem ocasionar uma determinada situação, como um aluno concluir ou não um curso ou ter um determinado desempenho em uma avaliação. Assim, podemos afirmar que em LA há uma constante busca pela identificação de como as partes envolvidas na aprendizagem tendem a se comportar, tendo no fim um método holístico.

Em LA, o processo é um ciclo contínuo que está sempre se aperfeiçoando e não

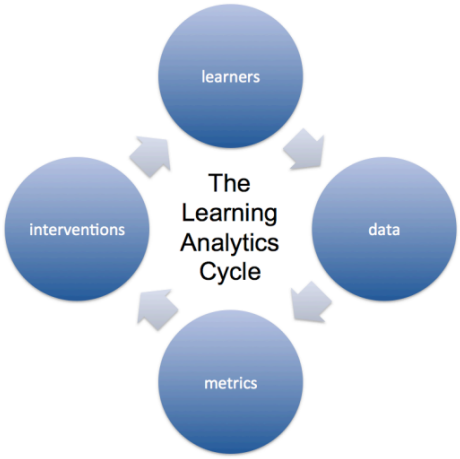


Figura 1 – Ciclo de Clow(CLOW, 2012)

tem um fim pré-determinado. Portanto, ele está em constante treinamento e reavaliação. Por esse motivo, a análise de aprendizagem mantém uma proximidade com outras áreas além da EDM, como Business Intelligence (BI), Web Semântica e os Sistemas de Recomendação(FERGUSON, 2012).

O ciclo estabelecido para o processo de learning analytics é proposto por Clow em (CLOW, 2012). Na figura 1, temos o diagrama do ciclo. Segundo (CHATTI et al., 2013), o conceito de LA está apoiado em 4 perguntas-chave: O QUE?; POR QUÊ?; COMO?; e QUEM?. Essas perguntas tendem a auxiliar na classificação das pesquisas desenvolvidas na área, propiciando o entendimento e a separação das subáreas de pesquisa (MOISSA; GASPARINI; KEMCZINSKI, 2014). Abaixo descrevemos de forma sucinta essas perguntas, entretanto, cada uma delas terá uma subseção mais adiante.

- O que? - Alusivo ao tipo de dado que será usado, bem como sua origem (exemplo: interações em um ambiente, redes sociais, dados históricos etc.);
- Quem? - Refere-se ao público-alvo da análise, podendo ser discentes, docentes, institucional, entre outros;
- Por quê? - Vinculado ao objetivo da utilização de LA, sendo alguns deles monitoramento, análise, predição, intervenção, entre outros;
- Como? - Refere-se a como é feito o processo, quais técnicas são utilizadas para isso. Por exemplo, mineração de dados, análise de redes sociais, técnicas estatísticas, entre outras.

Essas quatro perguntas são utilizadas como um modelo de referência para pesquisa e aplicação em LA e estão diretamente ligadas ao ciclo de quatro etapas proposto por Clow (CLOW, 2012)(MOISSA; GASPARINI; KEMCZINSKI, 2014). Dessa forma, podemos definir um escopo inicial para uma determinada pesquisa. Imagine o

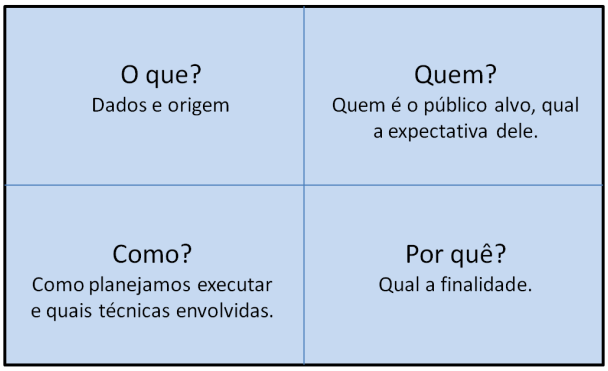

Figura 2 – Modelo de referência (CHATTI et al., 2013)

seguinte: uma pesquisa que busca utilizar as interações de acadêmicos em determinado ambiente, dados do histórico escolar, que busca monitorar e prever performance acadêmica utilizando dados de interações e histórico escolar, tendo como alvo um sistema automatizado de aviso de risco ou informe geral semanal. Como fazer isso? No escopo já estão definidas várias respostas das perguntas: O que? refere-se aos dados de interação e histórico; Quem? são os acadêmicos; Por quê? é o monitoramento e predição; mas ainda falta Como?, que seriam as técnicas que podemos utilizar para isso. A seguir, descreveremos de forma mais adequada cada uma das perguntas-chave.

O método proposto por (CHATTI et al., 2013) pode ser comparado a um planejamento estratégico, onde tem-se as etapas de definição do escopo, de detecção de qual problema se quer trabalhar, quais são os resultados esperados no processo, quais metas devem ser atendidas para se conseguir os resultados esperados e quais as técnicas que podem ser aplicadas naquele contexto.

### 2.1.1 O que?

Um aspecto relevante em LA é a avaliação dos dados que serão utilizados e qual sua origem(CHATTI et al., 2013). Os dados podem vir de diferentes bases, como ambientes virtuais de aprendizagem, sistemas de gestão institucional, sistemas acadêmicos, fontes externas, como redes sociais, e até mesmo dados oriundos de preenchimento manual (fomulários manuais, web, entre outros) (MOISSA; GASPARINI; KEMCZINSKI, 2015a).

Geralmente, essa etapa está diretamente vinculada a um processo de análise exploratória, que é feito buscando levantar os dados disponíveis, suas características, qualidades e quais as diferentes abordagens que podem ser empregadas no tratamento dos mesmos.

O processo de análise exploratória merece uma melhor análise, que será feita mais adiante nesta tese, mas de uma forma resumida podemos dizer que é o processo de levantamento e avaliação dos dados brutos disponíveis, bem como suas caracterísi-

cas, quantitativos e integridade (COX, 2017).

### 2.1.2 Quem?

Diferentes perspectivas podem ser abordadas para diferentes grupos, assim, essa dimensão da pesquisa está diretamente ligada ao seu público-alvo. Alguns dos diferentes públicos da pesquisa são estudantes, professores, tutores, orientadores educacionais, gestores educacionais ou até mesmo sistemas automatizados de avisos. Cada um deles tem expectativas diferentes quanto as informações que tendem a ser ofertadas (MOISSA; GASPARINI; KEMCZINSKI, 2014).

Um exemplo dessas expectativas é dado por Chatti (CHATTI et al., 2013), que relata que professores tendem a ter uma expectativa de métodos que possam aumentar o engajamento dos estudantes nas atividades e em formas de adaptação na abordagem de conteúdos para as necessidades específicas de determinados grupos de alunos. Enquanto isso, as expectativas dos alunos tendem a ser por métodos que possam auxiliar no incremento de suas notas.

Da mesma maneira, dificilmente grupos formados pelos stakeholders das instituições estarão interessados nos métodos anteriores. Possivelmente, esses grupos de tomada de decisão tenham expectativa em métodos sobre questões relativas aos custos, aos índices de evasão e retenção e formas de otimizar processos ou agrupamentos de estudantes (CHATTI et al., 2013) (CAMPBELL; DEBLOIS; OBLINGER, 2007).

Dessa forma, métodos que tenham uma boa aceitação em um determinado grupo dificilmente apresentarão os mesmos resultados em outros. Isso torna a escolha e a personalização da abordagem de suma importância para um projeto em LA, impactando diretamente no possível êxito da iniciativa.

### 2.1.3 Por quê?

A terceira dimensão proposta por (CHATTI et al., 2013) refere-se ao objetivo da aplicação da LA. Podemos expandir um pouco sua nomenclatura para "Por quê estamos fazendo isso? Qual a finalidade?" (MOISSA; GASPARINI; KEMCZINSKI, 2015b).

Os objetivos estão diretamente ligados aos resultados esperados. Assim, nessa etapa ocorre a discussão dos requisitos do projeto, bem como suas perspectivas e métricas de avaliação.

### 2.1.4 Como?

A quarta dimensão definida proposta por (CHATTI et al., 2013) tem como objetivo a definição das técnicas a serem empregadas para alcançar o objetivo. Apesar do modelo proposto não definir explicitamente uma sequência para a aplicação da mesma, entende-se que essa é a última parte do projeto (MOISSA; GASPARINI; KEMCZINSKI, 2014).

Para a definição da técnica empregada, precisamos ter uma leitura completa dos dados disponíveis, saber qual o público-alvo e ainda qual o objetivo e resultados pretendemos alcançar. Com um objetivo traçado, precisamos definir as técnicas a serem empregadas para obtê-lo. Entre uma variedade de métodos e técnicas para extração do conhecimento disponíveis na atualidade, podemos citar:

- Métodos Estatísticos e de Visualização de informações (DataViz);
- Academic Analytics;
- Business Intelligence (BI);
- Predictive Learning Analytics (Data Mining e Educational Data Mining);
  - Modelos de Predição e Classificação;
  - Clusterização;
  - Regras de associação;
  - Mineração de relacionamento;
- Análise de redes sociais;
- Modelagens relacionadas a avaliação e desenvolvimento de ontologias buscando o processamento de linguagem natural (conhecimento do usuário X domínio do conhecimento);
- Modelos de recomendação;
- Gamificação.

## 2.2 Mineração de Dados Educacionais

A mineração de dados educacionais (EDM) é uma crescente e emergente área de pesquisa científica, que está intimamente ligada à Análise de Aprendizagem (Learning Analytics), à Mineração de Dados e à Aprendizagem de Máquina (SIEMENS; BAKER, 2012). A EDM tem como objetivo o desenvolvimento e a adaptação de métodos que possam auxiliar na descoberta de informações em dados provenientes de múltiplas fontes e recursos educacionais. Assim, busca a compreensão dos múltiplos fatores que influenciam na aprendizagem, bem como o entendimento do comportamento do estudantes e as configurações em que eles tendem a aprender de uma forma mais eficaz (ROMERO; VENTURA, 2007) (BAKER; INVENTADO, 2014).

Com um enfoque reducionista e uma predileção pela automatização de processos, a mineração de dados educacionais analisa grandes bases de dados educacionais

com processos exploratórios que também são amplamente utilizados em Data Science (BAKER; INVENTADO, 2014), como o CRISP-DM (FERNANDES et al., 2019) e KDD (ROMERO; VENTURA, 2013).

A EDM apresenta potencial para se transformar em um método de auxílio às diversas etapas da educação. Baker e Inventado (BAKER; INVENTADO, 2014) ainda avaliam que a EDM pode ser vista de duas formas distintas, uma como comunidade de pesquisa e outra como uma área de investigação de dados científicos. Mais especificamente como área de pesquisa, busca a descoberta de conhecimento sobre as formas de aprendizagem e situações ligadas diretamente aos acadêmicos, como desempenho e evasão. Esse processo se dá a partir da análise das bases de dados geradas por diversos recursos e sistemas educacionais, como AVAs, sistemas acadêmicos, dados demográficos, entre outros (BAKER; INVENTADO, 2014).

Segundo Romero (ROMERO; VENTURA, 2010), a EDM tem como suas principais linhas de pesquisa a análise e visualização de dados, recomendações de conteúdo, predição de desempenho e evasão, modelagem e personalização de curso, detecção de comportamentos, agrupamento de estudantes e análise de redes sociais.

As pesquisas em EDM têm crescido gradativamente e apresentado destaque no contexto da tecnologia na educação. Na predição, tanto de desempenho quanto de evasão, pesquisas como (LYKOURENTZOU et al., 2009) propõem a combinação da utilização de dados acadêmicos com dados demográficos para a predição de estudantes em risco de evasão. Já (MÁRQUEZ-VERA et al., 2016) propõem, com resultados satisfatórios, os algoritmos evolutivos ICRM e ICRM2 baseados em programação genética gramatical (GBGP) na predição da evasão de alunos do ensino médio no México. Para isso, são utilizados dados têm que em torno de 60 atributos, que vão desde o teste de admissão até os dados de pesquisa distribuídos aos alunos.

No Brasil, o trabalho de (MANHÃES et al., 2011) propõe a utilização de EDM para reduzir os índices de evasão no ensino presencial em cursos que utilizem o AVA como método de apoio. Para isso, são utilizados dados das interações dos estudantes com os AVA e o desempenho nas atividades. Os alunos são classificados em três bandeiras (verde - baixo risco de evasão, amarela - risco de evasão moderado, vermelha - alto risco de evasão). Fernandes et al. (FERNANDES et al., 2019) apresentam uma metodologia que busca a predição do desempenho acadêmico de alunos de escolas públicas do Distrito Federal, utilizando dados demográficos e acadêmicos e o classificador Gradient Boosting Machine (XGBoost). Ainda são analisadas as variáveis de maior importância e seu impacto na evasão. Costa et al. (COSTA et al., 2017) avaliam a utilização de diversas técnicas de EDM na predição de desempenho em cursos presenciais e a distância, conseguindo identificar de forma precoce os alunos que tendem a evadir. Ainda, demonstram quea aplicação de técnicas de pré-processamento e hyperparametrização tendem a incrementar os resultados.

## 2.3 Learning Analytics X EDM

A LA se distingue da EDM por apresentar um foco maior no processo, em como ele se dá, e que ele seja um ciclo com fim indeterminado, composto geralmente por coleta de dados e pré-processamento, aplicação de métodos e avaliação, intervenção e aprendizagem sobre como se deu o processo. Após isso, geralmente o processo volta para o início, estando sempre em fase de aprimoramento. Além disso, a LA também utiliza outras técnicas e áreas, como a Análise de Redes Sociais e a Visualização de Dados.

Enquanto isso, a Mineração de Dados Educacionais busca utilizar os dados estudantis gerados pelos AVAs, sistemas acadêmicos, dados demográficos, entre outros, em informações. Esse processo é derivado da Mineração de Dados e Data Science e, geralmente, utiliza técnicas em comum com as mesmas.

Entretanto, duas perguntas ainda pairam sobre essas áreas de pesquisa. A primeira refere-se a aceitação das tecnologias criadas no ambiente real. Afinal, a aceitação das tecnologias propostas está diretamente ligada a seu sucesso e é um fator ainda pouco explorado. Podemos fazer uma analogia a outras áreas, como Bussiness Inteligence (BI), que já tem um longo caminho percorrido, mas ainda luta pela aceitação nas instituições. Assim, se tivéssemos uma forma de medir essa aceitação, possivelmente poderíamos criar ambientes customizados de acordo com a aceitação relativa ao tipo de usuário que temos.

A segunda pergunta é sobre como aumentar os níveis de precisão dos métodos preditivos. Atualmente, diversas abordagens buscam esse objetivo com diversas técnicas ou algoritmos novos. Afinal, o impacto da precisão é uma das chaves para a melhorar a aceitação tanto de EDM quanto de Learning Analytics. Para isso, é necessária a criação de uma metodologia ampla e que envolva o processo como um todo.

## 2.4 Predictive Learning Analytics

Predictive Learning Analytics (PLA) é uma subárea de pesquisa da Learning Analytics e da Mineração de Dados Educacionais. PLA busca a geração de modelos de predição a partir de diversos tipos de dados de estudantes, com ênfase na utilização dos Ambientes Virtuais de Aprendizagem e dados multimodais (SCLATER; PEASGOOD; MULLAN, 2016). Para isso, são utilizadas técnicas e metologias de diversas áreas de conhecimento, como Educational Data Mining, Machine Learning, Data Science e Estatística (HERODOTOU et al., 2019).

A aplicação de PLA foi rapidamente difundida na educação a distância, entretanto, no ensino presencial ainda restam diversas dúvidas quanto a sua eficiência. Algumas dessas dúvidas dizem respeito a sua aceitação por estudantes, professores e Sta-

keholders (HERODOTOU et al., 2019), seu impacto em ambientes reais e como os usuários utilizam as suas predições (HERODOTOU et al., 2017).

Nesta seção abordaremos os principais conceitos e técnicas utilizadas para modelagem preditiva em PLA, bem como os principais tipos de predição, trabalhos da área e metodologias de aceitação para implementação de tecnologias.

### 2.4.1 Compreendendo os dados

Consiste na coleta dos dados e na Análise Exploratória de Dados (EDA), bem como na busca por fontes relevantes que possam adicionar dados ao projeto. Nessa fase, os dados são coletados, os diferentes atributos são analisados e suas qualidades são medidas. A coleta de dados é uma etapa importante em qualquer projeto de análise e deve ser norteada a partir do objetivo. Assim, nela serão avaliadas as possíveis fontes de dados e os requisitos para a sua utilização (VIEIRA; PARSONS; BYRD, 2018).

#### *2.4.1.1 Tipos de dados*

Na PLA, diversos tipos de dados costumam ser utilizados, sendo os mais comuns: dados demográficos, dados de sistemas acadêmicos e dados de ambientes virtuais de aprendizagem. Assim, podemos basicamente dividir os trabalhos entre os que utilizam dados sociodemográficos, dados de interações e sistemas híbridos que utilizam os dois tipos de dados.

Dados sociodemográficos são a representação quantitativa das características de grupos de seres humanos, populações, apresentados utilizando estatística (KOVACIC, 2010). Eles são gerados a partir do conhecimento da população, geralmente são quantitativos e representam a forma de organização de uma população, tendo como objetivo buscar a análise e interpretação de diferentes tipos de dados quantitativos e qualitativos baseados em estatística.

Os dados gerados pelos ambientes virtuais de aprendizagem, geralmente utilizados em PLA, são os log's. Esses dados contém as ações efetuadas pelos estudantes com o ambiente, seus colegas, professores e conteúdos. Essas ações também são conhecidas como interações e, atualmente, existem diversas metodologias de transformação para sua utilização.

Ainda podemos destacar a utilização de dados multimodais. Esses dados são oriundos de diversos contextos e sua utilização é crescente na learning analytics (WORSLEY, 2018). Esses dados podem ser agregados às diferentes bases e revelar informações importantes sobre o processo de aprendizagem.

#### *2.4.1.2 Análise exploratória de dados - EDA*

Análise exploratória de dados é o processo de familiarização e análise dos conjuntos de dados a fim de levantar suas principais características. Nela são empregadas

diversas técnicas, principalmente visuais, para extrair informações iniciais.

Algumas técnicas empregadas são análise de distribuição e detecção de outliers. Seus objetivos principais são o levantamento de hipóteses sobre os dados e suas relações.

### 2.4.2 Integração

Com o uso de dados de diversas fontes, uma etapa importante na PLA é a integração dos dados. Ela é a combinação de processos técnicos com o objetivo de transformar os dados, estruturados ou não estruturados, em uma base única e coesa. Assim, pode-se transformar esses dados separados em informações e conhecimento para as instituições com a combinação de dados confiáveis de diversas origens em uma única.

### 2.4.3 Redução da dimensionalidade

Com a coleta de dados de múltiplas fontes, possivelmente a base de dados final será composta por uma quantidade expressiva de features. Isso gera um problema para os algoritmos de classificação, conhecido como a maldição da dimensionalidade. Essa é uma condição que ocorre quando a grande quantidade de features acaba confundindo os classificadores e, por consequência, influenciando negativamente nas previsões. Diversas técnicas são empregadas na tentativa de reduzir o problema (BACH, 2017).

Os principais métodos de redução da dimensionalidade podem ser divididos em dois, a seleção de recursos e a extração de recursos. Na seleção de recursos (Feature Selection), as variáveis de maior relevância na base são selecionadas a partir de 3 principais técnicas (BHAGOJI; CULLINA; MITTAL, 2017).

1. Filtro: seleção de variáveis baseada em recursos estatísticos, como correlação, ANOVA e Qui-quadrado.

2. Wrapper: Avaliação por subconjunto com 3 principais técnicas:

   (a) Seleção direta: as variáveis são adicionadas uma a uma e o impacto delas é medido até que o modelo deixe de ter ganho com adição;

   (b) Eliminação: as variáveis são removidas uma a uma a partir de sua significância até que o desempenho do modelo estabilize;

   (c) Eliminação recursiva: algoritmo guloso, que busca encontrar as variáveis mais importantes de acordo com uma quantidade pré-estabelecida. Assim, o algoritmo elimina recursivamente as variáveis menos importantes até que seja encontrado o número de features planejado;

3. Incorporação (embedded): Combinação dos métodos anteriores, normalmente feita dentro do próprio classificador (e.g. Random Forest e XGBOOST). Um exemplo é a seleção de atributos feita pelo Random Forest, no qual são geradas diversas árvores de baixa profundidade para subconjunto de dados, onde as features são selecionadas pela média de suas pontuações.

Na extração de recursos, a alta dimensionalidade das variáveis é transformada de forma que o conjunto final apresente uma quantidade menor de dimensões, a partir da construção de variáveis derivadas. O método mais utilizado é a Análise de Componentes Principais (PCA). Nele, as features são analisadas a partir de testes, geralmente estatísticos, e as variáveis são transformadas de forma ortogonal, onde cada conjunto de componentes apresente de forma decrescente a maior quantidade de variação possível (BHAGOJI; CULLINA; MITTAL, 2017).

Outros métodos de extração de recursos de destaque são o Escalonamento Mutidimensional (MDS) e o T-Distributed Stochastic Neighbor Embedding (T-SNE). O MDS consiste na busca pela análise da similaridade dos dados através da medida da sua distância em espaços geométricos. O T-SNE utiliza o método T-Student para a medição da afinidade entre as variáveis, assim convertendo-as em pontos de similaridades (BHAGOJI; CULLINA; MITTAL, 2017).

### 2.4.4 Transformação dos dados

Diversas técnicas buscam padronizar os dados em conjuntos que sigam uma distribuição centralizada zero. Isso se dá porque em alguns algoritmos de machine learning as medidas centrais dos conjuntos são próximas desse valor, principalmente em algoritmos que utilizem a distância euclidiana. Esse processo, também conhecido como normalização de dados, utiliza diversos algoritmos que buscam transformar os dados em distribuições entre -1 e 1, sendo 0 a média e -1 e 1 os desvios padrão. É uma técnica usada para tratar bases com uma faixa de alto valor, como é o caso dos atributos gerados pelo censo e de outliers, como os gerados na contagem de interações (OBERMEYER; EMANUEL, 2016).

Os principais algoritmos de normalização dos dados são Min-Max, Normalização pela Média (Mean Normalization) e Standardization Scaller.

### 2.4.5 Modelos de Predição baseados em Machine Learning

Machine Learning é um campo de pesquisa da inteligência artificial que busca a utilização de algoritmos para analisar dados e aprender com eles questões relativas a um determinado contexto. Possibilita que os computadores, assim como os humanos, possam ser capazes de aprender pela experiência (SAMUEL, 1959).

Uma das definições amplamente utilizada para machine learning é dada por Tom Mitchell (MITCHELL, 1997) como "Um programa de computador aprende com a expe-

riência e com relação a alguma tarefa T e alguma medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E".

Para que isso seja possível, são utilizados algoritmos, também chamados de classificadores, que buscam encontrar padrões em dados, para ao final serem capazes de fazer determinações ou previsões sobre situações sem que tenham sido explicitamente programados para aquilo. Como o processo de aprendizagem é capaz de lidar com grandes quantidades de dados, com ele é possível identificar padrões que passariam despercebidos para os olhos humanos (SAMUEL, 1959) (GÉRON, 2017).

Aprendizagem de máquina apresenta como uma de suas principais características a capacidade de resolver problemas sem que tenha sido programado especificamente para isso. Essa é uma vantagem em cenários complexos, onde a utilização de algoritmos especializados é impraticável, geralmente pela quantidade de dados envolvidos e a multiplicidade dos casos de uso.

Um exemplo amplamente utilizado é o filtro de spams dos e-mails. Com a utilização de machine learning, ele tende a se aprimorar com o passar do tempo, aprendendo padrões contidos nas mensagens, como palavras ou frases, que indicam que essas possam ser spam, tornando-se, assim, mais eficiente em bloquear mensagens indesejadas (GÉRON, 2017).

Em comparação com um programa especializado, onde possivelmente teríamos uma longa lista de regras e que a cada atualização necessitaria de um longo trabalho manual, a utilização de machine learning se torna possivelmente menos custosa, mais eficiente e de menor complexidade. É possível a utilização de diversas técnicas para treinamento constante, como aprendizagem em lotes ou on-line (BURRELL, 2016).

O processo de aprendizagem de máquina passa por diversas etapas como:

- Coleta de dados;
- Pré-processamento;
- Escolha do algoritmo;
- Ajuste de hyperparâmetros;
- Treinamento;
- Avaliação;
- Predição.

Em PLA são dois os tipos de aprendizagem aplicadas, a supervisionada e a não supervisionada. Os problemas de classificação são tarefas típicas da aprendizagem supervisionada. Nele, o conjunto de dados utilizado para treinamento é composto pela atributos da base e a variável-alvo, também chamada de rótulo. Esse alvo pode ser um dado numérico (e.g. a nota de um aluno), como a característica (e.g. a cor em um sistema de sinaleiro de risco de evasão), ou binário (e.g. a predição se o aluno irá ser aprovado ou reprovado).

Na aprendizagem não supervisionada os dados não possuem alvo, mas são agrupados de acordo com suas variáveis. Nesse processo, geralmente leva-se em conta a

similaridade entre as variáveis.

#### *2.4.5.1 Classificadores*

Dentro da aprendizagem supervisionada, a classificação é a técnica utilizada quando o problema envolve dados categóricos. O processo de classificação se dá em um conjunto de dados rotulados que serve como treinamento, ou seja, aprendizagem do problema, e o conjunto de dados não rotulados onde o algoritmo testará o que aprendeu atribuindo um rótulo às entradas. Em PLA, a classificação é o modelo mais utilizado pelas características do problema e dos dados envolvidos, geralmente porque a tarefa envolve alvos binários, como a aprovação ou reprovação e a evasão ou conclusão de um curso.

##### 2.4.5.1.1 Árvores de decisão

Árvores de decisão são algoritmos utilizados para classificação supervisionada, pois neles é necessário saber quais são as classes de cada registro do conjunto de treinamento. Esse tipo de algoritmo gera uma estrutura de árvore que classifica as amostras desconhecidas. Para isso, utiliza os dados dos conjuntos de treinamento, criando uma árvore e, a partir dessa, classificando as amostras desconhecidas sem necessariamente testar todos os valores dos seus atributos(MICHALSKI; CARBONELL; MITCHELL, 2013).

Na estrutura de uma árvore de decisão tradicional existe basicamente três tipos de nós: o nó raiz, que inicia a árvore; os nós comuns, que dividem um determinado atributo e geram ramificações; e os nós folha, que contêm as informações de classificação do algoritmo. Já as ramificações possuem todos os valores possíveis do atributo indicado no nó para facilitar a compreensão e interpretação(QUINLAN, 1986).

#### *2.4.5.2 Redes Neurais*

As redes neurais artificiais surgiram em 1943 como uma tentativa de criar um modelo matemático que imitasse o comportamento de um neurônio biológico (MCCULLOCH; PITTS, 1943). Elas podem ser definidas como técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência (CARVALHO, 2009).

As redes neurais artificiais basicamente são formadas por um conjunto de terminais de entrada, também conhecidos como camada de entrada, que repassam a informação para as camadas intermediárias, onde ocorre o processamento, e uma camada de saída, que é onde saem as informações processadas.

Essas redes podem ser compostas por várias camadas de processamento de simples funcionamento. Cada camada é conectada com a próxima e está associada a um peso. Assim, cada neurônio processa somente os dados que recebe em sua entrada

e repassa o resultado para a camada seguinte.

Assim, as Redes neurais artificias geralmente são apresentadas como sistemas de neurônios interconectados que podem computar valores de entradas (CARVALHO, 2009). O tamanho das redes neurais pode variar de acordo com a tarefa para a qual ela é utilizada, variando de uma rede de um neurônio até centenas ou milhares (CARVALHO, 2009).

Um tipo de rede neural amplamente utilizado em PLA são as redes de Perceptron Multi-Camadas (Multi Layer Perceptron - MLP). Esse modelo de rede foi criado buscando sanar alguns problemas que existiam nas redes de uma única camada (ZAFFAR et al., 2018). Na MLP, existem camadas intermediárias de neurônios e de um algoritmo de aprendizagem por retro-propagação (back-propagation). Nesse modelo de rede, todos os neurônios de uma camada estão ligados a todos os neurônios das camadas anterior e posterior. Assim, é possível um treinamento eficiente, pois cada camada tem uma função específica.

#### 2.4.5.2.1 Ensemble Learning

Métodos de ensemble são a combinação de diversos algoritmos em conjunto para o incremento dos resultados. Geralmente, esses métodos apresentam resultados de classificação superior aos métodos simples. Nesse tipo de algoritmo é considerado que um fator-chave é a independência dos classificadores individuais, pois essa tende a diminuir a probabilidade de que erros na classificação aconteçam. Os principais métodos de aprendizagem em conjunto são bagging, boosting e stacking (ensaque, reforço ou empilhamento) (GÉRON, 2017).

No método de empilhamento, também chamado de método por generalização, tem-se N preditores, onde cada um faz sua predição individualmente, combinando-se ao final. Essa combinação pode ocorrer por diversas técnicas, como a média ou a utilização de um sistema mais elaborado como a teoria dos votos.

No método por ensaque, um dos algoritmos mais utilizado e conhecido é o Random Forest. Nele, o conjunto de treinamento é dividido em diversos subconjuntos e cada um deles é treinado por um classificador independente. Esse método busca garantir um maior poder de generalização aos modelos e resultados e diminuir o viés de amostragem.

No método por reforço, a ideia principal é que a capacidade de aprendizagem individual de cada modelo seja combinada. Assim, diversos modelos são treinados e testados em sequência. Dessa forma, no primeiro modelo são identificadas as instâncias rotuladas incorretamente; no segundo modelo, após o treinamento, o classificador é testado com os erros do primeiro e assim sucessivamente até o último modelo. Alguns algoritmos de destaque nessa modalidade são o AdaBoost, Gradiente Boosting e XGBoost.

#### 2.4.5.2.2 Redes Neurais Recorrentes e LSTM

Redes neurais recorrentes (RNN) são redes persistentes projetadas para reconhecimento de padrões comportamentais. Essas redes são estruturas em loop com capacidade para a acumulação de conhecimento sobre eventos, de forma que, com as seguintes interações, possam usar do conhecimento agregado anteriormente para suas predições (OKUBO et al., 2017).

RNN são geralmente utilizadas para problemas temporais e sequenciais, em que a entrada de dados é recorrente e geralmente a saída é dependente das entradas anteriores. Um exemplo clássico é a predição da palavra que será digitada em um compilador de texto ou teclado e um smartfone, a próxima entrada é dependente das anteriores (OKUBO et al., 2017).

O principal tipo de RNN utilizada em PLA é a Redes de Memória de Longo Prazo (LSTM). LSTM são RNN de persistência de longo prazo. O diferencial delas para RNN normais é que com o passar do tempo elas ainda mantêm ativa um bom acúmulo de informações sobre séries passadas. Isso se dá em função de sua arquitetura, na qual a entrada de memórias está ligada a um barramento com as redes anteriores, diferentemente da estrutura normal de um RNN, no qual a entrada da memória está ligada somente à série anterior (OKUBO et al., 2017).

### *2.4.5.3 Métodos de Hiperparametrização*

Os hiperparâmetros são as variáveis de controle do processo de treinamento dos algoritmos de machine learning. Eles têm como objetivo definir questões pertinentes ao modelo que será treinado, como, por exemplo, o número de estimadores em um algoritmo Random Forest ou o número de camadas ocultas em uma rede neural.

Diferentemente da programação normal, onde estamos acostumados a utilizar o termo parâmetro para nos referirmos a entrada de uma determinada função, na aprendizagem de máquina os parâmetros são definidos pelo próprio modelo gerado por algoritmos. Por exemplo, podemos pegar os pesos e um nodo em uma rede neural, ele é ajustado pelo algoritmo a partir dos dados de entrada e dos hiperparâmetros e é chamado de parâmetro. Em machine learning, a precisão dos modelos está diretamente ligada a qualidade da hiperparametrização de entrada do algoritmo. Assim, quanto mais ajustados os algoritmos forem, a tendência é que as taxas de precisão dos modelos também sejam maiores.

Os principais métodos de hiperparametrização utilizados atualmente são o Grid Search, o Random Search e propostas que buscam a utilização de algoritmos genéticos. No gridsearch, uma lista de valores para cada parâmetro é definida como a entrada da busca. Assim, o algoritmo testa todas as combinações possíveis naquela lista e retorna a que obteve o melhor resultado. Já no Random Search, uma lista também é passada como entrada de forma conjunta ao número de testes, assim as

combinações dessa lista são testadas de forma randômica.

No algoritmo genético (GA), o conjunto de soluções é definido por um espaço em que ocorre a busca de uma solução ótima, que pode não ser a melhor solução global (SEBASTIANI, 2002). Esse fator depende diretamente do problema. O tempo que pode ser gasto pesquisando, o resultado esperado e o conjunto de dados de entrada, entre outros, devem ser considerados no momento em que o algoritmo é projetado (SEBASTIANI, 2002). Nesse trabalho, uma abordagem de pesquisa com tempo limitado é proposta para que o algoritmo crie um número N de gerações, onde N é pré-definido no momento da configuração. Ao final, o algoritmo retornará uma solução com a configuração que obteve o melhor desempenho de acordo com a métrica predefinida (SEBASTIANI, 2002), nesse caso, um modelo de máquina de aprendizado, juntamente com seus hiperparâmetros otimizados para a previsão de alunos em risco em cursos a distância técnicos. Como mencionado anteriormente, essa solução pode ser global ou local.

A abordagem utilizada para hiperparametrização com GA consiste na geração de quantitativo de indivíduos com código genético (hiperparâmetros) gerado de forma randômica entre uma range de valores iniciais, onde cada cromossomo é um hiperparâmetro. Ao final, o processo é repetido por N épocas, onde cada uma conta com funções de fitness, mutação e crossover.

Cada uma dessas três abordagens apresenta pontos fortes e fracos. No grid search, o principal ponto forte é a certeza da escolha da melhor solução e o ponto fraco é o tempo de processamento, que pode se tornar um problema quando se tem grandes bases de dados. No Random search, o ponto forte pode ser considerado o tempo de processamento, mas não temos a garantia de que encontramos a melhor solução. Já no GA, o tempo pode ser definido pelo número de épocas ou por um método de parada e a granularização de sua busca é um fator de que apresenta bons resultados, enquanto o problema da convergência para platos continua sendo um problema considerável.

### 2.4.6 Métricas de avaliação de desempenho

As métricas são utilizadas para analisar os resultados gerados pelos modelos de predição. Elas ainda podem ser utilizadas para verificar se os resultados apresentados são satisfatórios e onde os modelos podem ser melhorados. A escolha da métrica deve levar em conta diversos fatores, como o balanceamento do conjunto de dados.

As principais métricas utilizadas em PLA consistem na análise de classificações binárias. Abaixo, são apresentadas as técnicas mais usuais em PLA:

1. Verdadeiros Positivos (TP) - quantidade de itens que foram classificados como positivos e realmente pertenciam a essa classe;

2. Verdadeiros Negativos (TN) - quantidade de itens que foram classificados como negativos e realmente pertenciam a essa classe;

3. Falso Positivo (FP) - quantidade de itens que foram classificados como positivos e não pertenciam a essa classe;

4. Falso Negativo (FN) - quantidade de itens que foram classificados como negativos e não pertenciam a essa classe;

5. Acurácia - a acurácia é considerada a métrica mais comum, o que se deve em grande parte por ser a métrica mais simples de ser calculada. Ela é dada por (TP + TN) / (TP + TN + FP + FN), assim ela demonstra a razão entre o número de previsões corretas e o número total de amostras de entrada. Ela é uma métrica utilizada para medir o desempenho geral do modelo e pode ser utilizada em casos onde o dataset seja balanceado;

6. Precisão - razão entre TP e FP, dada pela fórmula TP/(TP + FP). A precisão é a medida da capacidade de um classificador apontar corretamente uma instância positiva como tal;

7. AUROC - também chamada de AUC, AUROC é calculada a partir do tamanho da área sob a curva plotada em que o eixo Y é representado por Taxa Positiva Verdadeira (TPR) ou Sensibilidade (TP/(TP+FN)) e o eixo X é Taxa Negativa Verdadeira (TNR) ou Especificidade (TN/(TN+FP)). De acordo com (GAŠEVIĆ et al., 2016), a AUC pode ser interpretada da seguinte forma:

   (a) AUC $\leq$ 0,50: má discriminação;

   (b) 0,50 $\geq$ AUC $\geq$ 0,70: discriminação aceitável;

   (c) 0,70 $\geq$ AUC $\leq$ 0,90: discriminação excelente;

   (d) AUC $\geq$ 0,90: discriminação notável.

8. Recall - também chamado de revocação, recall é a medida em que o classificador aponta TP corretamente. Ele pode ser definido por TP/TP+FN;

9. F1 Score - a média ponderada entre a precisão e o recall, dada pela fórmula F1 = 2 * (precision * recall)/(precision + recall). Utilizada para análise de generalização de modelos em dados desbalanceados, F1 pode ser considerada a métrica de distinção entre as classes.

## 2.5 Estado da arte

Esta seção apresenta trabalhos focados na predição de alunos em risco em diferentes cenários e no uso de técnicas de hiperparâmetros para melhorar os resultados.

Vários trabalhos na área de análise de aprendizagem e mineração de dados educacionais lidam com o problema da previsão precoce de alunos em risco. Os trabalhos costumam se diferenciar em vários aspectos, como: (1) as fontes de dados utilizadas para gerar os modelos de previsão (demográficos, AVAs, pesquisas, exames); (2) o nível de escolaridade dos cursos (ensino médio); (3) o objetivo dos modelos preditivos (por exemplo, prever desempenho ou evasão); (4) o escopo da previsão focado em um programa inteiro ou um curso ou disciplina específica; e (5) a modalidade do curso (formal ou informal, presencial, semipresencial ou a distância).

De acordo com Liz-domínguez et al. (2019), análise de dados é o conjunto de técnicas utilizadas para transformar dados em informação e conhecimento, revelando correlações e padrões ocultos. Os dados resultantes desse processo podem ser usados para criar sistemas de alerta antecipado para prever eventos futuros. Esse processo tem como objetivo principal apoiar a aprendizagem e mitigar alguns dos problemas, como desempenho acadêmico, retenção e evasão. A confiabilidade das predições pelo preditor é um dos principais fatores estabelecidos por Liz-domínguez et al. (2019) e Herodotou et al. (2017) para sua aplicação em larga escala.

Com o crescimento do interesse em Predictive Learning Analytics, diversas pesquisas buscam modelar dados e extrair informações e conhecimento para as instituições. Essas pesquisas basicamente se dividem entre predição de desempenho e evasão. Entretanto, segundo (ZOHAIR, 2019), ambos os tipos de predição estão atrelados, pois o desempenho é um fator relevante na retenção dos estudantes. Com esse objetivo são utilizados dados de diversas fontes, como dados acadêmicos (ZOHAIR, 2019), dos ambientes de aprendizagem (MÁRQUEZ-VERA et al., 2016), (FERNANDES et al., 2019), dados demográficos (LYKOURENTZOU et al., 2009), dados sobre gastos e renda (DAUD et al., 2017) e dados multimodais provenientes de diversas fontes.

Nesse contexto, as pesquisas tradicionais costumam utilizar dados dos sistemas acadêmicos e dos ambientes virtuais. Pesquisas, como a de (ZOHAIR, 2019), propõem a utilização de somente dados provenientes do sistema acadêmico na predição de performance em estudantes de pós-graduação. São extraídos dados como cursos extracurriculares realizados e as respectivas notas, curso de formação inicial e dados descritivos sobre as notas e a idade do estudante. Nesse estudo foi demonstrado que para conjuntos pequenos de alunos essa é uma abordagem válida, que pode apresentar bons resultados com poucas etapas de pré-processamento e um conjunto limitado de dados. O referido autor foca na utilização de algoritmos que apresentam bons resultados com baixas quantidades de dados, como Máquinas de Suporte Vetorial (SVM) e Redes Neurais de Múltiplos Perceptrons (MLP), obtendo bons resultados com ambos.

Uma pesquisa tradicional da área e que merece destaque é a da (LYKOURENT-

ZOU et al., 2009), na qual é proposto um método próprio de ensemble para predição de alunos em situação de risco de evasão. Esse método combina os resultados da aplicação de 3 algoritmos diferentes, sendo eles: MLP, SVM e Fuzzy ARTMAP (PES-FAM). São utilizados dados demográficos do curso, como sexo e residência; além de dados acadêmicos, como performance e nível escolar; e dados variantes, como número de interações com o ambiente virtual, notas e até mesmo a data da entrega dos trabalhos. Com a aplicação dos algoritmos são criados 3 esquemas diferentes buscando a predição da evasão, onde no primeiro um estudante é considerado evadido se pelo menos uma técnica o classificou como tal, no segundo o estudante é considerado evadido se pelo menos duas técnicas indicam essa situação e no terceiro e último necessita-se que as 3 técnicas classifiquem o aluno como evadido para que ele seja assim classificado. Os resultados obtidos variam de 73% a 94%, sendo que os mais satisfatórios foram obtidos pelo esquema 1, que chegou a atingir 94%.

A busca por métodos que possam ser generalizáveis, portanto, replicáveis a outros cursos, também apresenta uma significativa parcela da área de pesquisa. Assim, pesquisas como a de (WHITEHILL et al., 2017) propõem uma arquitetura não dependente de dados únicos, trabalhando com o fluxo de cliques que os acadêmicos efetuam em um MOOC. Para isso, ele captura dados de um curso e busca treinar modelos diferentes de predição e testar em outros cursos e ambientes. Nos experimentos são demonstradas taxas entre 87% testando em cursos diferentes e 90% se testados no mesmo curso, assim, não variando significativamente de acordo com o ambiente.

Os algoritmos genéticos também são amplamente usados na mineração de dados e, por consequência, na PLA, podendo ser implementados como o próprio classificador ou como otimizador dos recursos envolvidos na predição (MINAEI; PUNCH, 2003). (MINAEI-BIDGOLI; PUNCH, 2003) apresentam uma abordagem para classificação de alunos tentando prever a nota final em uma determinada disciplina utilizando uma combinação de algoritmos, como árvores de decisão, redes neurais e regressão lógica linear de forma concorrente, utilizando algoritmos genéticos para otimizar os resultados. Assim, conforme os autores demonstram, é possível obter resultados mais satisfatórios em relação às técnicas normais que eles chama de classificação bruta.

Em (MÁRQUEZ-VERA et al., 2016) é apresentada uma proposta de AG para otimização de hiperparâmetros em uma variante do algoritmo Grammar-Based Genetic Programming (GBGP), com o objetivo de melhorar a classificação de acadêmicos em risco de evasão. Essa técnica é aplicada sobre o algoritmo ICRM, proposto por (CANO; ROMERO; VENTURA, 2013), sendo ajustados os parâmetros do classificador até que se chegue a um método que apresente maior aptidão. Os experimentos utilizaram dados de cursos de rápida duração (4 – 6 semanas). Em comparação com os classificadores usuais, o algoritmo proposto apresenta maior taxa de predição, estabelecendo-se como uma alternativa a cursos que disponham das características

utilizadas na proposta.

Na proposta de (XING et al., 2015) é apresentada a utilização de AG para as etapas de seleção de variáveis e classificação de alunos quanto ao desempenho em uma disciplina de um curso. Os autores sugerem uma abordagem que quantifica as atividades dos alunos no MOOC em 6 variáveis predefinidas, com o objetivo de diminuir a dimensionalidade dos dados. Como classificador também é implementado o algoritmo ICRM proposto por (CANO; ROMERO; VENTURA, 2013). Assim, no estudo foi possível obter resultados superiores em até 6% se comparados às técnicas tradicionais, como Nave Bayes, Random Forest, MLP, entre outros, tanto na etapa de predição da situação final do aluno quanto na interpretação dos modelos gerados.

A utilização de dados multimodais, como a frequência cardíaca, a contagem de passos, as condições climáticas e atividades de aprendizado, é proposta por (DI MITRI et al., 2017). Esses tipos de dados podem ser recolhidos com a utilização de dispositivos biosensores, como pulseiras do estilo smart band e relógios smartwatches. Nessa proposta, é desenvolvido um sistema responsável pelo recolhimento e pré-processamento dos dados. É proposto que os dados multimodais sejam divididos em categorias, que tendem a medir o nível de stress, produtividade, desafio e habilidade nas atividades desenvolvidas. Após isso, ocorre a integração com dados acadêmicos dos estudantes, como as atividades anteriores e dados derivados das mesmas. São treinados modelos lineares de predição com o objetivo de medir a eficácia da técnica, demonstrando que na comparação com a utilização somente de dados acadêmicos a abordagem apresenta um ganho significativo.

A utilização de dados multimodais, como a movimentação do mouse e a frequência de utilização do teclado, é proposta por (WEI et al., 2020). Nesse sentido, os dados são extraídos através de um plugin de questionários no ambiente virtual, onde são postadas perguntas e o aluno tem que arrastar a resposta certa. Essas informações são pré-processadas e divididas pelo tempo que o aluno demorou até o primeiro clique, a quantidade de cliques, o tempo de movimentação do mouse, o tempo no arrastar a resposta, a trajetória do mouse e seu tempo e o último clique. Esses dados são acrescidos de dados acadêmicos e, posteriormente, são aplicados classificadores como redes neurais e SVM buscando a predição da performance escolar. São demonstrados resultados promissores na comparação com a utilização de dados apenas dos sistemas acadêmicos.

O estudo de (HERODOTOU et al., 2019) busca identificar diferenças no desempenho de turmas onde os professores utilizaram o sistema PLA. Esse sistema utilizado é disponibilizado aos professores por uma universidade que oferece cursos de graduação on-line. Nele são apresentadas informações sobre a evolução do aluno e a predição de performance e risco de evasão. A análise desenvolvida nesse trabalho demonstrou que as turmas onde os professores utilizaram o PLA tiveram uma per-

formance pelo menos 15% maior que as turmas sem o uso. Essa melhora também foi observada na comparação com turmas dos mesmos professores, mas de anos anteriores.

(DAUD et al., 2017) propõem uma metodologia de análise que utiliza dados de renda, como gastos familiares e informações dos gastos pessoais dos alunos, acrescidos de dados demográficos e de desempenho acadêmico prévio na predição de retenção. Para isso, são analisados diversos tipos de gastos, sendo eles divididos em 4 classes: despesas familiares, renda familiar, informações pessoais dos alunos e bens familiares. São aplicados os métodos de classificação SVM, BayesNet e diferentes modelos de árvores de decisão. São medidos os ganhos das diferentes variáveis que compõem o estudo, com a média do gasto com energia elétrica apresentando ganho significativo. Na predição, os resultados apresentam uma melhora com a adição dos dados propostos em comparação com a utilização de somente dados acadêmicos.

No experimento de (JAYAPRAKASH et al., 2014), busca-se criar um sistema de alerta de risco quanto ao desempenho do aluno, a fim de diminuir as taxas de evasão e retenção escolares, fornecendo ao aluno um feedback atualizado de seu possível rendimento escolar. Para isso, os autores utilizam dados demográficos, como sexo e idade, interações dos alunos com o ambiente virtual de aprendizagem, desempenho acadêmico anterior, tempo na universidade, tempo on-line no ambiente virtual, dados do teste de aptidão escolar (SAT Verbal e Matemático), entre outros. Assim, são analisados os dados de 9.938 alunos e aplicando árvores de decisão com o algoritmo J48, redes Bayesianas com o Naive Bayes, Máquinas de suporte Vetorial com o SVM/SMO e regressão logística. Na tarefa de predição, todos algoritmos apresentaram resultados muito próximos, tendo o classificador de regressão logística apresentado resultados ligeiramente maiores que os outros 3, com 94,20% de acurácia geral e 66,70% de precisão na predição de alunos em risco de evasão.

Abordagens diferentes e que utilizam somente a contagem de interações com o ambiente virtual também são propostas por (DETONI; CECHINEL; MATSUMURA ARAÚJO, 2015), (QUEIROGA; CECHINEL; ARAÚJO, 2017), (QUEIROGA; CECHINEL; ARAÚJO, 2015) e (QUEIROGA et al., 2016). O trabalho de (DETONI; CECHINEL; MATSUMURA ARAÚJO, 2015) busca utilizar unicamente a contagem de interações para predição de reprovação de alunos em disciplinas de EAD. Para isso, são utilizados dois cursos oferecidos pela UFPel: Licenciatura em Educação do Campo (CLEC) e Licenciatura em Pedagogia (CLPD). Em sua pesquisa, o autor opta por extrair as interações dos alunos, tutores e professores e agrupá-las de forma semanal. A partir das interações ainda são calculadas a média, mediana, média da diferença (média da diferença entre a semana i e a semana i+1), razão com professores (razão entre o total de interações do aluno e dos professores), razão com tutores (razão entre o total de interações do aluno e dos tutores) e fator de empenho (razão

entre as interações da semana do aluno e a média de interações da turma naquela semana). Após isso, o autor aplica os algoritmos Redes Bayesianas, Redes Neurais, J48 e Random Forest, obtendo resultados de até 67% de acurácia na predição do desempenho do aluno.

A abordagem proposta por (QUEIROGA; CECHINEL; ARAÚJO, 2017) busca a geração de modelos de acompanhamento do risco de evasão que utilizam somente dados das interações com o ambiente virtual. As interações são classificadas conforme o dia e semana, a média de interações, mediana e desvio padrão. Assim, ao final do pré-processamento são geradas diversas variáveis e modelos de acompanhamento semanal para o aluno. São aplicados classificadores, como Redes Neurais, BayesNet e árvores de decisão (J48 e Random Forest). Os resultados demonstram que a partir da quarta semana de curso essa abordagem já apresenta taxas de acerto consideradas excelentes. A característica mais relevante dessa proposta é a facilidade de implantação, pois depende de dados externos e um plugin desenvolvido para o moodle poderia ser implementado de forma simples.

No trabalho de (MACARINI et al., 2019) é proposta uma comparação entre diversas técnicas de pré-processamento de dados de interações com o ambiente virtual Moodle na predição de risco. O autor também faz uso de dados do Plugin Virtual Programming Laboratory (VPL), buscando a predição de risco em disciplinas de algoritmo e programação em cursos de graduação. São gerados dados como a contagem de interações semanal, a média das interações, mediana, quantitativo de semanas sem interações, desvio padrão e fator de compromisso, baseados na técnica proposta anteriormente por (DETONI; CECHINEL; MATSUMURA ARAÚJO, 2015) e (QUEIROGA; CECHINEL; ARAÚJO, 2015). Além disso, são acrescidos dados sobre a contagem de interações dos professores, contagem social e contagem cognitiva, baseadas na teoria proposta por Swan (SWAN, 2003). Com dados naturalmente desbalanceados, é aplicada a técnica de sobre-amostragem de minoria sintética (SMOTE) para seu balanceamento. São gerados diversos datasets com variáveis diferentes ou a totalidade com o objetivo de comparar as técnicas. Os resultados obtidos demonstram que a utilização somente da contagem de interações, como proposta por (DETONI; CECHINEL; MATSUMURA ARAÚJO, 2015) e (QUEIROGA et al., 2016), apresentou resultados superiores às demais técnicas, inclusive a união delas.

Modelos de predição baseados em Deep Learning vem gradativamente conquistando espaço na pesquisa em PLA (DING et al., 2019), (KIM, 2019) e (HASSAN et al., 2019). Hassan (HASSAN et al., 2019) buscam comparar os resultados obtidos pelos métodos clássicos de deep learning utilizando LSTM. Para isso, são combinados dados temporais e acadêmicos do dataset da Open University, com o objetivo de predizer precocemente o risco de evasão. O modelo proposto na comparação com MLP e Regressão Logística alcança resultados de precisão significativamente maiores.

O método proposto por (KIM, 2019) busca a predição em tempo real do desempenho de estudantes de acordo com seu comportamento em ambientes virtuais de aprendizagem. São utilizados dados reais de alunos de graduação na Udacity. É proposto um algoritmo chamado GritNet, que utiliza diversas redes do tipo LSTM de forma conjunta e concorrente para gerar predições por agrupamento dos estudantes. O modelo proposto na comparação do autor apresenta resultados superiores aos modelos clássicos, como SVM, Regressão Logística e árvores de decisão. Entretanto, o modelo apresenta um custo de processamento significativamente maior.

A utilização de RNN do LSTM é proposta por (OKUBO et al., 2017). Nela, os autores utilizam dados do ambiente virtual de aprendizagem de uma universidade para predição de estudantes em risco. São coletados dados de 108 alunos e divididos de forma temporal e por atividade. Nos resultados demonstrados pelos autores, as redes do tipo LSTM obtêm resultados significativamente superiores a modelos de regressão na tarefa de predição precoce.

## 2.6 Implicações econômicas relacionadas a evasão e retenção

Mensurar o impacto gerado por problemas educacionais como a evasão, a retenção e a qualidade do ensino é o tema principal de diversas pesquisas. Nesse sentido, uma parte significativa dessas pesquisas apontam que o avanço científico, tecnológico e econômico de um país está diretamente correlacionado à qualidade da educação oferecida e os níveis de sucesso obtidos pelos estudantes nos diferentes níveis de formação(ASIF; HAYAT; KHAN, 2021; LATIF; CHOUDHARY; HAMMAYUN, 2015; YAKUNINA; BYCHKOV, 2015; ELISTIA; SYAHZUNI, 2018).

As implicações e os efeitos dos índices de evasão e retenção escolar são sentidos na economia de diversos países, sobretudo naqueles em desenvolvimento(LATIF; CHOUDHARY; HAMMAYUN, 2015). Assim, entende-se que o investimento em educação e os diferentes índices baseados na formação de estudantes nos diferentes níveis educacionais são demonstrativos do futuro pretendido por diferentes nações(ALJOHANI, 2016; LATIF; CHOUDHARY; HAMMAYUN, 2015; WETZEL; O'TOOLE; PETERSON, 1999).

Esses fatores são amplificados quando falamos dos anos iniciais de ensino, onde se dá o processo de formação inicial do estudante e ele adquire conhecimentos e percepções que irá carregar ao longo de toda sua formação e vida(FALL; ROBERTS, 2012; HOSOKAWA; KATSURA, 2018).

Ademais, existem diversos custos envolvidos para a formação de um estudante. Se pegarmos o Brasil como um exemplo, o Estado investe atualmente entre R$ 296,00 e R$ 420,00 mensalmente na formação de um aluno de ensino básico e médio. Com uma média anual de investimento para formação de um estudante em R$

3.349.00(GALVÃO, 2021; FERREIRA et al., 2020; FERNANDES; BASSI, 2021). Ainda nesse contexto, diversas fontes apontam que no sistema universitário o investimento anual por aluno é de aproximadamente R$ 27.850,00(GALVÃO, 2021; SANTOS; CARVALHO PEREIRA, 2019). Certamente, são valores consideráveis e, de certa forma, perdidos quando temos um estudante em estágio de evasão ou retenção.

Assim, a cada dia que passa aumenta a necessidade de encontrarmos formas de combater a evasão e a retenção estudantil para garantirmos a prosperidade dos sistemas como um todo(HAGEDORN, 2005). Nesse sentido, explorar as oportunidades criadas pelos aspectos do crescimento da utilização de TICs na educação passa automaticamente pelos processos de Learning Analytics e Mineração de Dados Educacionais. Assim, propomos nesta tese três diferentes métodos que podem ser utilizados em diferentes estágios de formação para auxiliar no combate aos problemas aqui citados.

Salienta-se que os métodos propostos atualmente se encontram em diferentes estágios de implementação/utilização, com a metodologia para extração de conhecimento e geração de modelos de predição para o ensino médio em estágio de implantação, a metodologia para trabalhos com dados do ensino superior já auxiliando na formação de políticas institucionais e com o algoritmo genético para seleção de hiperparâmetros ainda em um estágio aperfeiçoamento e desenvolvimento.

# 3 CONCEPÇÃO DO TRABALHO

Este capítulo apresenta o panorama das diferentes aplicações práticas de Learning Analytics e Mineração de Dados Educacionais em diferentes contextos e níveis educacionais gerados nesta tese.

Dessa forma, este capítulo apresenta uma seção com uma breve contextualização dos ambientes de aplicação, bem como 3 outras seções, em que cada uma apresenta uma breve introdução e contextualização da aplicação prática que está relatada no artigo publicado referente ao tema e uma seção final com a comparação das técnicas e seus contextos de aplicação.

Os artigos presentes nesta tese e citados nas seções deste capítulo encontram-se nos apêndices A, B e C.

## 3.1 Sistema de identificação e alerta precoce de estudantes em risco de evasão e reprovação na educação secundária no Uruguai

A educação secundária tem um papel significativo na formação de jovens, auxiliando a consolidar e aprofundar os conhecimentos adquiridos durante a infância e o ensino primário (SILVA, 2012). Nessa etapa de formação, são estimulados o desenvolvimento do pensamento crítico, criativo e independente (AIZIKOVITSH-UDI; CHENG et al., 2015), bem como são trabalhados aspectos relevantes para a continuação da formação acadêmica e a iniciação ao mercado de trabalho (MARCHBANKS III et al., 2015).

No entanto, as taxas de evasão e retenção são alarmantes nessa etapa de ensino e causam diversos problemas sociais e econômicos, principalmente em países em desenvolvimento (MARCHBANKS III et al., 2015; LEE; CHOI, 2011). No contexto dos países em desenvolvimento, estudantes que necessitam repetir algum ano de sua formação tem entre 3 e 7 vezes mais chances de evasão escolar, sendo esse um fator significativo no abandono escolar (FINE; DAVIS, 2003).

Nesse contexto, o apêndice A apresenta o artigo, em fase de publicação, intitulado

"Early prediction of at-risk students at secondary education: a countrywide K-12 learning analytics initiative in Uruguay", resultante do trabalho desenvolvido em parceria com a Agência Nacional de Administração da Educação Pública do Uruguai (ANEP).

No trabalho desenvolvido no artigo, buscamos coletar e analisar dados de todos os estudantes em nível primário e secundário do Uruguai para gerar uma metodologia de coleta, processamento, análise de dados e geração de modelos de predição antecipados para o sistema público de ensino secundário uruguaio. Assim, o artigo descreve uma iniciativa nacional de análise de aprendizagem focada na predição precoce de estudantes em risco evasão e retenção, bem como em fornecer um feedback baseado em dados para a implementação de futuras políticas governamentais para mitigar o problema no ensino médio.

Com essa finalidade, foram recolhidos dados de 258.440 estudantes em diversas fontes de dados oficiais da ANEP. Esses dados apresentam diversas informações dos estudantes durante sua trajetória escolar no primário e secundário no período entre 2015 e 2020. Algumas das principais informações presentes nos dados são a trajetória do estudantes desde a primeira série do ensino primário até a segunda série do ensino secundário (avaliações dos alunos nas diferentes disciplinas ao longo dos anos), quantidade de faltas, participação em programas de assistência social, zona da escola e algumas informações sociodemográficas.

Ainda, no artigo presente no apêndice A foram levantadas 4 perguntas de pesquisa do experimento 1 (PP - E1) principais que nortearam o desenvolvimento da pesquisa, da implementação, da avaliação dos resultados e da discussão dos resultados.

- **PP1 - E1** - É possível gerar uma metodologia baseada em LA que englobe aquisição de dados, transformação de dados e geração de modelos que possam ajudar a identificar precocemente alunos em risco de evasão no ensino secundário?
- **PP2 - E1** - A transformação de dados de diferentes bases de dados em séries temporais é uma alternativa viável do ponto de vista do pré-processamento? Em caso afirmativo, os resultados finais gerados pelos modelos de previsão usando essa técnica são satisfatórios?
- **PP3 - E1** - É possível gerar e analisar modelos explicáveis baseados em aprendizado de máquina para que os vieses possam ser identificados e corrigidos quando necessário?
- **PP4 - E1** - Quais são as características mais importantes para prever precocemente alunos em risco de evasão no Uruguai no nível secundário?

Com os objetivos gerais e as PP - E1 definidas e após uma etapa de análise exploratória dos dados, foram montados diversos scripts para processamento dos dados,

que tinham como principal finalidade criar um lifetime sobre o estudante. Para isso, são processados os dados dos diferentes sistemas, dando origem entre 130 e 170 diferentes variáveis que foram geradas e testadas. Na sequência, é executada uma etapa de seleção dos atributos com maior contribuição na tarefa de identificação precoce dos estudantes em risco de desvinculação.

Ainda, de acordo com o estabelecido na **PP3 - E1**, as diferentes variáveis geradas foram analisadas para que fosse possível a identificação de possíveis vieses (bias), assim identificando possíveis grupos protegidos. Essa etapa resultou em 3 grupos protegidos, que são o gênero, a zona escolar e a participação em programas de assistência social. Esses grupos foram analisados quanto a suas distribuições e o comportamento de seus membros, e na etapa de geração dos modelos os mesmos foram testados para que possíveis vieses fossem identificados e corrigidos quando necessário.

Foram planejados 8 diferentes modelos com diferentes entradas de dados para serem aplicados em diferentes períodos do ano letivo para o ensino secundário regular e para o ensino secundário técnico. Esses modelos devem ser aplicados antes do início do ano letivo e após a primeira reunião de avaliação de cada duas séries (4 modelos por tipo de ensino X 2 modelos de ensino). Assim, os modelos preditivos foram desenvolvidos considerando essa abordagem temporal e, após uma análise de viés considerando os atributos protegidos, 7 deles foram aprovados para serem usados para predição.

Esses modelos gerados obtiveram desempenhos destacados de acordo com a escala utilizada por Gašević et al. (2016), tendo obtido valores de AUROC superior a 0.90 e F1-Macro superior a 0.88.

Abordando brevemente as hipóteses de pesquisa e respondendo brevemente a **PP1 - E1**, podemos afirmar que, apesar de ser uma tarefa árdua, é possível a geração de metodologia baseada em LA que englobe aquisição de dados, transformação de dados e geração de modelos que possam ajudar a identificar precocemente alunos em risco de evasão no ensino secundário. No entanto, com a metodologia em fase de implantação pela ANEP, surgiram diversos questionamentos e possibilidades de melhoria, como a falta de aquisição automática dos dados. Essas situações tendem a ser corrigidas com a continuidade do projeto e com as etapas futuras de retreinamento anual dos modelos, onde deverão ser implementadas novas features ao processo.

Os modelos gerados e que estão em fase de implantação utilizam a metodologia estabelecida para processamento dos dados, onde são formados lifetimes para os estudantes. Assim, foi possível obter os resultados citados anteriormente. Dessa forma, podemos afirmar que a resposta para a **PP2 - E1** é positiva e que a geração de modelos de predição utilizando essa abordagem de levantamento de dados temporais é viável.

Ainda, os modelos gerados podem ser analisados e transformados em modelos explicáveis. Assim, a resposta para a **PP3 - E1** também é positiva e, como citado anteriormente, entre os 8 modelos gerados só um apresentou algum tipo de viés dentro do grupo de variáveis protegidas.

Como citado na literatura, as variáveis que representam o desempenho dos estudantes em etapas de ensino anteriores têm alto poder preditivo, tanto nas tarefas de predição da evasão quanto para a predição da retenção de estudantes. Assim, como uma resposta breve a **PP4 - E1**, o trabalho demonstra a importância das variáveis ligadas ao desempenho do estudante em etapas anteriores da sua formação. Como um exemplo, podemos pegar o modelo M2G1-CES, onde as variáveis com maior poder preditivo são a zona escolar do estudante no primeiro ano de ensino primário e a variável que representa o agrupamento do estudante pelas notas obtidas no sexto ano do ensino primário.

## 3.2 Identificação precoce de estudantes em risco de reprovação no ensino universitário

O processo de atualização contínua é umas das bases da educação, principalmente nas universidades. Nos últimos anos, esse processo está passando por uma significativa mudança com a adoção em massa de sistemas de gerenciamento e de apoio ao conhecimento.

Esses sistemas têm como uma de suas características a capacidade de recolher uma infinidade de dados sobre as populações estudantis. No entanto, o processo de transformação desses dados em conhecimento e informações ainda necessita ser lapidado.

O artigo "Using Virtual Learning Environment Data for the Development of Institutional Educational Policies", disponível no apêndice B, demonstra o processo de transformação de dados de diversas fontes em modelos de predição e análise de dados sobre as populações educacionais na Universidad de la República del Uruguay (UDELAR). Descreve como técnicas de EDM e Data Science podem auxiliar na identificação de padrões de comportamento em estudantes de ensino superior.

Com esse objetivo, foram utilizados dados de 4.529 estudantes de quatro diferentes cursos de graduação. Esses cursos acontecem de forma presencial, com os ambientes virtuais de aprendizagem auxiliando como repositório de conteúdo e espaço para discussão. Assim, foram recolhidos dados de diferentes tipos e fontes, como os ambientes virtuais de aprendizagem, os sistemas de gerenciamento da UDELAR e os censos universitários feitos pela própria instituição.

Após isso, foram definidas três perguntas de pesquisa principais para esta abordagem.

- **PP1 - E2:** O uso do AVA está associado à aprovação do aluno?
- **PP2 - E2:** Quais recursos dos diferentes conjuntos de dados (AVA, censo e sistema acadêmico) são os mais importantes para a previsão antecipada do desempenho dos alunos?
- **PP3 - E2:** Quais padrões de aprendizagem a mineração de dados educacionais pode ajudar a desvendar nos cursos estudados?

Assim, este trabalho buscou a utilização dos processos de EDM e Data Science como ferramentas para desvendar o conhecimento educacional e possíveis padrões existentes relacionados à situação final dos alunos. Assim, no artigo presente no apêndice B, reportarmos resultados quantitativos sobre modelos preditivos, bem como os padrões nos dados e como eles podem auxiliar para entender melhor o papel que os AVAs e outras variáveis têm no desempenho dos estudantes.

Após a coleta, os dados passaram por um processamento buscando a criação de uma base única que pudesse servir tanto de treinamento para a geração de modelos de predição de risco quanto para uma profunda análise exploratória dos dados. Essas análises geradas no trabalho disponível no apêndice B serviram de embasamento para a confirmação e geração de políticas educacionais voltadas para a permanência dos estudantes na universidade.

Após isso, foram gerados e comparados diversos modelos de precisão para diferentes semanas dos 4 cursos. Nesses modelos, conseguimos alcançar discriminações excepcionais já na quarta semana de curso, distinguindo os estudantes com maior probabilidade de reprovação com uma AUROC > 0.90.

Ao final, essa abordagem apresentou bons resultados no geral, inclusive demonstrando que o engajamento na utilização dos ambientes virtuais de aprendizagem está ligado a aprovação dos alunos e auxiliando na formação de políticas institucionais baseadas em evidências. Logicamente, essa abordagem ainda apresenta algumas limitações, como a falta de integração automatizada dos dados e a falta de expansão para os demais cursos da universidade. Essas limitações podem ser mitigadas em um futuro, já que a metodologia pode ser facilmente adaptada.

## 3.3 Identificação precoce de estudantes em risco de evasão no ensino técnico a distância

No Brasil, diversas cidades encontram-se afastadas dos grandes centros universitários e acabam ficando isoladas de programas de graduação e cursos técnicos profissionalizantes. Dessa forma, uma das alternativas adotadas pelo governo federal para a expansão do acesso à educação foi a utilização da modalidade à distância

(Educação a Distância - EAD), que tem como um de seus objetivos levar o ensino a essas localidades, geralmente utilizando Ambientes Virtuais de Aprendizagem (AVAs) (DELANO; CORRÊA, 2013; QUEIROGA et al., 2016).

O AVA é o “local virtual” onde os cursos na modalidade a distância, ou semipresenciais, normalmente acontecem. São ambientes que utilizam plataformas especialmente planejadas para abrigar cursos. Uma das plataformas mais utilizadas no país é o Modular Object-Oriented Dynamic Learning Environment (Moodle[1]).

No Moodle existem diversas áreas para apresentação de conteúdos em diversos formatos, atividades de verificação da aprendizagem e espaços para interação síncrona, por meio de chats, e assíncrona, através de fóruns de discussão. Trata-se de recursos que permitem a interação dos estudantes entre si e com a equipe de tutores e professores. A organização do ambiente virtual permite ao aluno um acompanhamento organizado e sistematizado daquilo que é estudado a cada semana. A recuperação da informação e dos conteúdos estudados também é um dos benefícios proporcionados por cursos a distância que utilizam AVAs (SEGUNDO; RAMOS, 2005).

Um dos principais desafios da EAD é obter a diminuição do índice de evasão, que, conforme o Censo EAD (CENSO, 2018), foi de 18,6% em 2010, 20,5% em 2011, 11,74% em 2012 e 16,94% em 2013 nos cursos autorizados pelo Ministério da Educação (MEC). Num contexto onde, em 2013, havia 5.754 cursos autorizados pelo MEC e a taxa de matrículas anual foi de 882.843, temos em torno de 149.553 alunos evadidos.

Atualmente, o processo de detecção de estudantes em risco de evasão é complexo e envolve diretamente os professores (MANHÃES et al., 2011). Esse processo é ainda mais complexo na educação a distância, onde geralmente um professor tem uma quantidade significativamente maior de alunos.

Nesse contexto, a aplicação da EDM pode possibilitar o tratamento diferenciado entre os alunos, dedicando formas de auxílio diferenciadas a um determinado aluno que esteja com uma probabilidade maior de evasão. No entanto, a aceitação nos modelos de predição está atrelada a diversos fatores, entre eles podemos citar o entendimento das predições geradas pelos modelos e as taxas de acertos.

Assim, o artigo presente no apêndice C, intitulado "A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course", dá seguimento aos estudos realizados durante o mestrado em Computação defendido com êxito no Programa de Pós-Graduação em Computação da Universidade Federal de Pelotas. Objetivou métodos de hiperparametrização baseados em um algoritmo genético proposto na pesquisa, que demonstrou uma capacidade de melhora dos resultados gerais da predição. Assim, pode contribuir para métodos mais exitosos e confiáveis na predição de estudantes em risco de evasão, principalmente

[1] https://Moodle.org/

na EAD.

Os modelos de curso utilizados para este estudo são os cursos na modalidade híbrida, com aulas on-line e ocasionalmente encontros presenciais, oferecidos pelo Instituto Federal Sul-riograndense (IFSul), campus Visconde de Graça (CaVG). Esses cursos são ministrados em 18 polos espalhados pelo interior do estado do Rio Grande do Sul e funcionam com atividades semanais, que são postadas no ambiente pelo professor, com os alunos tendo uma semana para o desenvolvimento dessas com o auxílio dos tutores.

Cada curso tem um tempo de realização máximo de 103 semanas, com carga horária total de 1.215 horas divididas nas disciplinas do curso dentro do período de 24 meses, contando com 3 intervalos também chamados de férias, sendo que a situação final do aluno é determinada pelo seu resultado nas avaliações. Nesse modelo de curso, as disciplinas são oferecidas sempre de forma sequencial, com o estudante cursando somente uma disciplina por vez.

O trabalho desenvolvido anteriormente propõe modelos que possam ser de fácil generalização e que acredita-se que possam ser aplicados em outros cursos do IFSul ou até mesmo em outras instituições de ensino que utilizem o modelo da Rede e-TEC. Optou-se por utilizar as contagens diárias e semanais de interações dos alunos com o ambiente virtual.

Com esses objetivos, foram utilizados os mesmos conjuntos de dados de testes anteriores e publicados por (QUEIROGA; CECHINEL; ARAÚJO, 2015; QUEIROGA et al., 2016; QUEIROGA; CECHINEL; ARAÚJO, 2017). Esse conjunto de dados contém informações sobre as interações de 2.503 estudantes de 4 diferentes cursos oferecidos pelo IFSul. O esquema no banco de dados para cada um dos cursos é baseado no id do aluno, 103 campos com as interações semanais dos alunos, 721 com o dia, 103 com a média, 103 com a mediana, 103 com o desvio padrão semanal, o total de interações e a situação final do aluno no curso.

Com o conjunto de dados pronto, são aplicados os seguintes algoritmos de classificação em suas configurações básicas: Decision Tree (DT), Random Forest (RF), MultiLayer Perceptron (MLP), Logistic Regression (LG) e o meta algoritmo AdaBoost (ADA). Além disso, é aplicado o método de seleção de hiperparâmetros Grid Search. Os resultados obtidos por eses algoritmos são salvos para futura comparação com os resultados obtidos pelo GA.

O algoritmo genético proposto para a otimização do classificador (hiperparâmetro) e seleção do mais apto objetivava prever a evasão nos cursos. Nessa abordagem, diversos indivíduos concorrem entre si na busca por aquele que apresenta as taxas de acurácia, nesse caso em AUROC. Ao final, o classificador e os hiperparâmetros com os melhores resultados são selecionados por uma função de aptidão. Isso se dá por um número limitado de gerações, que é definido no momento de início da execução do

algoritmo, bem como o número de indivíduos por geração. Ao final, o algoritmo retorna uma solução com a configuração que produziu o melhor desempenho de acordo com a métrica predefinida.

O AG proposto foi capaz de atingir valores acima de 10% nos experimentos até a 20ª semana de curso em relação aos algoritmos em sua configuração padrão. Quando comparado ao outro método de otimização, Gridsearch, nesse mesmo período, o GA obteve valores sempre acima de 6%, chegando, às vezes, a 15% em determinados momentos. Esses resultados obtidos com essa solução baseada no GA foram significativamente melhores que os obtidos pelas soluções tradicionais e ainda ficam próximos aos da literatura na área.

Dessa forma, entende-se que a solução proposta, em combinação com as soluções desenvolvidas nos trabalhos anteriores, pode contribuir para o desenvolvimento da área de pesquisa, principalmente aumentando os resultados obtidos pelos classificadores utilizados para predição de risco.

Entretanto, essa solução ainda pode ser aperfeiçoada, como com a colocação de métodos de stop mais efetivos e a combinação com outros métodos de seleção de hiperparâmetros, como o Grid Search e o Random Search, ficando essas sugestões para o desenvolvimento futuro desta ferramenta.

## 3.4 Comparação entre as aplicações

Os diferentes experimentos realizados nesta tese buscaram a geração de metodologias para a aplicação prática de LA e EDM em diferentes níveis e contextos educacionais. Assim, nesta seção, buscamos apresentar o panorama de cada um dos experimentos, bem como os desafios práticos encontrados no decorrer dos projetos, os resultados práticos e a comparação entre as aplicações.

Como esperado, cada uma das aplicações práticas apresentou desafios consideráveis na implementação da metodologia desenvolvida. Grande parte desses desafios diz respeito ao acesso direto às fontes de dados. Nesses casos, geralmente o detentor das fontes de dados entregou um dump da base e os diagramas da bases ou um arquivo .csv com as colunas solicitadas.

A tabela 1 demonstra uma comparação entre as 7 dimensões principais definidas nesta tese. Essas dimensões são: Contexto, Abrangência, Objetivo, Metodologia, Técnicas, Modelos preditivos e Resultados Obtidos e Implementação e Interessados.

Essas dimensões são apresentadas nas subseções e se dividem em 19 aspectos primordiais das aplicações. Esses aspectos são: Contexto, Abrangência, Objetivo, Dados Disponíveis, Abundância de dados, Complexidade na integração dos dados, Técnica de Balanceamento, Tamanho das bases em quantidade de estudantes, Técnica de Modelagem, Complexidade da Modelagem, Temporalidade, Método de gera-

Tabela 1 – Comparativo entre as aplicações

| Aplicação | Contexto | Abrangência | Objetivo | Metodologia | Técnicas | Modelos preditivos e Resultados Obtidos | Implementação e Interessados |
|---|---|---|---|---|---|---|---|
| Ensino Secundário | Planos de ensino UTU e CBT do ensino secundário público do Uruguai. | Ensino Secundário | Sistema de alerta antecipado para predição de risco de evasão ou retenção | Aplicação de técnicas de learning analytics em dados de 258.440 estudantes recolhidos de nove diferentes sistemas acadêmicos. | Learning Analytics | Utilização doal. Random Forest com resultados com discriminação excelente | Implementação prática para a Agência Nacional de educacção pública do Uruguai (ANEP). |
| Ensino Universitário | Quatro cursos presenciais de ensino universitário da UDELAR - Uruguai | Ensino Universitário | Avaliação das bases de dados disponíveis e geração de modelos de predição para identificação precoce de estudantes em risco de evasão | Aplicação de EDM em dados das interações dos estudantes com o AVA, dados oriundos dos sistemas de gerenciamento acadêmico e censo escolar. Totalizando 4.529 estudantes de 3 cursos diferentes. | EDM | Utilização do alg. Random Forest com resultados com discriminação excelente | Implementação prática voltada para os professores, coordenadores de cursos e administração. |
| Ensino a Distância | Cursos de nível técnico subsequente oferecidos na modalidade à distância. | Ensino a Distância | Predição de evasão precoce | Aplicação de EDM nas interações dos estudantes com o AVA em 4 diferentes cursos na modalidade EAD. Totalizando 2.503 estudantes. | EDM | Algoritmo genético proposto pelo autor, com base nos alg. MLP, Random Forest, Nave Bayes, Logistic Regression, AdaBoost. Resultados com discriminação excelente | Implementação prática ainda não executada. No entanto os interessados são os coordenadores de tutoria, os analistas de sistemas e a administração. |

ção, Algoritmos de Classificação Utilizados, Redução de Dimensionalidade, Técnica de Hiperparâmetrização, Resultados Obtidos, Interessados e Implementação Prática.

Esses aspectos foram pensados para elucidar as diferenças e as semelhanças entre as aplicações. Assim, as seções abaixo e suas subseções utilizarão esses aspectos-base para as discussões.

### 3.4.1 Contextos de Aplicação

Para o desenvolvimento deste projeto, foram utilizados dados de três contextos e níveis educacionais diferentes em dois países da América Latina, Brasil e Uruguai. Cada um deles foi escolhido pelas particularidades envolvidas no processo do ensino daquele determinado contexto educacional. Além disso, outros fatores foram considerados nessas escolhas, como a facilidade de acesso aos dados, as possibilidades de geração de aplicações teóricas e práticas e o acesso ao financiamento para a pesquisa.

O primeiro contexto educacional é a educação de nível secundário no Uruguai. O projeto desenvolvido nesse contexto contou com o financiamento do Banco Interamericano de Desenvolvimento (BID) e da colaboração de Agência Nacional de Administração da Educação Pública do Uruguai (ANEP) e da Universidade da República do Uruguai (UDELAR).

O objetivo geral do projeto desenvolvido nesse contexto era a geração de uma metodologia para aquisição de dados de múltiplas fontes institucionais, pré-processamento, análise exploratória de dados e aplicação de machine learning para identificação precoce de estudantes com possíveis problemas acadêmicos (evasão e/ou retenção). Assim, foi gerada uma API para recolhimento, pré-processamento e alerta precoce para estudantes em risco de evasão ou reprovação.

Com esse objetivo, foi gerada uma metodologia para aquisição de dados das trajetórias educacionais dos estudantes desde o ensino primário. Essa metodologia cria

um lifetime do estudante, demonstrando suas avaliações, faltas, recebimento de benefícios sociais e variáveis derivadas, desde o primeiro ano de estudo do aluno.

Para a realização dessa aplicação, foram recolhidos dados de 258.440 estudantes oriundos de nove diferentes fontes de dados institucionais. Esses estudantes estão regularmente matriculados nos dois principais planos de ensino secundário do Uruguai, CES e UTU.

Ao final, os dados e os modelos de predição passam por diversas avaliações, como a análise de possíveis vieses, buscando garantir os resultados e uma maior equidade. Ainda, é gerada uma API para predição e avaliação dos resultados, bem como diversos treinamentos e manuais para a implementação e constante retreinamento dos modelos.

O segundo contexto educacional é a educação de nível universitário no Uruguai. A pesquisa referente a esse contexto foi possibilitada por meio de uma colaboração com a UDELAR. Nessa colaboração, foi possível realizar um estágio de pesquisa de em torno de 45 dias em Montevidéu no Uruguai. Assim, foi possível conhecer as instalações de pesquisa e o gerenciamento da Comissão Setorial de Educação (CSE-UDELAR), bem como outras faculdades da UDELAR e as particularidades de seu modelo de ensino.

Essa pesquisa foi financiada por meio do projeto de pesquisa denominado "10 años en EVA: siguiendo las huellas de los estudiantes en el Entorno Virtual de Aprendizaje" , que é financiado por recursos competitivos do edital de “Proyectos de Investigación para la Mejora de la Calidad de la Enseñanza Universitaria” (PIMCEU) na Universidade da República do Uruguai.

No estágio de pesquisa foi possibilitado o acesso aos dados de diferentes fontes educacionais de três diferentes faculdades da UDELAR. Como principal fonte de dados foram utilizados os ambientes virtuais de aprendizagem, que nesses cursos, alvos da pesquisa, são utilizados como repositório de conteúdos e ferramenta de apoio ao ensino presencial.

Nessa aplicação, foram utilizados dados de 4.529 estudantes oriundos de fontes institucionais, como o censo acadêmico anual da UDELAR, que contém informações sociodemográficas, e os sistemas de gerenciamento acadêmico, que contêm as notas dos estudantes. Podemos destacar a utilização de 1.129,392 interações com o ambiente virtual de aprendizagem, realizadas em 14 diferentes disciplinas alvo desse estudo.

O terceiro contexto educacional são quatro cursos técnicos na modalidade a distância (EAD) do Instituto Federal Sul-riograndense (IFSul) campus Visconde de Graça (CaVG). Esses cursos são ministrados em 18 polos espalhados pelo interior do estado do Rio Grande do Sul e funcionam com atividades semanais, que são postadas no ambiente pelo professor, com os alunos tendo uma semana para o desenvolvimento

dessas com o auxílio dos tutores.

Cada curso tem um tempo de realização máximo de 103 semanas, com carga horária total de 1.215 horas divididas nas disciplinas do curso dentro do período de 24 meses, contando com 3 intervalos também chamados de férias, sendo que a situação final do aluno é determinada pelo seu resultado nas avaliações.

O prazo máximo para a integralização do curso é de 4 anos, podendo o aluno repetir somente uma vez cada disciplina e, por consequência, o ano. Ele ainda tem a opção de levar até 2 disciplinas como dependência para o próximo ano e cursá-las de forma concomitante às outras disciplinas do curso. Para a aprovação, o aluno deverá ter média igual ou superior a seis em cada uma das disciplinas da matriz curricular. Considera-se evadido o aluno que passe um período de 365 dias sem interações com o ambiente virtual ou não efetue sua rematrícula anual, sendo desligado do curso.

Assim, o aluno pode assumir 2 estados diferentes no final das atividades, aprovado ou reprovado. No entanto, o estudo tem como objetivo a predição dos alunos que entrem em situação de evasão no decorrer do curso. Para tal, define-se que o aluno será considerado evadido caso abandone, não efetue as atividades no decorrer do curso e, também, sua matrícula no semestre seguinte.

Nessa abordagem foram coletados dados sobre as interações dos estudantes com os ambientes virtuais de aprendizagem. Sendo utilizados os mesmos conjuntos de dados anteriormente criados em outras pesquisas (QUEIROGA; CECHINEL; ARAÚJO, 2015; QUEIROGA et al., 2016; QUEIROGA; CECHINEL; ARAÚJO, 2017).

A metodologia utilizada para o trabalho com esses dados foi a contagem de interações. Ela consistiu na coleta e pré-processamento de 3.700.916 logs de interações dos alunos. Após isso, esses logs permitem calcular as interações por dia e semana de curso de acordo com as 103 semanas do calendário do mesmo. Ao final, são extraídas variáveis derivadas dessas contagens, como a média nas últimas 4 semanas, a média semanal, a média mensal, desvio padrão sobre a média da turma e os quartis onde o estudante se encaixa a partir da contagem de interações.

Nesse contexto, maximizar os resultados obtidos pelos diferentes classificadores baseados em aprendizagem de máquina é um desafio considerável. Isso se dá porque os diferentes algoritmos comumente apresentam uma grande variação nas taxas de desempenho que dependem da combinação de várias características (por exemplo, equilíbrio entre classes, quantidade de dados, variáveis de entrada e outros) e hiperparâmetros do algoritmo.

Dessa forma, propomos um algoritmo genético baseado em computação evolutiva que gera populações concorrentes, com cada indivíduo sendo um classificador com hiperparâmetros diferentes. Esses indivíduos concorrem entre si por um número limitado de eras até que, ao final, ocorre a seleção do mais apto.

Dessa forma, esses são os três contextos e níveis educacionais que buscamos

aplicar LA e EDM nesta tese, buscando avaliar os diferentes impactos que podem ser gerados pela sua aplicação e comparar as diferenças e semelhanças em cada um deles.

### 3.4.2 Abrangência de Aplicação

As três aplicações desenvolvidas nesta tese se diferenciam pela abrangência geral. O fator abrangência foi previsto para avaliar qual o nível de abrangência geográfica onde a aplicação foi testada e não necessariamente o nível de abrangência que ela pode escalar. Nesse sentido, as propostas se diferenciam em abrangência regional, estadual e nacional.

A proposta voltada para a educação básica teve uma abrangência nacional, utilizando dados de estudantes em todas as escolas pertencentes ao sistema UTU e CES no Uruguai. Assim, a metodologia criada foi aplicada em nível nacional.

A proposta voltada para a educação universitária foi classificada com abrangência regional. Isso se dá pelo fato de, apesar de a UDELAR oferecer cursos em graduação em nível nacional, os dados utilizados no desenvolvimento são de faculdades em Montevidéu.

Já a proposta voltada para a educação de nível técnico na modalidade a distância foi classificada como estadual, pois os dados utilizados são de 19 polos espalhados pelo interior do Rio Grande do Sul.

Assim, cada uma das metodologias criadas e aplicadas no decorrer desta tese possuem abrangências geográficas diferentes.

### 3.4.3 Objetivo

As aplicações desenvolvidas nesta tese contam com objetivos próprios, alinhados com as necessidades dos contextos e níveis de ensino em que estão inseridas. Nesse sentindo, de certa forma, elas refletem diversas características dos ambientes de aplicação.

A aplicação voltada para o ensino secundário no Uruguai tem como objetivo principal a criação de um sistema de identificação antecipado para estudantes em risco acadêmico (evasão e retenção), através da exploração de diferentes bases de dados educacionais, gerando, ao final, uma metodologia para transformação dos dados de diferentes sistemas em dados que possam servir para alimentar o sistema.

A segunda aplicação tem como objetivo principal avaliar como os rastros deixados pelos estudantes nos ambientes virtuais de aprendizagem podem auxiliar na identificação de padrões na educação universitária presencial, principalmente nos que dizem respeito a retenção de estudantes.

A terceira aplicação tem como objetivo principal gerar sistemas de identificação precoce de estudantes em risco de evasão, com foco principal na melhora dos resulta-

dos obtidos anteriormente através da exploração de técnicas de hiperparâmetrização.

Assim, podemos definir que a aplicação voltada para o ensino secundário é voltada ao processo, a aquisição dos dados, ao processamento e a geração dos modelos como um ciclo, muito próximo ao modelo definido por Chatti et al. (2013). Isso se difere das duas outras aplicações geradas nesta tese, que buscam um foco maior no processo de mineração de dados, tanto na análise exploratória quanto no incremento dos resultados.

### 3.4.4 Metodologia

Nesta subseção, concentraremos os aspectos referentes aos dados disponíveis, a abundância de dados, ao tamanho dos dados, ao método de trabalho utilizado, a complexidade de pré-processamento e integração, a modelagem, as técnicas de balanceamento utilizadas e a presença de temporalidade nos dados. Juntos, esses 9 aspectos representam a base principal da metodologia desenvolvida em cada uma das aplicações e serão tratados de forma conjunta nesta subseção.

O método de trabalho está diretamente ligado aos objetivos e resultados esperados naquele contexto. Assim, cada um dos trabalhos desenvolvidos nesta tese necessitou de uma metodologia própria, adaptada ao seu contexto e objetivos.

Na metodologia criada para a aplicação de LA em dados do ensino secundário no Uruguai, o contexto principal foi trabalhar com dados dos dois sistemas de ensino secundários majoritários no país. Esses sistemas correspondem conjuntamente em torno de 95% das matrículas de estudantes nessa etapa de ensino.

O trabalho para essa aplicação consistiu em uma breve adaptação do método CRISP-DM (Cross Industry Standard Process for Data Mining). No CRISP-DM, adaptado para o desenvolvimento da metologia criada no contexto da educação secundária, temos quatro diferentes estágios de treinamento. Em cada um desses estágios são adquiridos novos dados sobre os estudantes e os modelos referentes a esse estágio são retreinados.

Ao final da linha de processamento dos dados, adicionamos uma etapa de conclusão ao CRISP-DM, onde buscamos adquirir informações sobre a população estudantil, são avaliados possíveis correções no processo, estudadas formas de melhoria no processo como um todo, análise do feedback repassado pelo usuários do sistema de predição e discussões sobre a otimização do processo. Nela, os dados gerados pelos modelos podem ser analisados e agrupados, tendo, por exemplo, os percentuais de estudantes em risco por escola, região ou até mesmo turma.

A metodologia desenvolvida no contexto da educação universitária na UDELAR - Uruguai é baseada em uma aplicação de Mineração de dados educacionais em dados oriundos de diferentes fontes. Nesse contexto, a metodologia desenvolvida utiliza uma adaptação do método CRISP-DM muito próxima a do trabalho com dados

da educação secundária. Nessa adaptação, na última etapa da implementação, as informações que emergiram durante o processo são encaminhadas para a geração de conhecimento das equipes responsáveis pela formulação e adequação de políticas internas da universidade.

A terceira metodologia foi criada pensando em situações onde inexistem ou não estão acessíveis dados de múltiplas fontes. A aplicação principal pensada para essa metodologia são os cursos oferecidos na modalidade EAD e os MOOCs. Nesses cursos, geralmente, não existe uma grande integração ou recolhimento de dados externos, como, por exemplo, dados sociodemográficos, de trajetórias anteriores do estudantes, entre outros.

Sendo uma metodologia voltada para aplicação em dados dos ambientes virtuais de aprendizagem no contexto da educação a distância em que não é possível a agregação de dados de outras fontes, buscamos utilizar somente dados disponíveis nos AVAs. Assim, trabalhamos com variáveis derivadas da contagem de interações dos estudantes dentro do curso.

Dessa forma, o processo utilizado para essa aplicação consistiu na Descoberta de Conhecimento em Bases de Dados, do inglês para Knowledge Discovery in Databases (KDD). O KDD é um processo de extração de informações em grandes quantidades de dados, buscando encontrar informações até então desconhecidas, que possam ser úteis na identificação de padrões (QUEIROGA, 2017; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Nos trabalhos anteriores desenvolvidos pelo autor desta tese, essa metodologia baseada na contagem de interações apresentou bons resultados, mas que ainda careciam de uma melhora nos índices de predição obtidos nas primeiras semanas. Dessa forma, optamos por técnicas de hiperparâmetrização dos algoritmos classificadores (QUEIROGA et al., 2016; QUEIROGA, 2017).

Nesse sentido, as técnicas tradicionais baseadas em buscas exaustivas apresentam excelentes resultados, tendo a certeza de que encontraram os melhores HPs dentro do conjunto passado para teste. No entanto, essas técnicas apresentam alto custo computacional e tempo de processamento. Isso se agrava quando buscamos utilizar diferentes algoritmos, pois cada um necessita de uma busca exaustiva dentro dos HPs e um processo manual de seleção daqueles com melhores índices para cada algoritmo e o melhor geral.

Esses fatores se agravam em situações onde não dispomos de hardware adequado para essa tarefa, como em parte dos servidores onde os ambientes virtuais de aprendizagem estão instalados. O moodle define como configuração mínima para sua instalação um equipamento com processador de 1 ghz e 512mb de memória ram. Essa configuração mínima não é capaz de rodar os sistemas de busca exaustivas atuais de forma eficiente.

Dessa forma, buscamos a criação de um algoritmo que pudesse selecionar HPs de forma não exaustiva, de forma mais otimizada computacionalmente, que testasse diferentes classificadores e que obtivesse bons resultados sem comprometer o funcionamento do moodle. Assim, criamos um algoritmo genético baseado na teoria da evolução darwinista e na concorrência populacional.

Esse algoritmo criado e aplicado gera populações onde os indivíduos são instâncias de um classificador geradas com HPs criados randomicamente na primeira geração. Após essa primeira geração, é executada uma etapa de ranqueamento dos mais aptos de acordo com um parâmetro passado na inicialização do algoritmo. Após isso, é gerada uma nova população a partir de funções de seleção dos mais aptos, cruzamento, mutação e reinserção. Esse processo se dá por um número limitado de épocas, que é definido na inicialização do algoritmo. Ao final, o algoritmo apresenta uma listagem com os indivíduos mais aptos. Essa listagem apresenta o classificador, a época em que ele foi gerado, seus HPs e os índices obtido.

Assim, em cada projeto apresentamos uma metodologia diferente, baseada nos dados existentes, no contexto e nível educacional e nos objetivos previamente estabelecidos. A motivação para a utilização de diferentes métodos é que o processo de Learning Analytics não tem um final estabelecido e sim ciclos, o que se adapta melhor à estrutura do método CRISP-DM. Enquanto isso, os projetos voltados para a mineração de dados e mineração de dados educacionais têm um fim definido, assim eles são melhor adaptados ao método KDD, pois nesse contexto o projeto é finalizado no momento em que você entrega as informações aos interessados.

Cada uma das aplicações desenvolvidas nesta tese contou com dados oriundos de diferentes fontes. Na aplicação voltada para o ensino secundário foram recolhidos dados de 9 diferentes fontes. Além disso, foram avaliados outros sistemas que não apresentaram dados capazes de acrescentar informações relevantes à base final. Nesse metodologia, as fontes de dados foram agrupadas em três principais conjuntos, dados da educação primária, dados da educação secundária CES e dados da educação secundária UTU.

Após o agrupamento, os dados foram pré-processados e integrados por scripts gerados em python. Essas etapas foram de alta complexidade, tendo em vista o tamanho das bases com 258.440 estudantes, a quantidade de diferentes bases e a abundância de dados e fontes. Uma característica interessante nessa integração e que dificultou o processo foi que o id do estudante mudava em algumas bases, com bases salvando o id do estudante do sistema e outras o número do documento pessoal, sendo necessária a utilização de uma base externa para a validação.

Nessa metodologia, foi possível a criação de um lifetime do estudante, onde são utilizados os dados das avaliações desde o ensino primário até o momento da geração do modelo de predição. Assim, foi possível montar um aspecto de temporalidade na

base final utilizada para os modelos de predição.

Na metodologia para aplicação de mineração de dados educacionais no contexto universitário foram utilizados dados dos AVAs, do censo acadêmico e de diferentes sistemas de gerenciamento acadêmicos da universidade. A partir do censo acadêmico foram extraídos dados sociodemográficos dos estudantes, como estado civil, residência, renda familiar, número de filhos, escolaridade dos membros da família, entre outros.

Nos sistemas acadêmicos foram extraídas informações, como as notas dos estudantes, o número de disciplinas cursadas e a quantidade de disciplinas em que o estudante está matriculado naquele semestre.

Já nos ambientes virtuais de aprendizagem foram extraídas informações sobre as interações dos estudantes com os conteúdos disponibilizados no mesmo. Assim, foi possível criar um lifetime das interações do estudante no AVA. Esse lifetime demonstrou os principais momentos onde os estudantes buscam conteúdo e auxílio nos AVAs. Além disso, essas interações dos estudantes, quando combinadas com as informações de outras bases, demonstraram um alto poder preditivo e de melhora nos índices de predição.

No contexto dessa aplicação são inexistentes dados históricos que possam ser classificados como temporais. Assim, fica impossibilitada a criação de um lifetime para demonstrar o históricos do estudante antes da disciplina. Isso se dá, em partes, pela metodologia de ingresso nos cursos da UDELAR, onde o estudante pode somente efetuar a matrícula em uma disciplina dentro de um curso sem que seja aluno daquele curso, caso não existam pré-requisitos de disciplinas anteriores.

Ainda nesse contexto, inexiste uma integração com dados de níveis anteriores de ensino, como o primário e o secundário, e com bases de dados governamentais.

No contexto da educação a distância de nível técnico inexistem dados de outros fontes que não o moodle. Assim, foi possível trabalhar apenas com as contagens de interações. Essas interações estão presentes nos logs de acessos dos estudantes no ambiente virtual. Um modelo de log do moodle é apresentado na tabela 2.

O conjunto de dados utilizado nesse aplicação é o mesmo utilizado anteriormente em diversos estudos (QUEIROGA; CECHINEL; ARAÚJO, 2015; QUEIROGA et al., 2016; QUEIROGA; CECHINEL; ARAÚJO, 2017; QUEIROGA, 2017). Assim, são utilizadas 3.700.916 linhas de logs de interações dos de estudantes com o AVA de quatro diferentes cursos, conforme apresentado na Tabela 3.

Quanto ao balanceamento, a base da educação técnica apresenta um balanceamento natural entre a quantidade de estudantes concluintes e a quantidade de evadidos. Esse mesmo processo não ocorreu nas bases da educação universitária e educação secundária. Assim, nessas bases foi necessária a aplicação de técnicas de balanceamento, como o SMOOTH, que gera dados artificias para balancear os

Tabela 2 – Modelo de log do Moodle - Fonte: Queiroga (2017)

| Dado | Descrição | Exemplo |
|---|---|---|
| Curso | Turma e curso ao qual o aluno esta matriculado. | Informática Aplicada - 2013/2 Administração |
| Hora | Data e hora que foi efetuada a ação que gerou este determinado log. | 2012 abril 8 10:39 |
| Endereço IP | O endereço IP ao qual o computador utilizado para o acesso ao Moodle tinha no exato momento do acesso. | 187.86.133.66 |
| Nome Completo | Nome cadastrado do aluno no ambiente virtual. | José Alves da Cunha |
| Ação | Tipo da ação geradora do log que foi efetuada pelo usuário. | Course view (http://Moodle.cavg.ifsul.edu.br /course/view.php?id=75) |
| Atividade | Se refere ao local/link onde foi efetuada a ação | Download - LibreOffice (BrOffice) |

Tabela 3 – Quantitivo de dados utilizados - Fonte: Queiroga (2017)

| Cursos | Quant. Logs | Nº de alunos | Evadidos | Concluintes |
|---|---|---|---|---|
| Curso 1 | 682.773 | 407 | 212 | 195 |
| Curso 2 | 1.033.910 | 729 | 301 | 428 |
| Curso 3 | 933.221 | 615 | 246 | 369 |
| Curso 4 | 1.051.012 | 752 | 354 | 398 |
| Totais | 3.700.916 | 2503 | 1113 | 1390 |

conjuntos de treinamento.

Dessa forma, cada uma das aplicações foi metodologicamente pensada para se adaptar ao contexto de sua aplicação, sendo adaptada aos dados disponíveis, à quantidade de dados e suas caraterísticas.

### 3.4.5 Técnicas

No contexto desta tese, as aplicações geradas são baseadas em duas principais técnicas, a Learning Analytics e a Mineração de Dados Educacionais. Como explicado anteriormente, essas técnicas se diferenciam pelo seu ciclo, objetivo e finalidade, com a LA tendo um foco maior no processo educacional como um todo e a EDM tendo um maior enfoque no processo técnico envolvido.

No caso da metodologia desenvolvida para aplicação no ensino secundário, ela é

classificada como uma técnica de LA. Isso se dá porque ela carrega em si um foco muito alto no processo educacional como um todo. Em contrapartida, as metodologias que compõe a aplicação para o ensino universitário e o ensino técnico tem um enfoque maior na mineração de dados, na sua transformação e na extração de conhecimento bruto.

Dessa forma, para cada uma das metodologias criadas para as aplicações práticas foi definida uma técnica que norteou o desenvolvimento. No entanto, as aplicações desenvolvidas nesta tese, mesmo que classificadas em uma técnica, podem carregar um ou mais aspectos relacionados a outra técnica. Como é o caso da seleção de HP no processo de geração dos modelos de predição para o ensino secundário. Ainda assim, a técnica principal prevalece na quantidade de semelhanças.

### 3.4.6 Dimensões

Em pesquisas e implementações voltadas para aplicação de learning analytics, podemos apoiar o desenvolvimento no conceito de LA estabelecido por Chatti et al. (2013). Nesse conceito, a pesquisa é apoiada em 4 perguntas-chave: O QUE?, POR QUÊ?, COMO? e QUEM?. A teoria envolta nesse tema foi abordada anteriormente no capítulo 2, na seção 2.1.

Nesse sentido, buscamos adaptar as implementações voltadas para o ensino universitário e técnico para que possam ser comparadas com a implementação voltada para o ensino secundário, que é nativamente uma implementação de LA. Assim, na tabela 4 demonstramos as dimensões para cada uma das aplicações geradas.

### 3.4.7 Modelos Preditivos e Resultados Obtidos

Essa etapa do desenvolvimento das metodologias englobou a seleção dos algoritmos de classificação, a avaliação da dimensionalidade das bases, a geração dos modelos, as técnicas de hiperparâmetrização utilizadas e a avaliação dos resultados.

A busca pela redução da dimensionalidade foi uma preocupação latente no desenvolvimento das aplicações presentes nesta tese. Assim, procuramos testar técnicas já consolidadas que pudessem auxiliar nessa redução e no consequente incremento dos resultados.

Nesse contexto, a técnica que apresentou os melhores resultados foi a seleção de variáveis. Assim, optamos por utilizá-la nas aplicações para o ensino secundário e o ensino universitário.

Em ambas aplicações o algoritmo de machine learning selecionado para classificação depois de diversos testes foi o Random Forest. Esse algoritmo é conhecido pelos seus bons resultados em problemas no contexto desta tese e pelo sistema de poda que trabalha de forma robusta em casos de alta dimensionalidade de dados. Nesse sentido, em ambas as aplicações ainda foram utilizadas técnicas para hiperparâmetri-

Tabela 4 – Dimensões do ciclo de Chatti et al. (2013) nas aplicações desenvolvidas.

| Aplicação | O que? | Por quê? | Como? | Quem? |
|---|---|---|---|---|
| Ensino Secundário | Trajetórias dos estudantes na primaria, trajetória dos estudantes na secundária, faltas dos estudantes, recebimento de benefícios sociais, sistemas de gerenciamento acadêmico | Sistema de alerta antecipado para predição de risco de evasão ou retenção. | Combinação de dados de diversas fontes buscando formar um life time da trajetória do estudante desde a matricula no ensino primário até o momento da geração do modelo de identificação de risco para evasão ou retenção. | Tomadores de decisão da Agência Nacional de Educação Pública do Uruguai (ANEP). |
| Ensino Universitário | Interações dos estudantes com o AVA, dados oriundos dos sistemas de gerenciamento acadêmico e censo escolar. | Avaliação das bases de dados disponíveis e geração de modelos de predição para identificação precoce de e studantes em risco de evasão. | Recolhimento das interações dos estudantes com os ambientes virtuais de aprendizagem, que nesse modelo de curso são utilizados em sua grande maioria como repositório de conteúdo. Após isto, os dados são combinados com dados oriundos de diversas fontes para geração de modelos de classificação focados na predição precoce de estudantes em risco de retenção. | Professores, coordenadores de cursos e administração. |
| Ensino Técnico | Interações dos estudantes com o AVA | Predição de Evasão precoce | Modelagem de dados buscando a geração de um life time do estudantes durante as semanas de curso. Após o processamento, os dados são utilizados como entrada para geração de modelos de classificação de acompanhamento semanal. Esses modelos servem para acompanhamento do risco de abandono dos estudantes. Esses modelos ainda são hiperparâmetrizados utilizando o algoritmo genético proposto. | Coordenadores de tutoria, analistas de sistemas e administração. |

zação baseadas em buscas exaustivas, como o GridSearch.

Na aplicação voltada para o ensino técnico a distância foi o utilizado o algoritmo genético baseado na teoria darwinista. Nesse algoritmo, como citado anteriormente, diversos classificadores concorrem entre si e ao final somente o indivíduo mais apto é selecionado para a classificação. Assim, o algoritmo utilizado vai depender dos dados de entrada e do método de aptidão selecionado.

No quesito resultados, as 3 aplicações criadas apresentaram índices satisfatórios, com uma discriminação excepcional (AUROC > 0.90) ou próxima a esse valor. Esses valores, quando analisados a fundo, são mais satisfatórios, pois a depender do caso essas discriminações são obtidas em momento cruciais dos cursos para o tratamento dos possíveis causas da evasão e/ou retenção.

Nesse sentido, somente no modelo gerado no pré-ingresso do estudante no primeiro ano de ensino secundário na UTU o algoritmo não conseguiu estabelecer bons resultados. Isso se deu em partes pela falta de dados para treinamento dos modelos naquele momento do curso.

### 3.4.8 Implementação e Possíveis Interessados

As diferentes aplicações geradas no decorrer desta tese se encontram em diferentes momentos de implementação.

A aplicação voltada para o ensino secundário está em fase de avaliação e retreinamento pela ANEP, podendo ser aplicada na prática no início do próximo ano letivo. Para essa implementação foram criados diversos manuais técnicos e vídeos de treinamento que tem como objetivo demonstrar o funcionamento das funcionalidades desenvolvidas pela API. O planejamento de uso da informação gerada por essa API é para que ela tenha uma utilização pelos stackholders da ANEP, assim fomentando as tomadas de decisões baseadas em dados.

No entanto, as aplicações voltadas para a aplicação no ensino a distância e no ensino universitário são dependentes do dashboard para visualização e geração de predição baseado em machine learning, que atualmente encontra-se em desenvolvimento. Esse dashboard deve incluir diversas ferramentas de visualização de dados, acompanhamento dos estudantes e a geração de modelos de predição, voltados para os administradores dos ambientes virtuais de aprendizagem e os stackholders das instituições de ensino.

Assim, entende-se que as aplicações estão em estágios diferentes de implementação, com a API voltada para o ensino secundário significativamente mais adiantada. Ainda, que aplicações apresentam grupos de possíveis usuários e interessados diferentes, indo desde os professores até os administradores dos sistemas.

### 3.4.9 Comparação entre as aplicações

Nesta tese, as diferentes aplicações buscaram estabelecer um panorama da aplicação prática de LA e EDM em dados reais de diferentes contextos e níveis educacionais e com dados de diferentes tipos e origens. Isso, por si só, já é uma clara contribuição para a área de pesquisas. Isso se dá, pois, como citado anteriormente, os trabalhos tanto em LA quanto em EDM geralmente utilizam quantidades pequenas de dados, são restritos a um determinado ambiente de controle e dificilmente têm aplicações práticas envolvidas.

Dessa forma, esta tese busca demonstrar como a aplicação de LA e EDM pode contribuir com os diferentes contextos e níveis de ensino. Assim, abordamos a educação presencial de nível secundário e superior e a educação a distância de nível técnico subsequente em um modelo próximo ao que é utilizado em larga escala nos MOOCS.

As aplicações da educação presencial foram projetadas metodologicamente para trabalhar com dados oriundos de diferentes sistemas acadêmicos. Assim, em ambas foi possível demonstrar os impactos que a utilização de dados de diferentes sistemas pode trazer, como, por exemplo, o impacto positivo da utilização de dados dos AVAs mesmo na educação presencial e as diversas informações que esses podem revelar.

Já na aplicação da educação secundária, demonstramos o impacto na utilização dos dados de trajetórias anteriores dos estudantes.

Em ambas as metodologias presencias existia uma abundância de fontes de dados. Dessa forma, recolhemos a maior quantidade de dados possível para análise e modelagem. Foi demonstrado o impacto significativo que a modelagem de dados pode exercer no processo de transformação de dados em informações.

Essa abundância de dados não era presente na educação a distância. Assim, foi necessário adaptarmos metodologias para processamento e modelagem de dados, transformando as contagens de interações em um lifetime de acompanhamento do diário dos estudantes. Nesse processo, ficou demonstrado o alto poder preditivo desse lifetime, com valores de predição significativos já nas quatro primeiras semanas de curso.

No entanto, para alcançarmos índices de predição mais acurados dentro das possibilidades dos servidores que hoje mantêm o moodle em funcionamento, buscamos a criação do algoritmo genético proposto. Assim, foi criado um algoritmo que tenha menor consumo de hardware com resultados semelhantes aos obtidos por métodos de busca exaustiva, como o GridSearch.

Dessa forma, esta tese demonstra que o panorama de aplicação prática de LA e EDM é complexo e sinuoso, com os pesquisadores encontrando uma série de dificuldades no processo. Como exemplo dessas dificuldades, cito novamente os obstáculos presente no acesso aos dados nos diferentes contextos, as diferentes metodológicas

do ensino dentro de um mesmo país e as dificuldades para acesso a documentação sobre essas metodologias, a dificuldade de acesso a hardware específico para big data e mineração de dados e, principalmente, a qualidade dos dados disponíveis.

No entanto, independente do contexto, essas dificuldades podem ser rompidas com adaptações das técnicas existentes atualmente e um forte trabalho na modelagem de dados. Assim, as metodologias e aplicações práticas geradas no decorrer desta tese, apesar de terem objetivos internos diferentes, estão interligadas na demonstração de como as diferentes estratégias de EDM e LA podem ser aplicadas na prática como forma de auxílio aos processos de ensino.

# 4 CONTRIBUIÇÕES GERAIS

Esta tese tem como objetivo central a geração de diferentes métodos de aplicação de Learning Analytics e Mineração de Dados Educacionais para diferentes contextos educacionais. Assim, buscou-se demonstrar os panoramas práticos e teóricos da aplicação das técnicas baseadas em LA e EDM nos diferentes contextos educacionais, bem como as semelhanças e diferenças existentes nessas aplicações. Ainda, o trabalho desenvolvido e relatado nesta tese e nos artigos presentes nos apêndices A, B e C, busca contribuir com a pesquisa na área de LA e EDM com a geração de métodos de auxílio aos processos educacionais, principalmente no contexto regional da América Latina.

Com esse objetivo geral, e em cumprimento das metas específicas 4 e 6, este trabalho buscou abordar, através da primeira aplicação, a educação básica, voltada para crianças e adolescentes, com uma abordagem a nível nacional. Essa abordagem é uma tentativa de auxílio aos métodos de ensino na identificação dos estudantes com maior probabilidade de terem problemas em sua formação no ensino secundário. Ainda nessa linha, também foram testados modelos baseados na mesma metodologia para o ensino primário que obtiveram bons resultados e que podem ser implementados em um futuro próximo e com pequenas adaptações.

Em paralelo a geração da aplicação, o desenvolvimento da API auxilia no entendimento e na efetiva utilização dos modelos gerados. Nessa API, é possível que os usuários da metodologia visualizem informações básicas sobre a predição, efetuem novas predições, testem diferentes modelos e busquem pelo resultado de determinado estudante. Ainda, como um projeto futuro, imagina-se que essa API seja o start do desenvolvimento para um dashboard voltado para a visualização de dados estudantis no Uruguai.

A avaliação dos resultados provenientes dessa primeira aplicação demonstrou que os mesmos foram considerados satisfatórios, com 7 dos 8 modelos gerados obtendo uma discriminação excepcional (AUROC > 0.90) entre as duas classes (estudantes com e sem problema na formação).

Entre os modelos gerados, apenas um apresentou algum tipo de viés (UTU -

M!G1). Esse modelo conta com a menor quantidade de dados de entrada e apresenta vieses nas classes protegidas. Dessa forma, entre os oito modelos gerados nessa, aplicação somente esse último modelo citado não foi considerado apto para utilização. Outra limitação atual da metodologia é a falta de avaliação das implicações da pandemia da Covid-19 no processo educacional.

Nesse contexto, a ANEP atualmente está em fase de avaliação dos modelos, compreensão dos scripts, da metodologia gerada e retreinamento dos modelos. Isso só foi possível através de uma ampla documentação gerada para a metodologia, que envolve relatórios, manuais, vídeos de treinamento e acompanhamento da implantação. Com essas etapas, os resultados observados e a API, entende-se que se cumpriu o ciclo geral da Learning Analytics e o processo pode ser replicado e reajustado anualmente no momento estabelecido para retreinamento dos modelos.

Ainda, essa metodologia pode ser facilmente replicada para outros contextos com pequenas alterações, mesmo que nestes contextos existam dados diferentes dos disponíveis no Uruguai, o que contribui significativamente para a área de pesquisa. Ademais, como dito anteriormente, o processo de LA é um ciclo que deve ser melhorado e readaptado constantemente, assim como o processo educacional como um todo. Dessa forma, nas etapas atuais de implementação da metodologia para os anos de 2022 e 2023 surgiram novas perspectivas que podem ser implementadas para melhorar a metodologia como um todo, como a adoção de ferramentas de visualização robustas, melhorias gerais no processamento dos dados e automatização de processos.

A segunda aplicação, baseada no objetivo geral e nas metas 3 e 6, buscou a utilização de dados da educação superior oriundos de diversas fontes para que seja possível entender os padrões dos estudantes e, consequentemente, gerar conhecimento sobre a população estudantil. Essa aplicação foi focada principalmente em estudantes em risco de abandono escolar e, com esse objetivo específico, foram recolhidos diversos tipos de dados de diferentes sistemas, como ambientes virtuais de aprendizagem, sistemas de controle de notas, censos escolares, entre outros.

O cruzamento desses dados gerou uma base para processamento e análise que revelou que é possível identificar os estudantes mais propensos a reprovação em uma determinada disciplina a partir da quarta semana de curso, com uma excepcional discriminação (AUROC > 0,90). Nos experimentos realizados, esses resultados só foram obtidos nas bases criadas a partir da combinação de dados e pela contribuição dos logs dos ambientes virtuais de aprendizagem, mesmo quando utilizados somente como forma de apoio à educação presencial, onde, nesse caso em específico, os AVAs são utilizados apenas como repositório de conteúdo.

Dessa forma, entende-se que a metodologia criada foi capaz de demonstrar o impacto do engajamento dos estudantes no ambiente virtual, bem como contribuir na

identificação de diferentes aspectos envolvidos em sua formação. Ainda, a partir do processamento de dados e da EDA, emergiram diversas informações presentes nas bases de dados institucionais. Isso contribuiu ativamente para que ao final ao final do processo as informações se transformassem em conhecimento e, consequentemente, embasar a geração de políticas institucionais baseadas em dados, auxiliando efetivamente no processo educacional.

Nessa segunda aplicação, as contagens de interações semanais dos estudantes demonstraram um interessante poder preditivo, principalmente quando analisadas de acordo com o contexto de utilização dos AVAs na modalidade de ensino. Ainda, foram demonstradas questões importantes, como os momentos do dia, da semana e do curso onde os estudantes apresentam maior atividade nos ambientes virtuais.

Essa descoberta demonstrou-se importante para a UDELAR. Isso se deu porque, através dessa identificação, foi possível direcionar diferentes esforços para que os alunos tenham uma maior assistência nesse período, bem como que nesses momentos os AVAs tenham um redirecionamento de recursos computacionais, podendo mitigar problemas de quedas ou falhas no sistema.

Além disso, entende-se que o direcionamento desses esforços possa, futuramente, gerar a criação e a implementação de um dashboard que concentre dados de diversas fontes. Assim, ele deverá estar interligado aos ambientes virtuais de aprendizagem e aos outros sistemas de gerenciamento acadêmico, podendo servir como base de informações e avisos para os estudantes sobre o seu desempenho. Entende-se que a falta de automatização dos processos ainda é um problema a ser debatido, tornando-se, assim, um limitador atual da aplicação.

A terceira abordagem, baseada no objetivo geral desta tese e nas metas 5 e 6, foi uma expansão das pesquisas anteriormente desenvolvidas por Queiroga; Cechinel; Araújo (2015); Queiroga et al. (2016); Queiroga; Cechinel; Araújo (2017); Queiroga; Cechinel; Aguiar (2019); Detoni; Cechinel; Matsumura araújo (2015), buscando trabalhar com cursos de educação técnica híbrida sequenciais. Esses cursos são realizados de forma descentralizada, espalhado por diversos polos, e o estudante cursa somente uma disciplina por vez e tem encontros on-line e presenciais.

Nas etapas desenvolvidas anteriormente, grande parte dos comentários e sugestões versavam sobre como aumentar os resultados das predições. Isso se deu porque nesse modelo de curso massivo a quantidades de estudantes pode facilmente ultrapassar os 500 estudantes por turma. Assim, cada ponto percentual envolvido nos índices de acurácia dos classificadores pode representar uma quantidade expressiva de estudantes.

Dessa forma, buscamos diversas formas de otimizar os resultados, como melhorias no processamento, extração de dados, seleção de features e, principalmente, na seleção de hiperparâmetros. No entanto, nesse processo, esbarramos em questões-

chave, como a falta de recursos computacionais de maior capacidade. Esses recursos são de suma importância para que sejam feitas as buscas exaustivas por hiperparâmetros, que atualmente são o tipo de hiperparâmetrização mais utilizado.

Assim, buscamos o desenvolvimento de um algoritmo genético que utiliza a teoria da evolução darwinista para a escolha dos hiperparâmetros, apresentando um menor custo computacional. Essa abordagem foi publicada e está presente no apêndice C.

Nessa aplicação, foi possível a obtenção de resultados com uma AUROC até 10% mais acurada que o modelos com hiperparâmetrização padrão. Quando comparada as técnicas que utilizam buscas exaustivas, obtivemos resultados similares, mas com um menor custo computacional. Isso se deu, principalmente, pela não necessidade de testar uma quantidade expressiva de combinações, já que a técnica com GA converge para uma solução otimizada.

Nesse sentido, entende-se que essa terceira aplicação pode trazer benefícios a este tipo de curso massivo e a população no geral. Isso se deu, principalmente, quando analisamos que geralmente o tipo de estudante ingressante nesses cursos são jovens adultos que residem em cidades afastadas dos grandes centros universitários. Essa parcela da população necessita de formas de qualificação rápidas para otimizar seu acesso ao mercado de trabalho.

No entanto, essa abordagem ainda apresenta limitações, como a falta de implementação de métodos de parada mais efetivos, principalmente para os casos onde as soluções encontradas pelo algoritmo convergirem para um platô. Um trabalho futuro a ser implementado é a combinação com métodos como os algoritmos do tipo Random Search, onde poderia ser feita uma busca por grupos com o Random Search e depois uma busca nos vizinhos próximos com o GA. Assim, é possível maximizar o consumo computacional na busca dos hiperparâmetros.

Assim, entende-se que esta tese demonstrou os aspectos práticos e teóricos envolvidos no processo de aplicação de LA e EDM em diferentes contextos e níveis educacionais. Ainda, foram demonstradas as semelhanças e as diferenças nas diferentes aplicações e como as técnicas podem ser adaptadas ao contexto educacional.

Dessa forma, alcançamos o objetivo geral desta tese e de toda a pesquisa que vem sendo desenvolvida ao longo desta jornada. Mesmo assim, entende-se que existe um espaço em todas elas para trabalhos futuros que implementem melhorias em suas estruturas e, principalmente, nas tarefas de automatização.

Quanto à meta específica 1 definida anteriormente, esta tese apresenta no capítulo 2 a fundamentação teórica, bem como o estado da arte dos 3 principais temas ligados a pesquisas e projetos na área de predição educacional. Assim, a meta 1, "Investigar e documentar a fundamentação teórica e o estado da arte na utilização de Machine Learning, Learning Analytics e Educational Data Mining, principalmente como ferramentas de apoio e geração de conhecimento em diferentes contextos edu-

cacionais" foi cumprida.

Diante do exposto, entendemos que as contribuições científicas do projeto são claras e podem auxiliar o desenvolvimento de metodologias práticas para a compreensão dos dados gerados pelos diferentes sistemas educacionais. Além disso, essas abordagens contribuem com a teoria sobre as áreas relacionadas e geram aplicações práticas, que podem ser revertidas em prol dos processos educacionais.

# 5 CONSIDERAÇÕES FINAIS

Esta tese buscou apresentar os panoramas práticos e teóricos envolvidos no processo de desenvolvimento e aplicação de learning analytics e mineração de dados educacionais em diferentes níveis e contextos educacionais. Ademais, foram demonstradas as semelhanças e diferenças existentes nesse processo a depender de fatores como o contexto de aplicação, a diversidade de dados, os objetivos, entre outros.

Ainda nesse sentido, foram demonstradas como o processo de desenvolvimento de aplicações de LA e EDM necessitam de uma forte personalização para o contexto da aplicação. No entanto, apesar dessa personalização, a fundamentação básica para as propostas de aplicação pode ser aproveitada de outras aplicações e foi exposta nesta tese.

A detecção precoce de grupos de alunos com risco de evasão é uma condição importante para reduzir o problema da evasão, pois possibilita proporcionar algum tipo de atendimento direcionado a situações específicas do aluno. Atualmente, o processo de identificação desse grupo de alunos é manual, subjetivo, empírico e sujeito a falhas, dependendo primordialmente da experiência acadêmica e do envolvimento dos docentes.

Considerando que os docentes desempenham inúmeras atividades, é difícil acompanhar e reconhecer as necessidades de cada aluno e identificar aqueles que apresentam risco de evasão (MANHÃES et al., 2011; QUEIROGA; CECHINEL; ARAÚJO, 2017). Assim, a aplicação dos processos de Mineração de Dados Educacionais e Learning Analytics tem potencial para se tornar uma forma de auxílio transparente nos processos de ensino e aprendizagem, principalmente na identificação de estudantes com maior probabilidade de problemas em sua formação, como a evasão e a retenção.

Nesse sentido, o presente trabalhou buscou apresentar diferentes abordagens visando aplicações de LA e EDM, objetivando que as mesmas possam auxiliar a compreender os fatores que são indicativos de um possível problema na formação do estudante em diferentes níveis. Para isso, foram utilizados três níveis educacionais que têm como alvo públicos diferentes. Os resultados obtidos nos 3 casos são considerados satisfatórios, com uma provável aplicação e utilização dessas tecnologias

criadas em um mundo real.

Um ponto importante a ser salientado é que o algoritmo genético proposto não foi utilizado nas outras duas abordagens. Isso ocorre porque o AG ainda é uma abordagem experimental, carecendo de um maior desenvolvimento para aplicação prática como as das outras duas abordagens.

Contudo, na aplicação com dados do ensino primário e secundário não são apresentados dados específicos do sistema educacional de origem, bem como da análise exploratória executada. Isso ocorre porque esses dados podem ser considerados sensíveis, principalmente no contexto da Lei Geral de Proteção de Dados Pessoais (LGPD). No entanto, os relatórios entregues à ANEP apresentam esses dados com uma maior clareza e sua divulgação ainda é discutida internamente.

Assim, entende-se que esta tese se diferencia dos demais trabalhos produzidos na área, pois traz resultados teóricos e práticos, utilizando grandes volumes de dados e dados de estudantes de diferentes contextos educacionais. Ainda, são demonstradas que as aplicações práticas geradas, principalmente no estudo apresentado no apêndice B, estão resultando em benefícios aos sistemas educacionais onde estão sendo utilizadas, bem como as aplicações demonstradas no artigo presente no apêndice A estão em fase de avaliação final e implantação.

# REFERÊNCIAS


AFZAL, N. A study on vocabulary-learning problems encountered by BA English majors at the university level of education. **Arab World English Journal (AWEJ) Volume**, [S.l.], v.10, 2019.

AIZIKOVITSH-UDI, E.; CHENG, D. et al. Developing critical thinking skills from dispositions to abilities: mathematics education from early childhood to high school. **Creative education**, [S.l.], v.6, n.04, p.455, 2015.

ALJOHANI, O. A Review of the Contemporary International Literature on Student Retention in Higher Education. **International Journal of Education and Literacy Studies**, [S.l.], v.4, n.1, p.40–52, 2016.

ASIF, M.; HAYAT, M.; KHAN, S. Factors associated to high school students dropout in Malakand District Pakistan. , [S.l.], 2021.

AVVISATI, F. et al. Skills in Ibero-America: Insights from PISA 2015. **OECD Publishing**, [S.l.], 2018.

BACH, F. Breaking the curse of dimensionality with convex neural networks. **The Journal of Machine Learning Research**, [S.l.], v.18, n.1, p.629–681, 2017.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: **Learning analytics**. [S.l.]: Springer, 2014. p.61–75.

BARROSO, M. F.; FALCãO, E. B. Evasão universitária: O caso do Instituto de Física da UFRJ. **Encontro Nacional de Pesquisa em Ensino de Física**, [S.l.], v.9, p.1–14, 2004.

BASSI, M.; BUSSO, M.; MUÑOZ, J. S. Enrollment, graduation, and dropout rates in Latin America: is the glass half empty or half full? **Economía**, [S.l.], p.113–156, 2015.

BHAGOJI, A. N.; CULLINA, D.; MITTAL, P. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. **arXiv preprint arXiv:1704.02654**, [S.l.], v.2, 2017.

BLOSSFELD, H.-P.; KIERNAN, K. **The new role of women**: Family formation in modern societies. [S.l.]: Routledge, 2019.

BURRELL, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, [S.l.], v.3, n.1, p.2053951715622512, 2016.

BUSSO, M.; BASSI, M.; MUÑOZ, J. S. Is the glass half empty or half full? School enrollment, graduation, and dropout rates in Latin America. , [S.l.], 2013.

CAMPBELL, J. P.; DEBLOIS, P. B.; OBLINGER, D. G. Academic analytics: A new tool for a new era. **EDUCAUSE review**, [S.l.], v.42, n.4, p.40, 2007.

CANO, C. M.-v. A.; ROMERO, C.; VENTURA, S. Predicting School Failure and Dropout by Using Data Mining Techniques Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. , [S.l.], n.February, 2013.

CARVALHO, A. P. d. L. F. Algoritmos genéticos. **Instituto de Ciências**, [S.l.], 2009.

CECHINEL, C. et al. Mapping learning analytics initiatives in latin america. **British Journal of Educational Technology**, [S.l.], v.51, n.4, p.892–914, 2020.

CENSO, E. BR 2018-Relatório Analítico da Aprendizagem a Distância no Brasil. **Acesso em**, [S.l.], v.16, n.08, 2018.

CHATTI, M. A.; DYCKHOFF, A. L.; SCHROEDER, U.; THÜS, H. A reference model for learning analytics. **International Journal of Technology Enhanced Learning**, [S.l.], v.4, n.5-6, p.318–331, 2013.

CLOW, D. The learning analytics cycle: closing the loop effectively. , [S.l.], 2012.

COSTA, E. B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. **Computers in Human Behavior**, [S.l.], v.73, p.247–256, 2017.

COX, V. Exploratory data analysis. In: **Translating Statistics to Make Decisions**. [S.l.]: Springer, 2017. p.47–74.

CURY, C. R. J. A educação básica como direito. **Cadernos de pesquisa**, [S.l.], v.38, p.293–303, 2008.

DAHRENDORF, R. **Class and conflict in an industrial society**. [S.l.]: Routledge, 2022.

DAUD, A. et al. Predicting student performance using advanced learning analytics. In: OF THE 26TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB COMPANION, 2017. **Proceedings...** [S.l.: s.n.], 2017. p.415–421.

DELANO, R.; CORRêA, D. S. Redes na Educação a Distância: Uma Análise Estrutural do Sistema UAB em Minas Gerais. **Revista PRETEXTO.**, [S.l.], 2013.

DETONI, D.; CECHINEL, C.; MATSUMURA ARAÚJO, R. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. **Revista Brasileira de Informática na Educação**, [S.l.], v.23, n.3, 2015.

DI MITRI, D. et al. Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data. In: OF THE SEVENTH INTERNATIONAL LEARNING ANALYTICS & KNOWLEDGE CONFERENCE, 2017. **Proceedings...** [S.l.: s.n.], 2017. p.188–197.

DING, M.; YANG, K.; YEUNG, D.-Y.; PONG, T.-C. Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS & KNOWLEDGE, 9., 2019. **Proceedings...** [S.l.: s.n.], 2019. p.135–144.

DURYEA, S.; GALIANI, S.; ÑOPO, H.; PIRAS, C. C. The educational gender gap in Latin America and the Caribbean. , [S.l.], 2007.

ELISTIA, E.; SYAHZUNI, B. A. The correlation of the human development index (HDI) towards economic growth (GDP per capita) in 10 ASEAN member countries. **Jhss (journal of humanities and social studies)**, [S.l.], v.2, n.2, p.40–46, 2018.

FALL, A.-M.; ROBERTS, G. High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. **Journal of adolescence**, [S.l.], v.35, n.4, p.787–798, 2012.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [S.l.], v.17, n.3, p.37, 1996.

FERGUSON, R. Learning analytics: drivers, developments and challenges. **International Journal of Technology Enhanced Learning**, [S.l.], v.4, n.5/6, p.304–317, 2012.

FERNANDES, E. et al. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. **Journal of Business Research**, [S.l.], v.94, p.335–343, 2019.

FERNANDES, M. D. E.; BASSI, M. E. A disputa pela construção do Custo Aluno-Qualidade. **Retratos da Escola**, [S.l.], v.15, n.33, p.733–750, 2021.

FERREIRA, T. S. et al. Custos na administração pública: Análise dos custos educacionais em Luziânia/GO. **Revista Contabilidade e Controladoria**, [S.l.], v.11, n.2, 2020.

FINE, J. G.; DAVIS, J. M. Grade retention and enrollment in post-secondary education. **Journal of school psychology**, [S.l.], v.41, n.6, p.401–411, 2003.

GALVÃO, F. V. A Pesquisa sobre Custo-Aluno no Brasil: caminhos percorridos e possibilidades. **FINEDUCA-Revista de Financiamento da Educação**, [S.l.], v.11, 2021.

GARFIELD, J.; AHLGREN, A. Difficulties in learning basic concepts in probability and statistics: Implications for research. **Journal for research in Mathematics Education**, [S.l.], v.19, n.1, p.44–63, 1988.

GAŠEVIĆ, D.; DAWSON, S.; ROGERS, T.; GASEVIC, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. **The Internet and Higher Education**, [S.l.], v.28, p.68–84, 2016.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow**: concepts, tools, and techniques to build intelligent systems. [S.l.]: "O'Reilly Media, Inc.", 2017.

GRALKA, S. Persistent inefficiency in the higher education sector: evidence from Germany. **Education Economics**, [S.l.], v.26, n.4, p.373–392, 2018.

GREEN, A. Education and state formation. In: **Education and State Formation**. [S.l.]: Springer, 1990. p.76–110.

GREGORI, E. B.; ZHANG, J.; GALVÁN-FERNÁNDEZ, C.; FERNÁNDEZ-NAVARRO, F. d. A. Learner support in MOOCs: Identifying variables linked to completion. **Computers and Education**, [S.l.], v.122, p.153–168, 2018.

HAGEDORN, L. S. How to define retention. **College student retention formula for student success**, [S.l.], p.90–105, 2005.

HASSAN, S.-U. et al. Virtual learning environment to predict withdrawal by leveraging deep learning. **International Journal of Intelligent Systems**, [S.l.], v.34, n.8, p.1935–1952, 2019.

HERNÁNDEZ-LEAL, E.; DUQUE-MÉNDEZ, N. D.; CECHINEL, C. Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. **Heliyon**, [S.l.], v.7, n.9, p.e08017, 2021.

HERODOTOU, C. et al. Implementing predictive learning analytics on a large scale: the teacher's perspective. In: OF THE SEVENTH INTERNATIONAL LEARNING ANALYTICS & KNOWLEDGE CONFERENCE, 2017. **Proceedings...** [S.l.: s.n.], 2017. p.267–271.

HERODOTOU, C.; RIENTIES, B.; VERDIN, B.; BOROOWA, A. Predictive learning analytics 'at scale': Towards guidelines to successful implementation in Higher Education based on the case of the Open University UK. **Journal of Learning Analytics**, [S.l.], p.In–Press, 2019.

HOSOKAWA, R.; KATSURA, T. Effect of socioeconomic status on behavioral problems from preschool to early elementary school–A Japanese longitudinal study. **PLoS one**, [S.l.], v.13, n.5, p.e0197961, 2018.

JAYAPRAKASH, S. M. et al. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. **Journal of Learning Analytics**, [S.l.], v.1, n.1, p.6–47, 2014.

KIM, B.-H. Deep Learning to Predict Student Outcomes. **arXiv preprint arXiv:1905.02530**, [S.l.], 2019.

KOVACIC, Z. Early prediction of student success: Mining students' enrolment data. , [S.l.], 2010.

LATIF, A.; CHOUDHARY, A.; HAMMAYUN, A. Economic effects of student dropouts: A comparative study. **Journal of Global Economics**, [S.l.], 2015.

LEARNING ANALYTICS 1st International Conference on; KNOWLEDGE. **Banff, Alberta, February 27–March 1**. Disponível em: <https://tekri.athabascau.ca/analytics>. Acesso em: 2019-07-30.

LEE, Y.; CHOI, J. A review of online course dropout research: Implications for practice and future research. **Educational Technology Research and Development**, [S.l.], v.59, n.5, p.593–618, 2011.

LIZ-DOMÍNGUEZ, M.; CAEIRO-RODRÍGUEZ, M.; LLAMAS-NISTAL, M.; MIKIC-FONTE, F. A. Systematic Literature Review of Predictive Analysis Tools in Higher Education. **Applied Sciences**, [S.l.], v.9, n.24, p.5569, 2019.

LYKOURENTZOU, I. et al. Dropout prediction in e-learning courses through the combination of machine learning techniques. **Computers & Education**, [S.l.], v.53, n.3, p.950–965, 2009.

MACARINI, B. et al. Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. **Applied Sciences**, [S.l.], v.9, n.24, p.5523, 2019.

MANHãES, L. M. B. et al. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII SBIE-XVII WIE, Aracaju**, [S.l.], 2011.

MARCHBANKS III, M. P. et al. The economic effects of exclusionary discipline on grade retention and high school dropout. **Closing the school discipline gap: Equitable remedies for excessive exclusion**, [S.l.], p.59–74, 2015.

MÁRQUEZ-VERA, C. et al. Early dropout prediction using data mining: a case study with high school students. **Expert Systems**, [S.l.], v.33, n.1, p.107–124, 2016.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, [S.l.], v.5, n.4, p.115–133, 1943.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning**: An artificial intelligence approach. [S.l.]: Springer Science & Business Media, 2013.

MINAEI, B.; PUNCH, W. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System. In: 2003. **Anais...** [S.l.: s.n.], 2003. v.2724, p.2252–2263.

MINAEI-BIDGOLI, B.; PUNCH, W. F. Using genetic algorithms for data mining optimization in an educational web-based system. , [S.l.], p.2252–2263, 2003.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. (McGraw-Hill International Editions).

MOISSA, B.; GASPARINI, I.; KEMCZINSKI, A. Learning Analytics: um mapeamento sistemático. **Nuevas Ideas en Informática Educativa TISE**, [S.l.], v.2014, p.283–290, 2014.

MOISSA, B.; GASPARINI, I.; KEMCZINSKI, A. A systematic mapping on the learning analytics field and its analysis in the massive open online courses context. **International Journal of Distance Education Technologies (IJDET)**, [S.l.], v.13, n.3, p.1–24, 2015.

MOISSA, B.; GASPARINI, I.; KEMCZINSKI, A. Educational Data Mining versus Learning Analytics: estamos reinventando a roda? Um mapeamento sistemático. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE), 2015. **Anais...** [S.l.: s.n.], 2015. v.26, n.1, p.1167.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future—big data, machine learning, and clinical medicine. **The New England journal of medicine**, [S.l.], v.375, n.13, p.1216, 2016.

OCHOA, X. Learning analytics in Latin America present an opportunity not to be missed. **Nature human behaviour**, [S.l.], v.3, n.1, p.6–7, 2019.

OCHOA, X. et al. Need analysis of the students in programming courses in latin america. **Need Analysis Report. http:**//, [S.l.], 2011.

OECD. **Benchmarking Higher Education System Performance**. Paris, France: OECD Publishing, 2019. 644p.

OKUBO, F.; YAMASHITA, T.; SHIMADA, A.; OGATA, H. A neural network approach for students' performance prediction. In: SEVENTH INTERNATIONAL LEARNING ANALYTICS & KNOWLEDGE CONFERENCE, 2017. **Proceedings...** [S.l.: s.n.], 2017. p.598–599.

PARK, H. Prevalence and related risk factors of problem drinking in Korean adult population. **Journal of the Korea Academia-Industrial cooperation Society**, [S.l.], v.19, n.1, p.389–397, 2018.

PEÑA, L. P. de; PÉREZ, A. M. C. Review of some studies on university student dropout in Colombia and Latin America. **Acta universitaria**, [S.l.], v.23, n.4, p.37–46, 2013.

QUEIROGA, E.; CECHINEL, C.; AGUIAR, M. Uma abordagem para predição de estudantes em risco utilizando algoritmos genéticos e mineração de dados: um estudo de caso com dados de um curso técnico a distância. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2019. **Anais...** [S.l.: s.n.], 2019. v.8, n.1, p.119.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R. Um Estudo do Uso de Contagem de Interações Semanais para Predição Precoce de Evasão em Educação a Distância. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2015. **Anais...** [S.l.: s.n.], 2015. v.4, n.1, p.1074.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R. Predição de estudantes com risco de evasão em cursos técnicos a distância. In: XXVIII BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2017)), 2017, Recife, PE, Brazil. **Anais...** Sociedade Brasileira de Computação - SBC, 2017. p.1547–1556. ISSN: 2316-6533.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R.; COSTA BRETANHA, G. da. Generating models to predict at-risk students in technical e-learning courses. In: LEARNING OBJECTS AND TECHNOLOGY (LACLO), LATIN AMERICAN CONFERENCE ON, 2016. **Anais...** [S.l.: s.n.], 2016. p.1–8.

QUEIROGA, E. M. **Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados**. 2017. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Pelotas.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v.1, n.1, p.81–106, 1986.

RAITASALO, K.; ØSTERGAARD, J.; ANDRADE, S. B. Educational attainment by children with parental alcohol problems in Denmark and Finland. **Nordic Studies on Alcohol and Drugs**, [S.l.], v.38, n.3, p.227–242, 2021.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, [S.l.], v.33, n.1, p.135–146, 2007.

ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the State of the Art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, [S.l.], v.40, n.6, p.601–618, 2010.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [S.l.], v.10, n.3, p.e1355, 2020.

ROMERO; VENTURA. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [S.l.], v.3, n.1, p.12–27, 2013.

SALAZAR-FERNANDEZ, J. P.; SEPÚLVEDA, M.; MUNOZ-GAMA, J.; NUSSBAUM, M. Curricular Analytics to Characterize Educational Trajectories in High-Failure Rate Courses That Lead to Late Dropout. **Applied Sciences**, [S.l.], v.11, n.4, p.1436, 2021.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, [S.l.], v.3, p.210–229, 1959.

SANTOS, J. A.; CARVALHO PEREIRA, V. de. A destinação orçamentária da União e sua vinculação ao custo aluno nas Universidades Federais. In: CONGRESSO BRASILEIRO DE CUSTOS-ABC, 2019. **Anais...** [S.l.: s.n.], 2019.

SCLATER, N.; PEASGOOD, A.; MULLAN, J. Learning analytics in higher education. **London: Jisc. Accessed February**, [S.l.], v.8, n.2017, p.176, 2016.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, [S.l.], v.34, n.1, p.1–47, 2002.

SEGUNDO, F. R.; RAMOS, D. K. Soluções baseadas no uso de software livre: alternativas de suporte tecnológico à educação presencial e a distância. **Anais do 12 Congresso Internacional de Educação a Distância**, [S.l.], v.12, p.18–22, 2005.

SIEMENS, G. Learning analytics: The emergence of a discipline. **American Behavioral Scientist**, [S.l.], v.57, n.10, p.1380–1400, 2013.

SIEMENS, G.; BAKER, R. S. d. Learning analytics and educational data mining: towards communication and collaboration. In: OF THE 2ND INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 2012. **Proceedings...** [S.l.: s.n.], 2012. p.252–254.

SIEMENS, G.; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE review**, [S.l.], v.46, n.5, p.30, 2011.

SILVA FILHO, R. B.; LIMA ARAÚJO, R. M. de. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. **Educação por escrito**, [S.l.], v.8, n.1, p.35–48, 2017.

SILVA, M. R. d. O Ensino Médio após a LDB de 1996: trajetórias e perspectivas. **Ensino médio em diálogo**, [S.l.], 2012.

SWAN, K. Learning effectiveness online: What the research tells us. **Elements of quality online education, practice and direction**, [S.l.], v.4, n.1, p.13–47, 2003.

TORRACO, R. Economic Inequality, Educational Inequity, and Reduced Career Opportunity: A Self-perpetuating Cycle? **New horizons in adult education and human resource development**, [S.l.], v.30, n.1, p.19–29, 2018.

UNITED NATIONS EDUCATIONAL, S.; ORGANIZATION, C. **Global Education Monitoring Report 2020**. [S.l.]: United Nations, 2020.

VIEIRA, C.; PARSONS, P.; BYRD, V. Visual learning analytics of educational data: A systematic literature review and research agenda. **Computers & Education**, [S.l.], v.122, p.119–135, 2018.

WEI, H. et al. Predicting Student Performance in Interactive Online Question Pools Using Mouse Interaction Features. **arXiv preprint arXiv:2001.03012**, [S.l.], 2020.

WELLS, A. S.; CRAIN, R. L. Perpetuation theory and the long-term effects of school desegregation. **Review of educational research**, [S.l.], v.64, n.4, p.531–555, 1994.

WETZEL, J. N.; O'TOOLE, D.; PETERSON, S. Factors affecting student retention probabilities: A case study. **Journal of economics and finance**, [S.l.], v.23, n.1, p.45–55, 1999.

WHITEHILL, J. et al. Delving deeper into MOOC student dropout prediction. **arXiv preprint arXiv:1702.06404**, [S.l.], 2017.

WORSLEY, M. Multimodal Learning Analytics' Past, Present, and Potential Futures. In: CROSSMMLA@ LAK, 2018. **Anais...** [S.l.: s.n.], 2018.

XING, W.; GUO, R.; PETAKOVIC, E.; GOGGINS, S. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. **Computers in Human Behavior**, [S.l.], v.47, p.168–181, 2015.

YAKUNINA, R.; BYCHKOV, G. Correlation analysis of the components of the human development index across countries. **Procedia Economics and Finance**, [S.l.], v.24, p.766–771, 2015.

YOUNGMAN, F. **Adult education and socialist pedagogy**. [S.l.]: Routledge, 2018.

ZAFFAR, M.; SAVITA, K.; HASHMANI, M. A.; RIZVI, S. S. H. A study of feature selection algorithms for predicting students academic performance. **Int. J. Adv. Comput. Sci. Appl**, [S.l.], v.9, n.5, p.541–549, 2018.

ZEICHNER, K. Rethinking the connections between campus courses and field experiences in college-and university-based teacher education. **Journal of teacher education**, [S.l.], v.61, n.1-2, p.89–99, 2010.

ZOHAIR, L. M. A. Prediction of Student's performance by modelling small dataset size. **International Journal of Educational Technology in Higher Education**, [S.l.], v.16, n.1, p.27, 2019.

# Apêndices

# APÊNDICE A – Early prediction of at-risk students at secondary education: a countrywide K-12 learning analytics initiative in Uruguay

MDPI

*Article*

# Early prediction of at-risk students at secondary education: a countrywide K-12 learning analytics initiative in Uruguay

**Emanuel Marques Queiroga [1,*], Matheus Francisco Batista Machado [2], Virgínia Rodés Paragarino[3], and Cristian Cechinel [2]**

[1] Instituto Federal do Rio Grande do Sul, IFSul, Pelotas 96015560, Brazil; emanuelmqueiroga@gmail.com (E.M.Q)
[2] Centro de Ciências, Tecnologias e Saúde (CTS), Universidade Federal de Santa Catarina, UFSC, Araranguá 88906072, Brazil; contato@cristiancechinel.pro.br (C.C.); matheusmachadoufsc@gmail.com (M.F.M.)
[3] Comisión Sectorial de Enseñanza, Universidad de la República, Udelar, Montevideo 11200, Uruguay; virginia.rodes@gmail.com (V.R.P.)
* Correspondence: emanuelmqueiroga@gmail.com; virginia.rodes@gmail.com

**Abstract:** This paper describes a national wide Learning Analytics initiative in Uruguay focused on the future implementation of governmental policies to mitigate students retention and dropout at secondary education. For that, data from a total of 258,440 students was used to generate automated models to predict students at-risk of failure and dropout. Data was collected from primary and secondary education from different sources and for the period between 2015 and 2020. Such data contains demographic information about the students and their trajectory from the first grade of primary to the second grade of secondary school (e.g. students assessments in the different subjects during the years, amount of absences, participation in social welfare programs, zone of the school, among others). Predictive models using Random Forest algorithm were trained and their performances evaluated with F1-Macro and AUROC measures. The models were planned to be applied in different periods of the school year for the regular secondary school and for the technical secondary school ((before the beginning of the school year and after the first evaluation meeting for each grade). A total of 8 (eight) predictive models were developed considering this temporal approach and after an analysis of bias considering three protected attributes (gender, school zone, social welfare program participation), 7 (seven) of them were approved to be used for prediction. The models achieved outstanding performances according to the literature, with AUROC higher than 0.90 and F1-Macro higher than 0.88. The paper describes in depth the characteristics of the data gathered, the specifics of data preprocessing, the methodology followed for models generation and bias analysis, together with the architecture developed for the deployment of the predictive models. Among other findings, results of the paper corroborate the importance given by the literature of using the previous performances of the students in order to predict their future performances.

**Keywords:** classification; educational strategies; secondary education; learning analytics; at-risk prediction; dropout prediction; bias analysis; fairness in machine learning

**Citation:** Queiroga, E.M.; Machado, M.F.; Paragarino, V.R.; Cechinel, C.; . *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Academic Editor: Firstname Lastname

Received:
Accepted:
Published:



## 1. Introduction

The educational system of Uruguay has experienced important problems associated with backwardness and disengagement in the recent decades [1]. Even though the system is characterized by universal coverage at the primary level, it is possible to observe that students grade retention, dropout rates and non-enrollment rates increase as the education system progresses, while age-appropriate coverage decreases [2]. As a result, a signification part of the students have difficulties to remain enrolled in the educational system [3,4].

For instance, during the transition from primary to secondary education, the educational system of Uruguay usually experiences a drop of 10% of students. Moreover, from the total of students at the age of 13 years old, 26% are over-aged for their grade, and 3%

are dropping out the system. In secondary education, during the transition from basic secondary education to upper secondary education (when students are from 15 to 17 years old), there is an increase of 20% of students that are over-aged for their grade, and the proportion of students who drop out the educational system increases 27%. In the year 2015, Uruguay experienced the lowest graduation rates during the 12 years of compulsory education. At last, 31% of the students graduations occurred at the age of 19 years old and 40% at the age of 24 years old [3].

Previous work conducted by [5] explored social, economical, historical and political aspects associated with this situation in Uruguay. According to [5], lag, dropout and absenteeism are the three most important explanatory factors related to educational disengagement. Therefore, the identification of these aspects in educational trajectories allows one to establish early action in order to mitigate the risks and increase the chances of academic success.

The abundant amount of data generated by the digitalization of academic management systems has opened new perspectives for the analysis of educational data. The approach known as Learning Analytics [6] seeks to understand and improve educational processes through the multi-technical processing of data and products generated by students and teachers [7]. The field of Learning Analytics aims to develop data based educational solutions that can be useful for the many stakeholders involved in the teaching and learning processes so that such process can be constantly improved [8].

Among the techniques used by Learning Analytics one can mention statistical models, educational data mining (EDM), machine learning, natural language processing (PLN), computer vision and new algorithms resulting from research in Artificial Intelligence. These techniques allow the processing of large volumes of data from different educational systems to generate solutions to support decisions focused on the improvement of the different educational scenarios [9–13].

The present paper presents the methodology followed for the development of automated models to detect students at-risk of dropout at the secondary level in Uruguay. For that, a Clow cycle method [12] was adopted as baseline for the steps executed. In addition to the creation of the predictive models, the paper covers a deep exploratory analysis of the data used for the work together with the description of the resulting system developed to identify students at-risk of disengagement. The experiments and implementation developed in this work continue previous work conducted by [14,15]. Moreover, the work presented here was conducted under the fAIrLAC initiative of the Inter-American Development Bank (IDB). The fAIrLAC intends to influence public policies by promoting the development of artificial intelligence (AI) solutions in a responsible and ethical way[16][17].

The present paper intends to answer the following research questions:

- **RQ1** - Is it possible to generate a LA-based methodology that encompasses data acquisition, data transformation and the generation of models that can help to early identify students at risk of dropout at secondary level?
- **RQ2** - Is the transformation of data from different databases into time series a viable alternative from a preprocessing point of view? If so, are the final results generated by the prediction models using this technique satisfactory?
- **RQ3** - Is it possible to generate and analyze explainable models based on machine learning so that biases can be identified and corrected when necessary?
- **RQ4** - Which features are the most important to early predict students at risk of dropout in Uruguay at the secondary level?

The remainder of this paper is organized as follows. Section 2 describes some characteristics of the Educational System in Uruguay and section 3 presents the theoretical background and related work. Section 4 explains the methodology followed in the present work and section 5 describes the models generated for predicting students at-risk. Section 6 presents the most important results achieved by this project and section 7 depicts how the predictive models were deployed to the authorities. Finally, section 9 remarks the most important findings of this work.

## 2. Context Understanding: overview about education in Uruguay

Uruguay is located in the extreme south of Latin American, with a population around 3.4 million inhabitants and comprising 176,215 million square kilometres. Uruguay presents a huge concentration of population in urban areas (92% of population). Moreover, about 50% of the population live in the metropolitan region of the capital (Montevideo). In the context of Latin America, Uruguay is the third country in the Human Development Index (HDI) with a rate of 0.817 [18,19], and currently has one of the highest levels of connectivity in Latin America, with more than 80% of the population with access to the internet [20].

The Uruguayan basic education system comprises preschool, primary and secondary education, with public schools accounting for around 85% of enrollments [21]. In addition to this, university education is characterized by a policy of free and unrestricted admission, with no other condition than the completion of high school to be admitted to the university. University of the Republic (UDELAR) is the most important player with 90% of enrollments in higher education [11,22]. The educational system as a whole is managed by the National Administration of Public Education (ANEP[1]), a government agency responsible for planning, and managing public educational policies. For the present initiative, ANEP is the key stakeholder interested in the predictive models, and it was responsible for providing all the databases required for that.

Uruguay has been developing a series of social policies to combat inequality. Within these policies, one can highlight the Ceibal Plan [23,24]. The Ceibal Plan is a series of educational programs aimed at the digital inclusion of the Uruguayan population. These programs are based on a tripod of proposals aimed at students, teachers and students' families. In this context, a series of activities are developed, seeking to improve the quality of education through technological systems based on Information and Communication Technologies (ICT's).

For instance, one of the outstanding programs within Ceibal is the called "One laptop per child", where since 2007 the government has being distributing a laptop to each child enrolled in basic education and creating a network of technological assistance for such equipment throughout the country. In addition, there are several other programs that seek to include the tripod involved in the project, with programs aimed at training and qualifying teachers, involving families in educational activities, producing technological educational resources, providing free internet to students in schools and at home and technological educational activities aimed at student development, such as teaching robotics.

However, despite these multiple efforts, the Uruguayan educational system still faces high rates of students retention and disengagement. This situation is already felt in the early years of primary education. For instance, in 2012 around 27% of fourth-year students of primary experienced some kind of delay in their training [1].

Primary education in Uruguay begins at the first grade (for children at the age of 6) and ends at the sixth grade (for children at the age of 11). Secondary education is divided into two cycles (basic and upper secondary education) each with a duration of 3 years. The basic cycle of secondary education comprises the seventh, eighth and ninth grades, for children from 12 and 14 years old. Upper secondary education is also known as bachillerato, it lasts 3 years and it completes the education cycle for young people. This cycle can be compared to the high school in Brazil and United States.

This work focuses on the basic secondary education (seventh, eighth and ninth grades). Education for children in these groups is divided into two different models in Uruguay: normal secondary education (named as CES) and Technical Vocational Education (named as UTU). These different teaching models have their own characteristics, such as different methodology, calendar, schools and courses. Still, these educational models present several sub-models of their own, which will be briefly mentioned later in this work.

[1] Administración Nacional de Educación Pública (ANEP) - https://www.anep.edu.uy/acerca-anep

## 3. Theoretical Background

Learning Analytics (LA) is a recent area of research that emerged during the early 2000s [25] and which established itself as a new field during the first LAK (Learning Analytics and Knowledge conference) in 2011 [8,26]. According to [27], Learning Analytics can be defined as the "the measurement, collection, analysis and description of data about the students and their contexts, to understand and optimizing learning and the environments in which it takes place"[26].

LA is considered a multidisciplinary research area that encompasses a number of other research fields such as Machine Learning, Artificial Intelligence, Statistics, Data Visualization, among others [28][8]). LA seeks to make use of different techniques from these fields to develop methods that can help to improve learning in the different educational scenarios.

According to [29], LA aims to fully understand the many dimensions related to learning, and it seeks to analyze the different aspects associated to specific situations and problems faced in the education. These problems may be a student finishing or not a given course, or achieving or not a certain performance in a given assessment, for instance. The idea is to observe and analyze the scenarios observing the behavior of the different parties (student, professor, coordinator) by using a more holistic method [29]. Therefore, LA involves a continuous cycle process that is always improving itself and that does not have a predetermined end. LA solutions and strategies should be constantly being tested and re-evaluated. This is one of the reasons why [25] considers the field of LA maintains a deep proximity to other areas other than Educational Data Mining, such as Business Intelligence (BI) and Semantic Web and Recommendation Systems.

The present work can be classified under the scope of Predictive Learning Analytics as it is focused on the development of automated models to early predict students at-risk of dropout. The remainder of this section will present a brief bibliographical review of related works.

### *3.1. Related Work*

With the growing interest in Predictive Learning Analytics, several researches seek to model data coming from educational institutions in order to extract information and knowledge that can be used to improve teaching and learning processes. Predictive Learning Analytics problems are basically divided between performance prediction [30,31] and dropout prediction (evasion prediction) [32–34]. However, according to [35] these both types of prediction are linked, as performance is a relevant factor for student retention, with some studies pointing out that poor performance can lead students to disengagement [11, 36,37].

Predictive Learning Analytics may use data coming from different sources and types, such as, academic systems [31,35], learning environments [32,33,38,39], demographic information [30,31,40] expenditure and income data [30,41] and multimodal data coming from sensors and other sources [42,43].

Existing works usually test several classifiers in order to select the ones with the best performances. Among them, it is possible to see some converging towards the use of decision trees with emphasis on the random forest approach [30,34,43–45]. Decision trees are algorithms used for supervised classification that generate a tree structure that sorts the unknown samples. The approach uses the data coming from the training dataset in order to create a tree able to classify the unknown samples without necessarily testing all the values of their attributes[43,46]. Decision trees are considered a white-box approach, with models that are understandable and readable by humans. This is an important feature for educational scenarios as it allows some sort of explicability about the reasoning behind a given decision or prediction.

Although the majority of Learning Analytics initiatives are normally restricted to smaller datasets related to disciplines, courses, institutions or case studies, a number

of works have also started to explore information related to wider contexts and using educational data covering an entire country or state [47].

This is the case of the work developed by Frostad *et al.* [48] which evaluated the chances of a secondary student to dropout. The authors developed regression models using socio-demographic data from 2,045 students from secondary school of the Sør-Trøndelag Norwegian region. The authors identified a number of factors associated to students who dropped-out, such as the mother's instruction level, the level of support provided by the school and the teacher, and the amount of friends the student has inside his/her school class.

The use of data about the performance of the students in previous years to predict dropout is an approach that is also being adopted in the literature. For instance, Nagy and Molontay [49] obtained satisfactory results for predicting dropout at tertiary level by using demographic information and data about the performances of the students in secondary school. The authors used data from 15,825 undergraduate students (from Economics program) and achieved a AUC between 0.808 and 0.811 to predict students dropout. In the same direction, Lehrl *et al.* [50] used data from 554 students since pre-primary school to evaluate how learning and performance in the early years affects future educational problems such as dropout and retention at secondary school. The authors demonstrated that results in early years of school are directly related to the performances in secondary, specially when considering the topics of reading, language and alphabetization. This is an important finding that encourages the use of data from primary education to predict possible problems on secondary education.

In Latin America, researches such as [15,33,51] seek to map large amounts of data in academic systems in order to predict results and the situation of students. Marquez [33], for example, proposes a system based on evolutionary algorithms to predict the dropout rate of high school students in Mexico. For this, data with 60 attributes are used, ranging from the admission test to the research data distributed to students and obtaining satisfactory results in the prediction. Moreover, Macarini *et al.* [15] describes a countrywide K-12 learning analytics initiative in Uruguay, focusing on better understanding Uruguay's educational data and the secondary level students trajectory inside the educational system. In that work, several databases were used to generate association rules related to students at-risk of failure. Clustering techniques were also applied to better understand the characteristics of the different groups of students. The authors reported important findings such as that the amount of absences (non-attendances both non-justified and justified) can be used as a predictor of risk of failure. Dashboards were also provided for visualize students trajectories along the school years, and to compare students performances in the different subjects between schools. At last, the authors described a total of 8 main challenges faced during the implementation of a countrywide LA initiative. The work of [15] was used as a basis for the implementation of the present project.

Another remarkable initiative, is the work of Hernández-Leal *et al.* [51] which uses educational data from several sources of primary and secondary education at the State of Santander (Colombia) The authors integrated data originated from different educational levels to search students patterns related to their performances. The authors used different data modeling techniques, such as, decision trees and t-SNE clustering. Among the results, the authors demonstrated that the performance of the students in previous years is associated to their current performances, and that some socio-demographic features (such as social level and zone residence) are also important predictors of students failure.

## 4. Methodology

This section introduces an Exploratory Data Analysis (EDA) and feature engineering to build a set of data (variables) that allows the development of automatic models for the early identification of students at risk of dropout. The methodology applied here is based on the Cross Industry Standard Process for Data Mining (CRISP-DM - Figure 1). The sequence explains the CRISP-DM model used and its 6 stages.

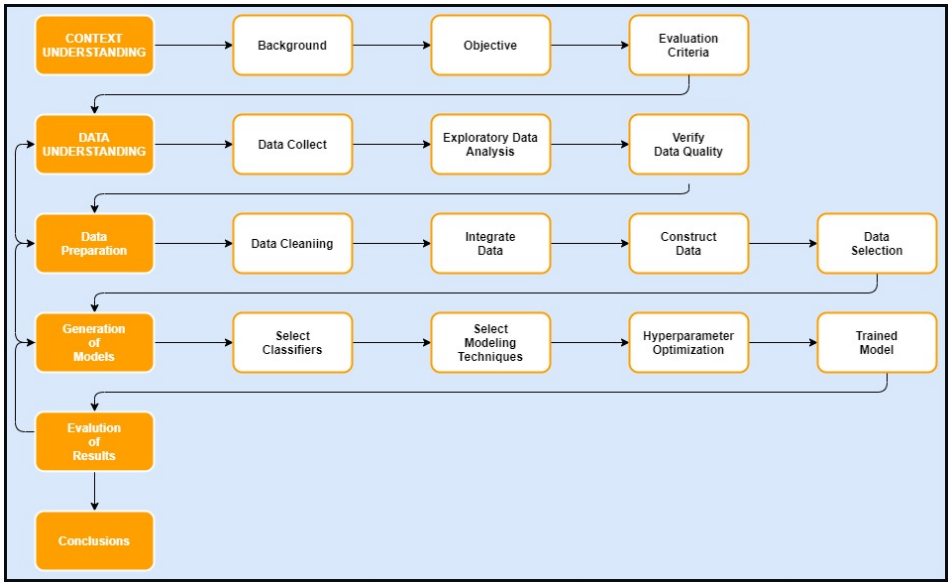


**Figure 1.** Modelo CRISP-DM.

- Context Understanding: identification and understanding the problem context, as well as defining the research hypotheses and the project requirements.
  - Background: understanding of the problem to be worked on and formulation of research hypotheses.
  - Project objective: definition of research objectives and questions.
  - Evaluation criteria: definition of the metrics that will be used to evaluate the results.
- Data Understanding: Consists of data collection and Exploratory Data Analysis (EDA), as well as the search for relevant sources that can add data to the project. In this phase, data is collected, different attributes are analyzed and their qualities are measured.
- Data Preparation: Consists of the 4-step feature engineering process:
  - Integrating Data: is the process of combining data from different databases into an integrated database.
  - Data Cleansing: is the process of detecting and correcting or removing incorrect or corrupt records as well as inconsistent data.
  - Data Building: is the process of creating variables (resources) that do not exist in the original data.
  - Data Selection: is the process of selecting and fitting the data that will be used as input in the predictive models. It can contemplate stages of handling outliers and deleting irrelevant data.
- Model Generation (Modeling): It is an iterative step that occurs in conjunction with data preparation and in which different models are tested with different input sets and hyperparameters.
- Results evaluation: In this step, the selected models are evaluated based on the metrics and objectives established in the previous steps. Models that meet the success criteria are delivered.
- Delivery and Conclusions: This stage consists on the delivery of the models, together with the manuals and the training of the ANEP technical team to use the solutions (to generate databases and retrain the models for the coming years).

### *4.1. Data Understanding*

The data used in this work were provided by the National Administration of Public Education (ANEP), and collected from nine different educational management systems.

The different databases gathered for this project were preprocessed and transformed to generate three main datasets used for model generation: 1)Primary Education database (PE), 2)Regular Secondary Education database (CES)[2] and 3) Technical Secondary Education database (UTU)[3]. These databases were built from information collected from several other secondary databases, such as: database of students trajectories and performances, database about social welfare programs, database related to the absences of the students, database with information about schools, among others. The data was available for the period from 2015 to 2020. During this period, 261,446 students completed their primary education. From these, 258,440 are present in the secondary databases (194,636 at CES and 63,804 at UTU), while 3,006 do not appear in the secondary databases. For these 261,446, we have the complete information cycle (complete primary education + first and second grades of secondary education). The specifics of each database are presented in the following sections.

#### 4.1.1. Primary Education Database (PE)

The objective of the work with the primary education database was the creation of a data structure that would allow the integration of the trajectory of students during the primary education with the data of students in secondary. Such integration allows the development of models to predict students dropout before they begin their secondary studies. For that, data were collected from 614,307 students born between 2004 and 2013. These students belong to 2,088 schools distributed in the 19 departments (states) of Uruguay. From the total of students in the database, 62,601 presented information for complete primary education cycle (data from the first The data was available for the period from 2015 to 2020. During this period, 261,446 students completed their education. From these, 258,440 are present in the secondary databases (194,636) and UTU (63,804), while 3,006 do not appear in any of these databases. So, for these 261.446, we have the complete information cycle (primary data + secondary data) in the two main planes.to the sixth year of primary). Examples of data stored at PE are: student scores on assessments, school code, class, department, jurisdiction, type of school zone (rural or urban), area and subarea, among others.

#### 4.1.2. Regular Secondary Education Database (CES)

The work with the Regular Secondary Education Database (CES) was guided by the creation of a data structure that would allow the early identification of students who showed indicators of failure or dropout in the first and second year of secondary education. Examples of information from this database and that were used for modelling students at-risk are: subjects (disciplines) taken by the students together with the performances achieved by them, presences and absences during the courses, where the students were part of social welfare programs or not, data about the students school. In total, data from 213,620 students were used from the period between 2016 and 2019. It is important to mention that the school year in secondary education in Uruguay is organized in three trimesters and that at the end of each trimester the teachers evaluate their students in the so called meetings (three meetings per year). After each trimester, students receive their assessments (ratings) in each subject. At these meetings, the absences of the students (justified and non-justified) are also computed. All these information are available at CES database.

#### 4.1.3. Technical Secondary Education Database (UTU)

The UTU database stored information about technical education that is offered in Uruguay integrated to the regular secondary education. In this educational model, students attend to the secondary and technical school at the same time (after finishing primary education). Technical education in Uruguay is organized in three years. In turn, each

[2] The acronym CES comes from Consejo de Educación Secundaria, i.e., Secondary Education Council
[3] The acronym UTU comes from Universidad del Trabajo del Uruguay, i.e. Labor University of Uruguay

**Table 1.** Number and percentage of students per database and final situation according to their gender

| Gender | UTU | | | | | | CES | | | | | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | | Apr. | | Prob. | | Total | | Apr. | | Prob. | | Total | |
| | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % |
| Female | 11.25 | 39,78 | 7.677 | 68,23 | 35,74 | 31,77 | 111.97 | 52,42 | 69.82 | 62,35 | 42.154 | 37,65 | 123.22 | 50,94% |
| Male | 17.02 | 60,22 | 10.449 | 61,36 | 65,80 | 38,64 | 101.64 | 47,58 | 60.26 | 59,29 | 41.378 | 40,71 | 118.67 | 49,06% |

year (grade) is organized in bimesters (four bimesters per year). The creation of the UTU database had similar data to the CES. However, these two teaching models contain different internal structures, such as the calendar and the number of evaluation meetings. Thus, it was necessary that the databases were generated separately despite having the same origin. In the end, the UTU database contained 46,994 students, of which 17,923 presented the complete education cycle in the database (ie, data from the first, second and third grades) and were considered in the forecasting process.

More details about how each database was created is given in subsection 4.3.

### *4.2. Fairness and Exploratory Analysis of Protected Groups*

The results obtained by machine learning algorithms are a direct reflection of the input data and the treatment dedicated to them. Some attributes (variables) can generate bias in the predictive models, generating wrong assumptions in the learning process and in the final result of the models. To avoid this kind of situation, [52] recommend the use of methods to assess the fairness of the results generated by the prediction algorithms. Precisely, [52] proposes to evaluate the datasets and to define those attributes that can generate some kind of unfairness in the prediction process (e.g. gender and other demographic data). After the creation of the predictive models, the performances are then compared with the fairness in relation to the so-called protected groups (groups related to the attributes previously selected).

For this work, the following attributes related to protected groups were defined to be taken into account for the identification of biases in the models: gender, social welfare program and school zone (location). The following subsections present an exploratory analysis of these attributes and the section 6.3 presents an overview of how the bias analysis was performed considering them.

#### 4.2.1. Gender

Assuring fairness in learning machine predictive models is a complex task [53]. Several authors point out that one of the most significant issues to be tackled in this direction is to ensure equity between genders during automated predictions. Considering that there one normally observes inequity between genders inside the data [53–55], it is expected the predictive models reflect unequal behavior towards one category in detriment of another. However, such predictions do not always accurate represent the behavior of these categories [55]; therefore, it is necessary to analyze gender as an input variable beforehand in order to avoid hidden biases inside the predictive models.

Table 1 presents the absolute frequencies and percentages of students according to gender in the three databases, together with the percentages of approval and/or "possible problem". As it can be seen in table 1, the UTU database has a higher percentage of male students than female students, precisely a difference of 20.44%. On the CES database, the percentages of male and female students is very close, with only 4.84% of difference between the two genders.

Regarding the percentage of students who engage in the educational system and complete their studies, the UTU database has 68.2% of female students completing their studies, against 61.4% of male students. The CES database presents similar data to the UTU database in this respect. For the CES, 62.3% of female students complete their studies, against 59.2% of male students. The data presented here refers only to the students whose trajectories are being used to generate the predictive models.


**Figure 2.** Correlation between students gender and final status for the CES database

**Table 2.** Number of students on social welfare programs per database

| Social Welfare Program in Primary | UTU | | CES | | Total | |
|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % |
| Yes | 19.858 | 70,21 | 87.866 | 41,13 | 107.724 | 44,53 |
| No | 8.422 | 29,79 | 125.754 | 58,87 | 134.176 | 55,47 |

Figure 2 shows Pearson's correlation between "Gender" and the students final status in the first and second grades of secondary (ResultYear1 and ResultYear2 respectively) for the CES database. As it is possible to see, the correlation coefficients are close to zero, indicating that there is no correlation between the students gender and their performances.

#### 4.2.2. Social Welfare Program

The attribute "Social Welfare Program" is a binary information representing whether the student or his/her family was part of governmental compensatory social policies throughout the student primary education trajectory. This information was integrated with CES and both UTU databases. Table 2 shows the number of students (or families) who were part of compensatory social policies during primary school. As it is possible to see in the table, in percentage terms, UTU students have participated more of compensatory social policies during primary education than CES students, precisely 70.2% at UTU against 41.1% at CES.

Table 3 presents the combination of gender and social welfare program attributes in primary school and the students final status at the first grade of secondary education. It is possible to observe that students of both genders who participated in social welfare programs had fewer problems in their education. When one analyzes the female gender, only 3.8% present a possible problem in their training, against 4.7% for males.

#### 4.2.3. School Zone

The third attribute considered to protect specific groups against possible bias is the area in which the school is located. This attribute indicates whether the student attends a rural or an urban school. Table 4 presents the number of students and their final status by grade and database. As it is possible to see in the table, there is a huge concentration of students in urban areas.

### *4.3. Data Preparation*

Data preparation was done using Python programming language and the following main libraries: NumPY, Pandas and Scikit-learn. This step included data integration, data cleaning, the derivation of new features, and data selection. The EDA stage played a significant role in data preparation, collaborating with insights and helping to identify

**Table 3.** Percentages of students with benefits in primary school considering gender and problematic situation.

| Gender | Result Grade 1 | Social Welfare Program in Primary | Students | |
|---|---|---|---|---|
| | | | Amount | % |
| F | Approved | No | 14,907 | 8.24 |
| | | Yes | 37,068 | 20.50 |
| | Possible Problem | No | 35,237 | 19.50 |
| | | Yes | 6,917 | 3.83 |
| M | Approved | No | 13,091 | 7.24 |
| | | Yes | 32,187 | 17.80 |
| | Possible Problem | No | 32,900 | 18.20 |
| | | Yes | 8,478 | 4.69 |

**Table 4.** Number of students in urban and rural areas for CES and UTU

| Database | Grade | Urban Zone | | | Rural Zone | | | Missing data | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aprov. | Failure. | Total | Aprov. | Failure. | Total | Aprov. | Failure. | Total | |
| CES | 1 | 89,758 | 46,876 | 136,634 | 3,752 | 1,703 | 5,455 | 841 | 606 | 1,447 | 143,536 |
| | 2 | 86,683 | 10,854 | 97,537 | 3,720 | 320 | 4,040 | 687 | 68 | 755 | 102,332 |
| UTU | 1 | 16,046 | 15,372 | 31,418 | 1,936 | 1,242 | 3,178 | 144 | 205 | 349 | 34,945 |
| | 2 | 9,102 | 6,944 | 16,046 | 1,215 | 721 | 1,936 | 66 | 78 | 144 | 18,126 |

issues such as the characteristics of the attributes (distribution, type, categories, etc), the impact of each database in the process and the importance of inserting new variables.

#### 4.3.1. Data Integration and Data Construction

The first step for data integration was to identify the educational path that students took after finish primary education. For that, the identification of the students on the PE database were compared to the identification of the students on the CES and UTU databases. After this step, it was possible to verify the following situations: students who dropped the educational system after primary education, students who engaged on the regular secondary education, and students who engaged on technical secondary education. In a second step, the following additional information was generated: number of years the student is in the databases, the first and last years of the student in the databases, and the first and final grade of the student in the databases. All these information helped to further consolidate the student's school life cycle for the years available in the databases (from primary school to the two first years of secondary).

The next step involved the transformation of the data contained in the three databases into time series. Time series is characterized by data collections that are directly interconnected by time, being widely used in various areas such as economics, statistics, finance and epidemiology [56]. For the context of this project, the most important sequences of data generated were: student evaluations during the years for the different subjects (disciplines), students/family in social welfare programs during the years, and students presences and absences during the academic cycle [57]. The motivation behind this strategy is to represent student's progress over time.

Here it follows a brief description of each database after transformation and integration:

- The PE database had 137 columns with information such as: students' evaluations in the different subjects separated by grade (from the first to the sixth grade), information about the school (location, code, department, among others), averages (means) of the students evaluations in each grade, student's final evaluations in each grade and the quartile where the student's evaluation is in comparison to the school. Finally, information from a total of 62,601 students was available on the database and with a temporal effect.

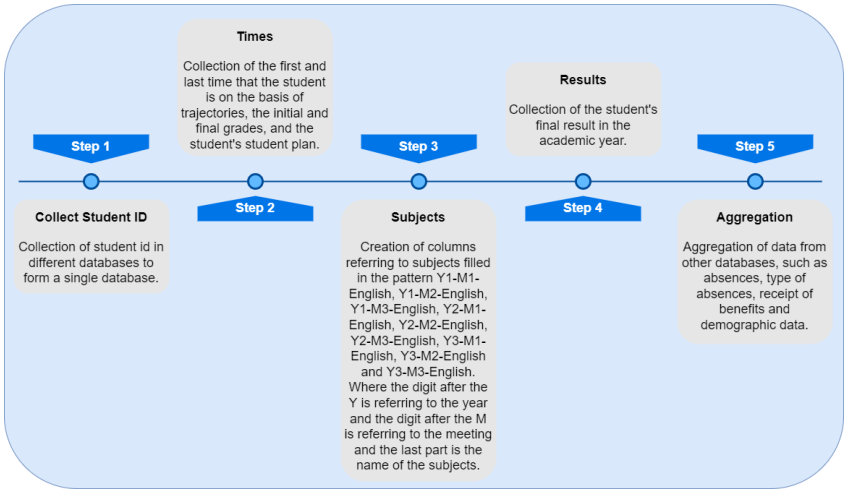


**Figure 3.** Database Structure

- At secondary educational level, there is a higher number of subjects (disciplines). Moreover, students are evaluated in meetings that take place every 2-3 months. In these meetings, students receive a evaluation rating (grade) for each subject they are enrolled (mathematics, foreign language, arts, among others). Schools perform regularly 3 meetings per year, and a fourth meeting at the end of the year when it is necessary and for the students who still need to take extra exams. Considering this scenario, the top 10 available subjects were filtered. New attributes were then generated for each evaluation of each meeting of each subject using the following syntax: Yi-Mj-Subject, where i stands for the number of the year (grade), and j stands for the number of the meeting. For instance, the attribute Y1-M1-English represents the assessment of the student in English subject at the first meeting of the first year. In addition, the number of absences of the students in each class of each subject were also computed. Absences in the context of this project could be justified absences and unjustified. absences.
- The generation of the UTU final database followed the same principles as for the CES generation. The difference here is that UTU has a greater number of disciplines, and the top 12 subjects were selected.

## 5. Generation of the predictive models

This stage involves a number of different aspects, such as: tests with different algorithms, selection of the algorithms to be used, filtering the data considering its characteristics and contribution to the performance of the models, and configuration of hyperparameters.

The target attribute (dependent variable) in this project is the student's final status at the end of the year. This status can be "approved", "failed" or "dropped out". Each one of these status was inferred from the databases available as it follows:

- To be considered "approved" in a given year (grade), the student must be enrolled in the courses of the next year (grade) in the database.
- The student is considered as "failed" in a given year if he/she is enrolled in the same grade in the database of the following year.
- The student is classified as "dropped out" if he/she does not appear enrolled in any courses in the database in the following year.

The categories "failed" and "dropped out" were grouped on a single category named "possible problem" in order to allow a binary prediction.

### 5.1. Selection of Algorithms

The first step for the generation of the models was the selection of algorithms that could meet the requirements established by the Responsible AI manual of the IDB fAIrLAC [58] and that presented good performances. According to the fAIrLAC manual, the predictive models needed to be explainable and auditable with the suggestion of using white-box models. The manual highlighted the importance of understanding the reasoning behind the automated decisions/classifications.

The following algorithms were initially tested with the first raw databases for CES and UTU obtained during preprocessing: Random Forest (RF), Decision Tree (DT), Adaboost (ADA), MultiLayer Perceptron (MLP), Naive Bayes Gaussian (NB) and Logistic Regression (RL). The neural network (MLP) was included to make a performance comparison, since machine learning studies usually show that MLP presents good results in this type of application [59].

The algorithms that presented the best results were Random Forest and MLP, both with very similar performances. Some tests were also performed with more advanced ensemble algorithms, such as Gradient Boost [60] and XGBoost [61], but both were discarded because they did not present significantly higher performances than the previous ones. Considering these results, Random Forest was selected to be used in the sequence of the project. This decision was made based on RF model architecture and the performance achieved in the first tests. Furthermore, even though the random forest is considered a black-box model, its models can be easily transformed into interpretable ones. For this transformation, we chose to use the TreeInterpreter package[4], which generates visualizations of the trees through the decomposition of the models.

### 5.2. Data preprocessing configurations for training

A total of 8 different combinations of configurations and data preprocessing were tested with Random Forest models to evaluate which ones presented the best performances. These different combinations are described as follows:

- I1 - Raw database - Application of the Random Forest algorithm with its default configuration using the raw data base.
- I2 - Weights (target variable) - Application of the algorithm in its default configuration using weights of the target variable with SKLearn's class-weight parameter.
- I3 - Feature Selection - feature selection application for selecting the top 20 attributes and training algorithm using these variables.
- I4 - Resource Selection + Database Balancing - Application of combination I3 in conjunction with the application of database balancing techniques.
- I5 - Resource Selection + Database Balancing + Weights (variable target) - Application of combination I4 with the increase of stage I2.
- I6 - Resource Selection + Balancing + Weights + GridSearch - Application of combination I5, adding the hyperparameterization of the algorithms with the GridSearch application [34,62].
- I7 - Pipeline generated from the TPOT automated learning library - use of the Python automated machine learning tool using TPOT genetic programming [63].
- I8 - Using the ImbLearn library [64] with the EditedNearestNeighbours, SMOTE and PCA methods.

### 5.3. Evaluation of predictive models

The strategy defined for the application and evaluation of the models followed the "Technical Manual of Responsible AI - AI Life Cycle" provided by fAIrLAC [65]. Thus, the algorithms were trained and tested using $k-10$ cross-validation. For the combinations where data balance was applied (iterations I4, I5 and I6), this was manually programmed

[4] https://github.com/andosa/treeinterpreter

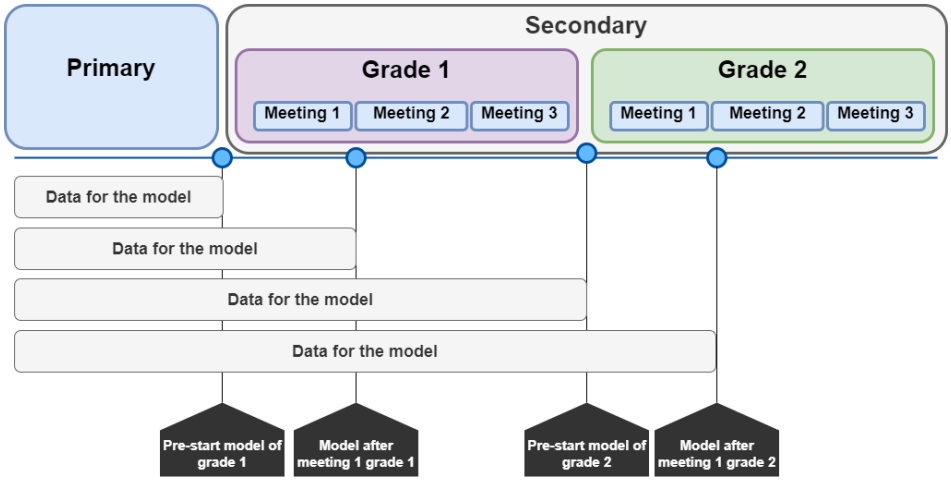


**Figure 4.** Temporal approach for the models and retraining periods

and applied only to the training dataset. In other words, data was divided into 10 folds and the one used for testing was not balanced.

Three different metrics were used to evaluate the performance of the models: F1-Macro, F1-Micro and AUROC. AUROC stands for the the area under the ROC curve, where the Y axis represents the True Positive Rate (TPR) or Sensitivity $(TP/(TP+FN))$ and the X axis represents the True Negative Rate (TNR) or Specificity $(TN/(TN+FP))$.

### *5.4. Temporal approach for the models and retraining periods*

The main idea of the work is to generate predictive models do early identify students at-risk of dropout/failure so that it is possible for professors and school coordinators to take actions in order to mitigate this situation. For that, it is necessary that the output of the models are provided in time for such actions to be taken. Precisely, 4 predictive models were generated for each database related to secondary education (CES and UTU). Figure 4 helps to illustrate the temporal approach adopted for the models. As it can be seen from the figure, two of these models are focused on predicting students at-risk in the beginning of the school year (one model for each grade), and the other two models are intended to be used after the first evaluation meeting of the school year. Here it follows a more in-depth explanation about each one of the models:

1. Grade 1 Pre-Start Model (M1G1): This model must be used before the 1st grade classes start and would be used from the primary data and social welfare program that students receive.
2. Grade 1 Post-meeting 1 model (M2G1): This model must be used after the first evaluation meeting and with the incorporation of new data obtained at that meeting (grades and absences results).
3. Grade 2 Pre-Start Model (M1G2): This model must be used prior to the start of Grade 2 classes and would be used from primary school data, 1st grade student outcomes, and data on social welfare program for high school students .
4. Grade 2 Post-meeting 1 model (M2G2): This model must be used after the first evaluation meeting of the 2nd grade and with the incorporation of the new data obtained in that meeting (grades and absences results).

It is suggested that all models should be retrained once a year, after the end of each school year.

## 6. Results

This subsection presents the best results obtained for each model considering the F1-Macro and AUROC evaluation metrics.

**Table 5.** Best results for CES predictive models

| Model | Best preprocessing | F1-Macro | AUCROC |
|---|---|---|---|
| M1G1-CES | I1 | 0.91 | 0.91 |
| M2G1-CES | I1 | 0.91 | 0.91 |
| M1G2-CES | I8 | 0.88 | 0.95 |
| M2G2-CES | I1 | 0.92 | 0.93 |

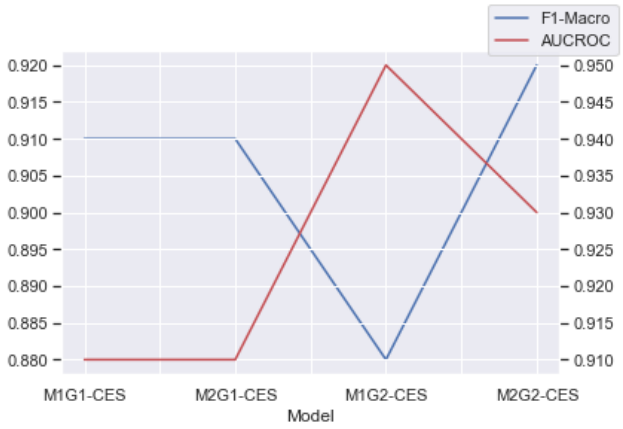


**Figure 5.** Performance of the models along the grades for CES database

### *6.1. Results for CES predictive models*

Table 5 presents the best results obtained for each of the CES models and the respective preprocessing combination that generated these results. As one can see in the table, all models have an F1-macro and an AUROC greater than 0.87. Figure 5 presents a temporal view of how the performance of the models evolves along the grades.

### *6.2. Results for UTU predictive models*

Table 6 presents the results of the models for the UTU database. As it can be seen in the table, the initial model M1G1-UTU presents the worst results with F1-Macro and AUCROC of 0.68. However, from the second model (M2G1-UTU ) the performances grow with results above 0.93 (F1-macro) and 0.95 (AUCROC). Figure 6 presents a temporal view of how the performances of the models evolves over the grades.

### *6.3. Analysis of Bias*

Bias analysis seeks to evaluate the predictive models regarding their ability to provide unbiased decisions towards any protected group. For this, issues such as the behavior of the models for the input data and whether the behavior is somehow biased are evaluated [66].

All resulting predictive models that used any attribute related to the protected groups previously defined (see section 4.2) were evaluated here using the What-if-tool[5]. Table 7

[5] https://pair-code.github.io/what-if-tool/

**Table 6.** Best results for UTU predictive models.

| Model | Best preprocessing | F1-Macro | AUCROC |
|---|---|---|---|
| M1G1-UTU | I7 | 0.68 | 0.68 |
| M2G1-UTU | I6 | 0.95 | 0.95 |
| M1G2-UTU | I8 | 0.93 | 0.95 |
| M2G2-UTU | I8 | 0.93 | 0.96 |

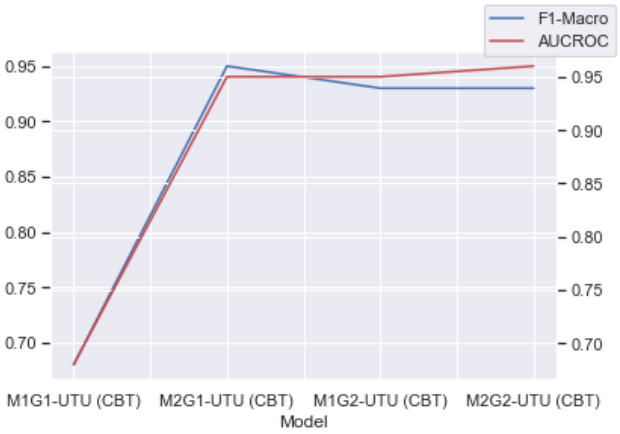


**Figure 6.** Performance of the models along the grades for UTU database

**Table 7.** Protected group attributes and the existence of bias

| Data Base | Model | Used protected group attributes (marked with X) | | | Bias |
|---|---|---|---|---|---|
| | | Gender | School Zone | Social Welfare Program. | |
| CES | M1G1-CES | - | - | - | - |
| | M2G1-CES | - | - | - | - |
| | M1G2-CES | X | X | X | No |
| | M2G2-CES | X | X | X | No |
| UTU | M1G1 | X | - | X | Yes |
| | M2G1 | X | X | X | No |
| | M1G2 | - | - | - | - |
| | M2G2 | X | X | X | No |

describes the models generated, the protected group attributes used by them and the bias found in the analysis.

Figure 7 presents the visualization of the bias analysis for the Social welfare program attribute in the M1G1 model. As it can be seen in the figure, the F1-Macro for category 1 (yes - social welfare program received) is 0.80, while F1-Macro for category 0 (no social welfare program received) is 0 (zero). This indicates a bias towards the students who participated in Social welfare programs.

With the detection of bias towards the Social welfare program and Gender attributes in the M1G1-UTU model, the remaining predictive models generated from the other pre-processing combinations were also tested. However, all remaining models also presented bias towards these attributes. Considering this, a new round of predictive models were generated, but removing the protected attributes. However, the resulting models for this round did not reach acceptable performances. Table 8 presents the confusion matrix for the model with the best performances when the attributes "gender" and "social welfare program" were not considered in the input. This model obtained an AUCROC of 0.49, an F1-Macro of 0.06 and an F1-Micro of 0.06. Due to the limitations of the models for M1G1-UTU, they were not recommended to be used in practice.

**Table 8.** Confusion matrix of model not using gender and social welfare program attributes

| | | Prediction | |
|---|---|---|---|
| | | 0 (Possible Problem) | 1 (Approved) |
| Real status | 0 (Possible Problem) | 179 | 6,534 |
| | 1 (Approved) | 9 | 267 |

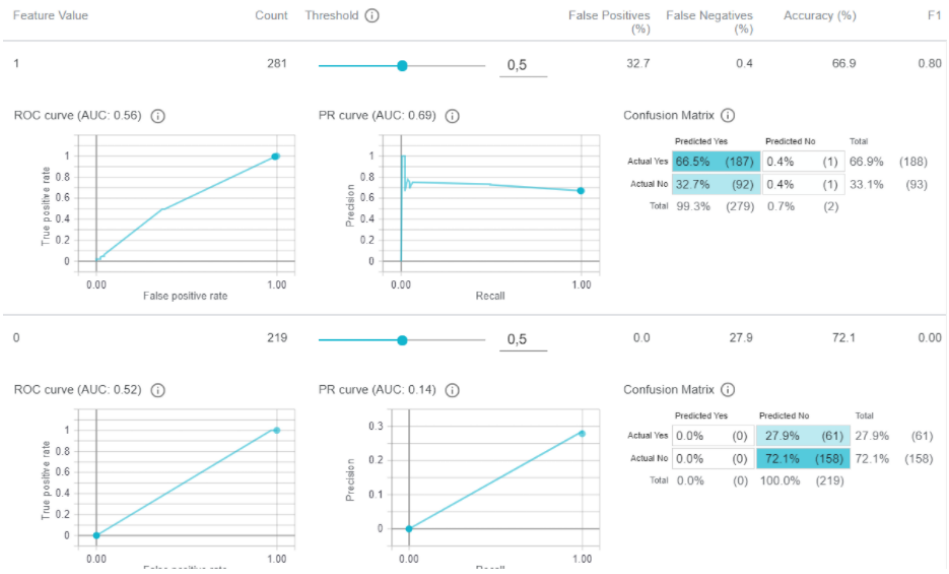


**Figure 7.** Bias analysis for Social welfare program attribute in the M1G1-UTU model

## 7. Predictive models deployment

LA is distinguished for defending a greater focus on the process and how the developed solutions are used to improve teaching and learning on a continuous way. The results providing by LA solutions should be incorporated on the teaching and learning cycle allowing interventions and providing new improved scenarios that are again continuously improved by these solutions.

The deployment of the predictive models and the strategies for retraining them are essential for completing a fruitful LA solution. As previously mentioned, the models developed here are recommended to be retrained twice a year (in the beginning of the school year and after the first evaluation meeting). Together with these recommendations, the project also developed a web API to use the models and provide the classification of the students according to their risk.

The API was developed using Python together with the Flask framework[6] to build the web server. Pandas and Celery[7] libraries were used in the API. Pandas is a library that facilitates the manipulation and treatment of data and Celery is an asynchronous queue of tasks implemented in Python, oriented to the passing of distributed messages in real time.

Queue handling was done using RabbitMQ [8] as a broker to transport messages between processes. Version v4 of RabbitMq was used, as in later versions messages larger than 128 MB are not handled by default as required by this project. A Redis[9] was used as a Backend in Celery as it is a very efficient key-value database for searching the results of tasks. Docker and Docker Compose were used to create Celery containers for Redis, Rabbit and API applications. To interact with the API, Python scripts and the application's front-end were developed. On the frontend, the javascript programming language and the ReactJS framework[10] were used to style the CSS3 frontend application. The Machine Learning API architecture is designed to support both asynchronous and synchronous predictions. Figure 8 presents an overview of how the API operates.

In asynchronous prediction, the user will send the CSV file containing the students data and will immediately receive a token to check the prediction results afterwards. The

6 https://flask.palletsprojects.com/en/2.1.x/
7 https://docs.celeryq.dev/en/stable/
8 https://www.rabbitmq.com/.
9 https://redis.io/
10 https://reactjs.org/

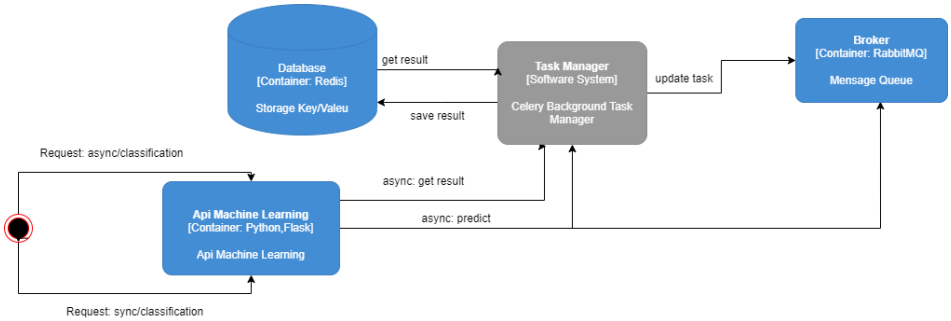


**Figure 8.** API Operation.

API will receive a HTTP/POST request with the CSV appended and the information of which model to use. This information may be sent from the web interface or from a terminal running a Python script. Celery is used by the asynchronous system to process the prediction task in the background, which will collect the information received in the HTTP request from the queue in RabbitMq. Then Celery will start processing the forecast. When processing is complete, a message is sent to the queue with the results of the prediction. Therefore, when the query is made by the user through the token, the prediction result will be retrieved. At this point, a cache of results will be created in Redis and the result will be sent to the user in HTML and CSV formats, along with the information displayed in the interface. In case the API has not finished to process the predictions while the user is consulting the results, the user will receive a message that the prediction is in process and that the user will need to try to retrieve the results latter again.

At the end of the prediction process, the user can consult the results and use them to make descriptions, such as the final status of the students (approved or possible problem) by region, by school, by participating in social welfare program, among others.

In synchronous prediction, the API will receive a HTTP/POST request. Together with the request it will be sent the CSV file containing the data and the information about which model should perform the prediction (e.g. M1G1-CES, M1G2-UTU). In this model, the API will process the CSV file and start making the prediction. The user who requested the prediction must wait for the process to finish before receiving the results. Once the prediction is completed, the user will receive the results in HTML and CSV together with other information about the prediction.

## 8. Discussion

**RQ1 - Is it possible to generate a LA-based methodology that encompasses data acquisition, data transformation and the generation of models that can help to early identify students at risk of dropout at secondary level?**

Yes, it is possible to generate this methodology. In general, the results found were satisfactory, with only one model (among 8) not showing good results and being discarded. All other models achieved AUROC values higher than 0.91 which is an outstanding discrimination when considering the scale provided by Gašević *et al.* [67]. These results are also confirmed by the F1-macro values where the worst value was 0.88. Specifically, when one analyzes only the four Pre-Start Models for CES and UTU (M1G1 and M1G2), three of them were able to classify students who would face a possible problem (failure or dropout) at that given grade and with great performances. Moreover, the four Post-Meeting 1 Models for CES and UTU (M2G1 and M2G2) presented great performances and were able to be used to classify students at-risk. These results confirm the viability of the proposed methodology to early identify students at risk at the beginning of the school year and after the first evaluation meeting of the school year. Moreover, from the results obtained by the models, it is possible to see that their performances increase as more information are provided as input for them.

However, this methodology still has some limitations, such as the need for annual manual collection of data, annual preprocessing and retraining of the predictive models. Future work will be focused on the direct integration of the predictive models with the different databases so that the data collection step could be automated, thus facilitating the process and optimizing the time spent in this part of the workflow.

Another issue to be deeply discussed in the next phases of the project is related to which stakeholders should have access to the prediction results. From the beginning, this project was designed to grant access to the results solely to ANEP's managers, so that these results could help the development of institutional and educational policies based on data. Considering that, teachers and students would not have access to the results at this initial phase, which is a practice aligned with the current Learning Analytics literature and recommendations for this kind of work [68]. Whether or not other stakeholders should also access the results of these predictions is still subject to future discussion.

**RQ2 - Is the transformation of data from different databases into time series a viable alternative from a preprocessing point of view? If so, are the final results generated by the prediction models using this technique satisfactory?**

Yes, from the preprocessing point of view, it was possible to generate time series from the collection and integration of data from the different databases, thus generating information and knowledge about the educational system and students.

Regarding the results, the models presented very good performances (with the exception of M1G1-UTU). Results found in the experiments show that it is possible to generate predictive models that can help in the identification of students with a tendency to face some problems (dropout or failure) during secondary school. However, these models need to be trained annually with new data that can represent the changes taking place in the students population. This can generate a complex situation, since these models use data prior to the Covid-19 pandemic and may not present good results with data from the pandemic period. It is understood that new educational scenarios that emerged from the pandemic will possible require future adaptations in the predictive models.

**RQ3 - Is it possible to generate and analyze explainable models based on machine learning so that biases can be identified and corrected when necessary?**

Yes, it is possible. In this project Random Forest algorithm was chosen considering as the algorithm to generate the models so that the reasoning of the models could be open and understood by humans. Moreover, currently there are several techniques and libraries that can assist in testing and verifying possible biases in machine learning models. In this work the What-If tool was used to help in this part of the analysis. The tool allowed to analyze the models regarding the bias in the attributes that were previously selected as protected. In the analyzes performed, only one model generated bias (M1G1-UTU). This model was eliminated from the work, as it was not possible to correct this bias after several interactions.

**RQ4 - Which features are the most important to early predict students at risk at secondary school in Uruguay?**

A large number of attributes were generated that served as input for the models. The strategy adopted to avoid the curse of dimensionality was the application of procedures for selecting input variables and to reduce them to the 20 most important ones, together with the use of the Random Forest algorithm, particularly suited to dealing with this problem [69].

Thus, for each predictive model, a prior step was carried out selecting the top-19 most important features that could help in the classification. To calculate the most important features to be used as input for the models, the Predictive Power Score (PPS) was used. This metric calculates a value between 0 (no predictive power) to 1 (perfect predictive power) that represents the relationship between the different attributes against the target [70,71]. This metric is widely used in time series, as it has the ability to point out how much a given variable says about another.

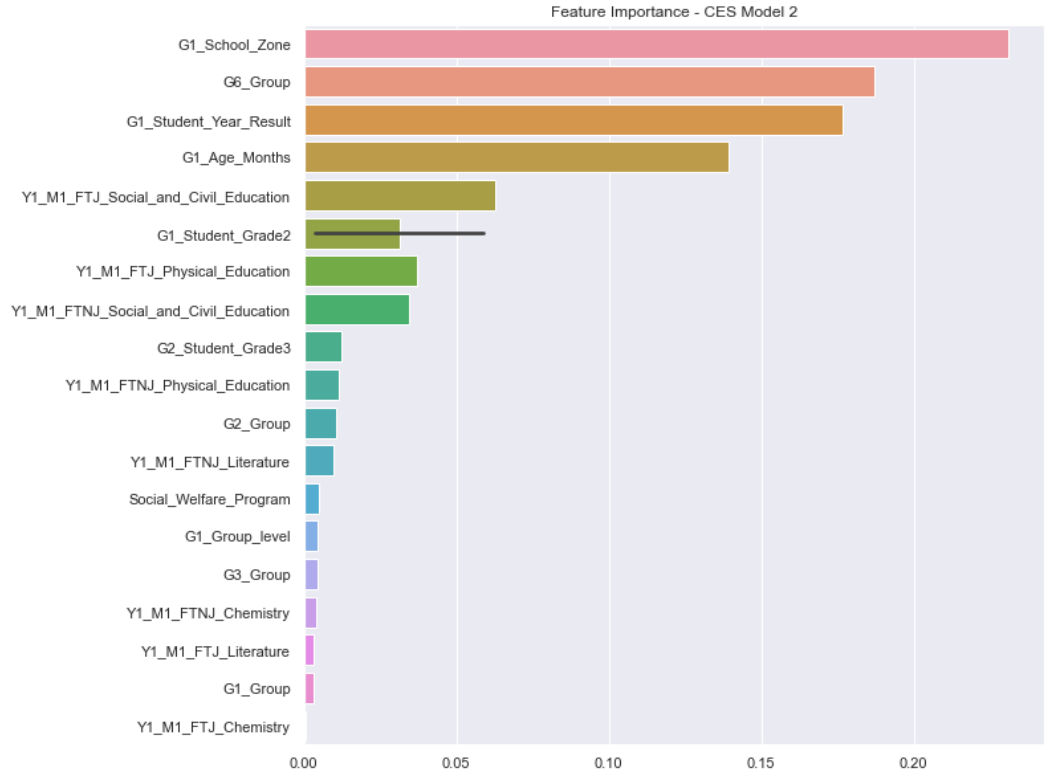


**Figure 9.** Feature Importance for M2G1-CES.[11]

Figure 9 presents the list of the most important features for M2G1-CES as an example. As it can bee seen in the figure, the most important features for this model combine information related to primary education together with information about the first meeting of secondary education. The two most important attributes for this model is related to the school zone (rural or urban) in the first year in primary, and to the students grouping based on their assessments in the sixth year of primary school.

From the analysis it is also possible to see that from the 10 most important features (attributes) the first five of them and the tenth are related to information from primary education. This demonstrates that educational problems in the studied context may have their origins in the first years of school. This finding corroborates previous findings of the literature [72–74]. Besides, this also confirms the finds of Nagy and Molontay [49], Hernández-Leal *et al.* [51] which highlighted the importance of using information about the performance of the students in the early years of education in order to predict their performance in the secondary education.

For instance, for this model, the attribute G1_School_Zone presented a PPS of 0.23 followed by the attribute G6_Group with PPS of 0.19. Moreover, the findings for the M2G1-CES model are very similar to the ones for the M2G1-UTU model in terms of the most important features. For the second year (grade 2), the assessment of the students in some of the subjects (disciplines) were among the top-10 most important features to be used as input by the models (M1G2-CES, M2G2-CES, M1G2-UTU and M2G2-UTU). This again confirmed the importance of using data from the primary education to predict students at-risk at secondary level.

[11] FTJ stands for justified absences. FTNJ stands for non-justified absences. For instance, Y1_M1_FTNJ_Literature means the number of non-justified absences in Literature until Meeting 1 during Year 1 (grade 1). Group stands for the classification of the student performance according to the quartile of the performances of all students at that grade

## 9. Conclusions

Learning analytics is a new research area that is gradually growing and consolidating itself. However, the main focus of the researches in this field is still towards higher education, with less attention to primary and secondary educational levels [13,44,45,75]. The present research specifically covers the adoption of LA in secondary education, at the same time seeks to assist Uruguay in the formation of institutional and governmental policies by early detecting students at-risk.

The present work proposed a methodology to predict students at-risk in secondary education at a national level. Together with the proposal, it was also possible to present the performances of the models running with real data collected from students and covering their school cycle from the first year of primary education to the second year of secondary education. A total of 8 models were generated and tested to avoid any bias, and 7 of them were approved to be adopted. Moreover, an API was developed and described so that these models could be deployed to the authorities responsible for running them. As the learning analytics process is cyclical, several manuals, reports and video training were also generated to facilitate the annual retraining of the models by the stakeholders of ANEP.

The data understanding stage allowed the establishment of an initial set of main variables that can be used in the process of generating early prediction models for students at risk in secondary level. Initial results suggest that primary school data, together with sociocultural student data, help to partially improve the performance of predictive models by approximately 4%. Moreover, Exploratory data analysis revealed sensitive issues that were hidden in the data, such as, for example, that the population of students who participated in some kind of social welfare program during primary school had fewer problems during secondary school than the population that did not participated in the social welfare programs. This situation was observed for both CES and UTU and may indicate that current social policies are in the right direction.

Throughout the process, several limitations were encountered for the advancement of the project. We can highlight some of them, such as the failure to obtain budget data, which could reveal new information about schools and the relationship between investment and results. Moreover, issues related to the crossing of data with the states and regions of the country and the Gross domestic product (GDP) were not explored in this project and should be considered in future improvements.

Future work should also focus on adding new functionalities to the developed API. Possible improvements could be the development of graphical visualization of the results, the analysis and cross-referencing of the data, and the automation of tasks related to preprocessing. Ideas of reports and dashboards for this context were already previously proposed by [15].

At the current stage of the project, it was possible to verify the efficiency of the predictive models in the task they are proposed to perform, at the same time one guarantee their fairness and explainability. However, it is still necessary to assess how the adoption and the interpretation of the predictive results will be effective to allow the governmental institutions to take actions to prevent dropouts and foster public policies. It is expected the process of adoption of the LA solution to be arduous, as it was already mentioned by previous works in the field [45,76]. It is important to highlight The work developed here is the first initiative towards the adoption of a Learning Analytics solution in secondary education at a national level in Latin America [44]

**Author Contributions:** E.M.Q.: experimental data analysis, algorithms development, experiments conduction, results description, manuscript writing;; M.M.F.: API development, manuscript writing; V.R.P: manuscript writing, editing, review, educational policies proposals, project coordination; C.C.: methodology definition, experiments setup, manuscript writing, editing, review, project coordination. The manuscript was written and approved to submit by all authors.

**Funding:** This work was supported by Udelar (University of the Republic) and Inter-American Development Bank through contract number RG-T3450-P004 (2020) - "Desarrollo de un modelo predictivo de riesgos de desvinculación educativa". Cristian Cechinel was partially supported by

CNPq (Brazilian National Council for Scientific and Technological Development) [DT-2 Productivity in Technological Development and Innovative Extension scholarship, proc.305731/2021-1]

## References


1. Santín, D.; Sicilia, G. Measuring the efficiency of public schools in Uruguay: main drivers and policy implications. *Latin American Economic Review* **2015**, *24*, 1–28.
2. Filgueira, F.; Gutiérrez, M.; Papadópulos, J. A perfect storm? Welfare, care, gender and generations in Uruguay. *Development and Change* **2011**, *42*, 1023–1048.
3. INEED. INEED. Informe sobre el estado de la educación en Uruguay 2015-2016. INEED, Montevideo, 2017. *Imprenta Blueprint* **2017**.
4. Ravela, P. A formative approach to national assessments: The case of Uruguay. *Prospects* **2005**, *35*, 21–43.
5. Pereda, T.F.C. Explicar/intervenir sobre la desafiliación educativa en la enseñanza media. *EL URUGUAY DESDE LA SOCIOLOGÍA VIII*, *165*.
6. Siemens, G.; Long, P. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review* **2011**, *46*, 30.
7. Hilliger, I.; Ortiz-Rojas, M.; Pesántez-Cabrera, P.; Scheihing, E.; Tsai, Y.S.; Muñoz-Merino, P.J.; Broos, T.; Whitelock-Wainwright, A.; Pérez-Sanagustín, M. Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. *The Internet and Higher Education* **2020**, *45*, 100726.
8. Baker, R.S.; Inventado, P.S. Educational data mining and learning analytics. In *Learning analytics*; Springer, 2014; pp. 61–75.
9. Campbell, J.P.; DeBlois, P.B.; Oblinger, D.G. Academic analytics: A new tool for a new era. *EDUCAUSE review* **2007**, *42*, 40.
10. Cano, C.M.v.A.; Romero, C.; Ventura, S. Predicting School Failure and Dropout by Using Data Mining Techniques Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data **2013**. doi:10.1109/RITA.2013.2244695.
11. Queiroga, E.M.; Enríquez, C.R.; Cechinel, C.; Casas, A.P.; Paragarino, V.R.; Bencke, L.R.; Ramos, V.F.C. Using Virtual Learning Environment Data for the Development of Institutional Educational Policies. *Applied Sciences* **2021**, *11*, 6811.
12. Clow, D. The learning analytics cycle: closing the loop effectively **2012**.
13. Kovanovic, V.; Mazziotti, C.; Lodge, J. Learning Analytics for Primary and Secondary Schools. *Journal of Learning Analytics* **2021**, *8*, 1–5.
14. Macarini, L.A.; dos Santos, H.L.; Cechinel, C.; Ochoa, X.; Rodés, V.; Casas, A.P.; Lucas, P.P.; Maya, R.; Alonso, G.E.; Díaz, P. Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay. *Interactive Learning Environments* **2019**, *0*, 1–25, [https://doi.org/10.1080/10494820.2019.1636082]. doi:10.1080/10494820.2019.1636082.
15. Macarini, B.; Antonio, L.; Cechinel, C.; Batista Machado, M.F.; Faria Culmant Ramos, V.; Munoz, R. Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Applied Sciences* **2019**, *9*, 5523.
16. Pombo, C.; Cabrol, M.; González Alarcón, N.; Roberto, S.Á. fAIr LAC: Responsible and Widespread Adoption of Artificial Intelligence in Latin America and the Caribbean **2020**. doi:10.18235/0002169.
17. Arias Ortiz, E.; Giambruno, C.; Muñoz Stuardo, G.; Pérez Alfaro, M. Camino hacia la inclusión educativa: 4 pasos para la construcción de sistemas de protección de trayectorias: Paso 1: Exclusión educativa en ALC:¿ cómo los sistemas de protección de trayectorias pueden ayudar? **2021**. doi:10.18235/0003455.
18. Bogliaccini, J.A.; Rodríguez, F. Education system institutions and educational inequalities in Uruguay. *Cepal Review* **2015**.
19. Bozkurt, A.; Jung, I.; Xiao, J.; Vladimirschi, V.; Schuwer, R.; Egorov, G.; Lambert, S.; Al-Freih, M.; Pete, J.; Olcott Jr, D.; et al. A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education* **2020**, *15*, 1–126.
20. Silveira, I.F.; Casali, A.; Bezeira, A.V.M.; Sprock, A.S.; Collazos, C.A.; Cechinel, C.; Muñoz-Arteaga, J.; Maldonado-Mahauad, J.; Chacón-Rivas, M.; Motz, R.; et al. Iguales en las diferencias: iniciativas de investigación transnacionales sobre Informática Educativa en Latinoamérica en el periodo 2010-2020. *Revista Brasileira de Informática na Educação* **2021**, *29*, 1060–1090.

21. Bucheli, M.; Lustig, N.; Rossi, M.; Amábile, F. Social spending, taxes, and income redistribution in Uruguay. *Public Finance Review* **2014**, *42*, 413–433.
22. Dirección General de Planeamiento. Estadísticas Básicas 2018 de la Universidad de la República. Technical report, Universidad de la República, 2018.
23. Rivoir, A.L. Innovación para la inclusión digital. El Plan Ceibal en Uruguay **2009**.
24. RIVERA VARGAS, P.; Cobo, C. Plan Ceibal en Uruguay: una política pública que conecta inclusión e innovación. *Políticas Públicas para le Equidad Social. Santiago de Chile: Colección Políticas Públicas* **2018**.
25. Ferguson, R. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning* **2012**, *4*, 304–317.
26. 1st International Conference on Learning Analytics and Knowledge 2011, 2011.
27. Siemens, G. Learning analytics: The emergence of a discipline. *American Behavioral Scientist* **2013**, *57*, 1380–1400.
28. Chatti, M.A.; Dyckhoff, A.L.; Schroeder, U.; Thüs, H. A reference model for learning analytics. *International Journal of Technology Enhanced Learning* **2013**, *4*, 318–331.
29. Siemens, G.; Baker, R.S.d. Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the Proceedings of the 2nd international conference on learning analytics and knowledge, 2012, pp. 252–254.
30. Phauk, S.; Okazaki, T. Integration of Educational Data Mining Models to a Web-Based Support System for Predicting High School Student Performance. *International Journal of Computer and Information Engineering* **2021**, *15*, 131–144.
31. Cortez, P.; Silva, A.M.G. Using data mining to predict secondary school student performance **2008**.
32. Detoni, D.; Cechinel, C.; Matsumura Araújo, R. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. *Revista Brasileira de Informática na Educação* **2015**, *23*.
33. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Mousa Fardoun, H.; Ventura, S. Early dropout prediction using data mining: a case study with high school students. *Expert Systems* **2016**, *33*, 107–124.
34. Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, 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* **2020**, *10*, 3998.
35. Zohair, L.M.A. Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education* **2019**, *16*, 27.
36. Aldowah, H.; Al-Samarraie, H.; Fauzy, W.M. "Educational data mining and learning analytics for 21st century higher education: A review and synthesis". *Telematics and Informatics* **2019**, *37*, 13–49. doi:10.1016/j.tele.2019.01.007.
37. Saqr, M.; López-Pernas, S. The Dire Cost of Early Disengagement: A Four-Year Learning Analytics Study over a Full Program. In Proceedings of the European Conference on Technology Enhanced Learning. Springer, 2021, pp. 122–136.
38. Queiroga, E.; Cechinel, C.; Araújo, R.; da Costa Bretanha, G. Generating models to predict at-risk students in technical e-learning courses. In Proceedings of the Learning Objects and Technology (LACLO), Latin American Conference on. IEEE, 2016, pp. 1–8.
39. Fernandes, E.; Holanda, M.; Victorino, M.; Borges, V.; Carvalho, R.; Van Erven, G. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research* **2019**, *94*, 335–343.
40. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education* **2009**, *53*, 950–965. doi:10.1016/j.compedu.2009.05.010.
41. Daud, A.; Aljohani, N.R.; Abbasi, R.A.; Lytras, M.D.; Abbas, F.; Alowibdi, J.S. Predicting student performance using advanced learning analytics. In Proceedings of the Proceedings of the 26th international conference on world wide web companion, 2017, pp. 415–421.
42. Di Mitri, D.; Scheffel, M.; Drachsler, H.; Börner, D.; Ternier, S.; Specht, M. Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data. In Proceedings of the Proceedings of the seventh international learning analytics & knowledge conference, 2017, pp. 188–197.
43. Camacho, V.L.; de la Guía, E.; Olivares, T.; Flores, M.J.; Orozco-Barbosa, L. Data Capture and Multimodal Learning Analytics Focused on Engagement With a New Wearable IoT Approach. *IEEE Transactions on Learning Technologies* **2020**, *13*, 704–717.

44. Cechinel, C.; Ochoa, X.; Lemos dos Santos, H.; Carvalho Nunes, J.B.; Rodés, V.; Marques Queiroga, E. Mapping learning analytics initiatives in latin america. *British Journal of Educational Technology* **2020**, *51*, 892–914.
45. Bruno, E.; Alexandre, B.; Ferreira Mello, R.; Falcão, T.P.; Vesin, B.; Gašević, D. Applications of learning analytics in high schools: a Systematic Literature review. *Frontiers in Artificial Intelligence* **2021**, p. 132.
46. Michalski, R.S.; Carbonell, J.G.; Mitchell, T.M. *Machine learning: An artificial intelligence approach*; Springer Science & Business Media, 2013.
47. Sclater, N.; Peasgood, A.; Mullan, J. Learning analytics in higher education. *London: Jisc. Accessed February* **2016**, *8*, 176.
48. Frostad, P.; Pijl, S.J.; Mjaavatn, P.E. Losing all interest in school: Social participation as a predictor of the intention to leave upper secondary school early. *Scandinavian journal of educational research* **2015**, *59*, 110–122.
49. Nagy, M.; Molontay, R. Predicting dropout in higher education based on secondary school performance. In Proceedings of the 2018 IEEE 22nd international conference on intelligent engineering systems (INES). IEEE, 2018, pp. 000389–000394.
50. Lehrl, S.; Ebert, S.; Blaurock, S.; Rossbach, H.G.; Weinert, S. Long-term and domain-specific relations between the early years home learning environment and students' academic outcomes in secondary school. *School Effectiveness and School Improvement* **2020**, *31*, 102–124.
51. Hernández-Leal, E.; Duque-Méndez, N.D.; Cechinel, C. Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. *Heliyon* **2021**, *7*, e08017.
52. Gardner, J.; Brooks, C.; Baker, R. Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the Proceedings of the 9th international conference on learning analytics & knowledge, 2019, pp. 225–234.
53. Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.W.; Wang, W.Y. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* **2019**.
54. Cao, Y.T.; Daumé III, H. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics* **2021**, *47*, 615–661.
55. Leavy, S. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the Proceedings of the 1st international workshop on gender equality in software engineering, 2018, pp. 14–16.
56. Wei, W.W. Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*; 2006.
57. Diggle, P.; Al-Wasel, I. Time series. **1990**.
58. González, F.; Ortiz, T.; Sánchez, R. IA Responsable **2020**.
59. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing* **2019**, *12*, 86.
60. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, pp. 1189–1232.
61. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
62. Bergstra, J.S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In Proceedings of the Advances in neural information processing systems, 2011, pp. 2546–2554.
63. Olson, R.S.; Urbanowicz, R.J.; Andrews, P.C.; Lavender, N.A.; Kidd, L.C.; Moore, J.H., Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I; Springer International Publishing, 2016; chapter Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137. doi:10.1007/978-3-319-31204-0_9.
64. 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* **2017**, *18*, 1–5.
65. González, F.; Ortiz, T.; Ávalos, R.S. *IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial*; Inter-American Development Bank, 2020. doi:10.18235/0002876.
66. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–35.

67. Gašević, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education* **2016**, *28*, 68–84.
68. Herodotou, C.; Rienties, B.; Verdin, B.; Boroowa, A. Predictive learning analytics 'at scale': Towards guidelines to successful implementation in Higher Education based on the case of the Open University UK. *Journal of Learning Analytics* **2019**, pp. In–Press.
69. Athey, S.; Tibshirani, J.; Wager, S. Generalized random forests. *The Annals of Statistics* **2019**, *47*, 1148–1178.
70. Mai-Nguyen, A.V.; Tran, V.L.; Dao, M.S.; Zettsu, K. Leverage the Predictive Power Score of Lifelog Data's Attributes to Predict the Expected Athlete Performance. In Proceedings of the CLEF (Working Notes), 2020.
71. Oksanen, T.; Tiainen, M.; Skrifvars, M.B.; Varpula, T.; Kuitunen, A.; Castrén, M.; Pettilä, V. Predictive power of serum NSE and OHCA score regarding 6-month neurologic outcome after out-of-hospital ventricular fibrillation and therapeutic hypothermia. *Resuscitation* **2009**, *80*, 165–170.
72. Zeichner, K. Rethinking the connections between campus courses and field experiences in college-and university-based teacher education. *Journal of teacher education* **2010**, *61*, 89–99.
73. Fall, A.M.; Roberts, G. High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of adolescence* **2012**, *35*, 787–798.
74. Hosokawa, R.; Katsura, T. Effect of socioeconomic status on behavioral problems from preschool to early elementary school–A Japanese longitudinal study. *PLoS one* **2018**, *13*, e0197961.
75. Queiroga, E.; Cechinel, C.; Aguiar, M. Uma abordagem para predição de estudantes em risco utilizando algoritmos genéticos e mineração de dados: um estudo de caso com dados de um curso técnico a distância. In Proceedings of the Anais dos Workshops do Congresso Brasileiro de Informática na Educação, 2019, Vol. 8, p. 119.
76. Brown, M. Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data. *Teaching in Higher Education* **2020**, *25*, 384–400.

# APÊNDICE B – Using Virtual Learning Environment Data for the Development of Institutional Educational Policies

applied sciences

MDPI

*Article*

# Using Virtual Learning Environment Data for the Development of Institutional Educational Policies

**Emanuel Marques Queiroga [1,*], Carolina Rodríguez Enríquez [2], Cristian Cechinel [3], Alén Perez Casas [4], Virgínia Rodés Paragarino [5], Luciana Regina Bencke [6] and Vinicius Faria Culmant Ramos [3]**

1 Instituto Federal do Rio Grande do Sul, IFSul, Pelotas 96015560, Brazil
2 Facultad de Enfermería, Universidad de la República, Udelar, Montevideo 11600, Uruguay; carolinacabocla@gmail.com
3 Centro de Ciências, Tecnologias e Saúde (CTS), Universidade Federal de Santa Catarina, UFSC, Araranguá 88906072, Brazil; contato@cristiancechinel.pro.br (C.C.); email@viniciusramos.pro.br (V.F.C.R.)
4 Facultad de Información y Comunicación, Universidad de la Republica, Udelar, Montevideo 11200, Uruguay; alen.perez@fic.edu.uy
5 Comisión Sectorial de Enseñanza, Universidad de la República, Udelar, Montevideo 11200, Uruguay; virginia.rodes@gmail.com
6 Instituto de Informática, Universidade Federal do Rio Grande do Sul, UFGRS, Porto Alegre 91501970, Brazil; luciana.bencke@gmail.com
* Correspondence: emanuelmqueiroga@gmail.com

**Featured Application: Combining different data sources has high power to predict students at-risk of failure and to identify behavior patterns to develop institutional polices based on evidence.**

**Abstract:** This paper describes the application of Data Science and Educational Data Mining techniques to data from 4529 students, seeking to identify behavior patterns and generate early predictive models at the Universidad de la República del Uruguay. The paper describes the use of data from different sources (a Virtual Learning Environment, survey, and academic system) to generate predictive models and discover the most impactful variables linked to student success. The combination of different data sources demonstrated a high predictive power, achieving prediction rates with outstanding discrimination at the fourth week of a course. The analysis showed that students with more interactions inside the Virtual Learning Environment tended to have more success in their disciplines. The results also revealed some relevant attributes that influenced the students' success, such as the number of subjects the student was enrolled in, the students' mother's education, and the students' neighborhood. From the results emerged some institutional policies, such as the allocation of computational resources for the Virtual Learning Environment infrastructure and its widespread use, the development of tools for following the trajectory of students, and the detection of students at-risk of failure. The construction of an interdisciplinary exchange bridge between sociology, education, and data science is also a significant contribution to the academic community that may help in constructing university educational policies.

**Keywords:** classification; educational strategies; higher education; learning analytics



**Citation:** Queiroga, E.M.; Enríquez, C.R.; Cechinel, C.; Casas, A.P.; Paragarino, V.R.; Bencke, L.R.; Ramos, V.F.C. Using Virtual Learning Environment Data for the Development of Institutional Educational Policies. *Appl. Sci.* **2021**, *11*, 6811. https://doi.org/10.3390/app11156811

Academic Editor: Andrea Prati

Received: 8 June 2021
Accepted: 19 July 2021
Published: 24 July 2021



## 1. Introduction

Universities have been concerned with using the extensive data produced by their educational systems in aiming to improve the overall performance of students [1–6]. According to [7], the scope of contemporary higher education is vast, and concerns about the performance of higher education systems are widespread. Among several challenges that have been faced by universities, one can mention low completion rates, which are commonly associated with inefficiencies in higher education, even though they also depend on other factors, such as the student profiles and their paths to completion [5,7,8].

Data mining techniques can be used to overcome some of these challenges. Two specific areas are used to refer to the application of data mining in educational settings: Educational Data Mining (EDM) and Learning Analytics (LA) [9,10]. EDM is an interdisciplinary research field that deals with the development of methods to explore data sourced from the educational context [11,12]. LA seeks to measure, collect, analyze, and report data about students and their contexts to understand and optimize their learning and learning environment [13]. Student and teacher interactions within Virtual Learning Environments (VLEs) provide data that feed the research in these areas, thus, enabling the discovery of new knowledge [14].

Learning Management Systems (LMSs) and student information systems containing socio-demographic and student enrollment data can be considered the technological foundation for higher education institutions [15]. Modern educational systems use VLEs to support classroom activities, even in face-to-face courses. In these environments, it is possible to share materials, perform tasks, and interact with other users with the ultimate goal of generating and acquiring knowledge, both individually and collectively [14,16,17]. Modular Object-Oriented Dynamic Learning Environment (Moodle) is one of the most widely used VLEs worldwide. In Uruguay, there are 413 installation sites [18].

Data mining in higher education is mainly used for techniques, such as classification, clustering, and association rules as well as to predict, group, model, and monitor various learning activities [5,9,19]. Current studies on LA vary in several dimensions, covering, for instance, the techniques employed (data mining, visualization, social network analysis, and statistics), the source of the data (LMSs, surveys, and sensors), the stakeholders involved (students, professors, and administrators), and the educational level to which the systems/experiments are directed [20].

This work aims to unveil educational patterns of student interactions with the VLE in higher education courses that use Moodle as a complementary tool for teaching and learning processes. Hence, a series of data mining experiments are applied to the data from the VLE and also to data from other sources, such as surveys and academic systems. The experiments intend to better understand the VLE's role in helping students' education inside the studied courses and to discover educational patterns and knowledge that can further help in planning future actions and policies inside the institution. For the present work, we propose the following research questions (RQ):

- RQ1: Is the use of VLE associated with student approval?
- RQ2: Which features from the different datasets (VLE, census, and academic system) are the most important for the early prediction of student performance?
- RQ3: Which learning patterns can educational data mining help to unveil in the studied courses?

In this work, data mining was used as a tool to unveil educational knowledge and possible existing patterns related to the final status of the students. Even though we report quantitative results about predictive models, our main goal is to uncover these patterns to better understand the role that VLEs and other variables have in students' performance so that future educational policies can be built based on empirical findings. The process followed here can also be defined as Knowledge Discovery in Databases (KDD).

The context of the study is the University of the Republic (Udelar), the main institution of higher education in Uruguay. The remainder of this work is organized as follows: Section 2 presents related works, and Section 3.2 describes the context of the present study. Section 3 depicts the methodology followed in the paper (data collection, model generation, and evaluation). Section 4 presents the results, and Section 5 discusses the research questions based on the results. Section 6 presents possibilities of institutional polices based on the evidence, and Section 7 indicates our conclusions, limitations, and future research.

## 2. Related Work

This section presents an overview of the research problem topic. Also in this section, Table 1 presents a summary of the aboarded studies.

Leitner et al. [21] presented a practical tool that can be used to identify the risks and challenges that arise when implementing LA and explained how to approach the same. The authors propose a framework with seven main categories for LA initiatives: Purpose and Gain, Representation and Actions, Data, IT Infrastructure, Development and Operation, Privacy, and Ethics. They remarked that the order of implementation depends on each institution. The Data dimension encompasses the application of the advantages of modern technology and the various data sources available, looking for the right analysis to improve the quality of learning and teaching, as well as to enhance the chances of student success.

In a global context, the prediction of performance and dropout is concentrated at the university level, with about 70% of the research focused on this purpose [22]. This trend is the same in Latin America [23]; however, according to [1], Latin American universities still have considerably lower adoption rates compared to institutions in other regions. Thus, Latin American educational institutions can use LA to combat disparities in teaching quality, performance problems, and high dropout rates.

The potential of using predictive methods in education has already been demonstrated by numerous works in the literature [4,14,24–32].

Our work focuses on the data dimension, as it is essential to analyze practical case studies and understand which are the key metrics and the processes they are applying. As there is already another work summarizing the important findings up to 2017 (i.e., [9]), we concentrated our exploratory search on papers after 2017. The systematic review from [9] covered the most relevant studies related to four main dimensions: computer-supported learning analytics, computer-supported predictive analytics, computer-supported behavioral analytics, and computer-supported visualization analytics from 2000 to 2017.

The authors identified twelve relevant EDM/LA techniques that researchers normally combine: classification (26.25%), clustering (21.25%), visual data mining (15%), statistics (14.25%), association rule mining (14%), regression (10.25%), sequential pattern mining (6.50%), text mining (4.75%), correlation mining (3%), outlier detection (2.25%), causal mining (1%), and density estimation (1%). Searching EDM/LA works after 2017, we found research applying different techniques and using different sources of data that we mention here.

A practical application of early prediction is proposed by [29]. The authors implemented an alert system to predict performance in some classes at the university. The research demonstrated that the use of predictive methods in education allowed an increase of up to 15% on te students' performance compared to those in classes that did not use the models.

Gutiérrez et al. [27] proposed the use of the Learning Analytics Dashboard for Advisers (LADA) as a tool to support the learning process and the students' final success. This tool seeks to assist educational counselors in the decision-making process through comparative and predictive analyses of the student data. The use of the predictive methods of this tool showed significant results, especially in complex cases, in student success.

Foster and Siddle [26] investigated the effectiveness of LA in identifying at-risk students in higher education. To this end, the authors compared the low-engagement alerts of an LA tool with the results of students at the end of the first year of graduation. In addition, different methodologies for generating alerts have been compared, such as the use of demographic data and only VLE participation data. The tests demonstrated that the VLE-data approach was more efficient at generating alerts than using socio-demographic data. In the end, the authors demonstrated that students who had performance problems or dropped out at the end of the first year received an average of 43% more alerts on the tool.

The problem of college-going students taking longer to graduate than their parental generations was tackled by [33]. The authors presented a prediction model to identify students at-risk of failing courses that they plan to take in the next term (or future). Different

models are learned from different courses. To predict a student's grades in the next courses, his grades from prior courses are fed into corresponding models. To capture the sequential characteristics of students' grades in prior courses, they modeled the learning behavior and performance using recurrent neural networks with long short term memory (LSTM).

In Latin America, a number of initiatives proposed approaches to the use of Educational Data Mining and Learning Analytics at the higher educational level [23]. In this context, [4] presented a proposal that aimed at early prediction of university student retention at Chile. For that, the authors applied a number of different data mining algorithms (decision trees, k-nearest neighbors, logistic regression, naive Bayes, random forest, and support vector machines) over students socioeconomic information and previous achievements in their courses. The results demonstrated an accuracy on the classification higher than 80% in all tested scenarios.

**Table 1.** Summary of related works.

| Work | Goal | Technique | Algorithm | Edu. Level | Features Used |
|---|---|---|---|---|---|
| [29] | To predict students at-risk of fail | Classification | Logistic regression | Higher | Student factors (IMD area, price area, and disability), Previous studies (highest qualification on entry), Student course (total credits studying in a year and late registration), Previous progress at the university (best previous score and number of fails) |
| [27] | To support academic advising scenarios | Multilevel clustering | Fuzzy C-means | Higher | Grades and the number of courses students took during the semester |
| [26] | To predict students at-risk of fail | Not-mentioned | Not-mentioned | Higher | Demographic data versus only VLE participation data |
| [4] | To predict student retention | Classification | Decision trees, k-nearest neighbors, logistic regression, naive Bayes, random forest, and SVM | Higher | Educational score and the community poverty index and university grades. |
| [30] | To predict students at-risk of fail | Statistical Analysis | Correlation and regression analysis | Higher | Click stream data, self-reported measures, and course performance. |
| [24] | To predict both marginal and at-risk students of fail | Classification | Training vector-based SVM | Higher | Demographic data and interaction with a virtual learning environment. |
| [25] | To select best features to improve predicting students performance | Feature selection to improve supervised learning classifiers | Deep learning with LSTM | Higher | Metrics from navigation events that are combined in the LSTM network. |
| [28] | To predict students at-risk of dropout | Classification | Random forest and boosted decision | School | Attendance and course performance. |
| [32] | To identify learners personality | Classification | Naive bayes | Higher | Participation in forums and chats, access to supplementary course materials, delay in assignment delivering, score, accomplishment of assignments, time solving of quizzes, and number of entrances in the system. |
| [33] | To predict students at-risk of fail | Classification | LSTM and RNN | Higher | Previous grades. |

The learning process in which students are responsible for defining their goals and constantly auto-regulating their objectives towards some content or course is named Self-Regulated Learning (SLR) [34]. Li et al. [30] evaluated SRL in face-to-face courses that are supported by online activities/courses to demonstrate the extent to which LMS interactions may be used to better understand how students manage their time and regulate their efforts. By doing so, the authors aim to improve their performance on the identification of at-risk students.

They collected questionnaire data (pre- and post-course) from freshmen university students enrolled in a 10-week course. The questions were based on the following: the Motivated Strategies for Learning Questionnaire (MSLQ), the students' interactions with the VLE, and socio-demographic data. Their findings showed a moderate positive correlation between the VLE clicks and students' SRL, as well as between VLE clicks and the students' final performance. Moreover, the authors reported that the combination of demographic attributes with SRL variables significantly impacted the model's ability to predict at-risk students.

According to [25], a significant challenge faced when building predictive models of student learning behaviors is to use handcrafted features that are effective for the prediction task at hand. The authors, then, adopted an unsupervised learning approach to learn a compact representation of the raw features. They sought to capture the underlying learning patterns in the content domain and the temporal nature of the click-stream data. The authors used Deep Learning the training and a modified auto-encoder combined with the LSTM network to obtain a fixed-length embedding for each input sequence.

The selected features used in supervised learning models achieved superior results. Identifying at-risk students is the main goal of [28]. Dropout reasons include not only poor performance but also other events, such as violation of school rules, illness, etc. The authors addressed the class imbalance problem in the binary classification (dropout corresponds to 1% of the labeled dataset) through oversampling techniques.

They trained the embedded methods of random forest and boosted decision trees using the big data samples of the 165.715 high school students. The 15 features used referred to attendance (for example, unauthorized early leave in the first four weeks), behavior (number of volunteer activities), and course performance (normalized ranking on Math). A ROC and PR curve analysis was presented, showing that the boosted decision tree achieved the best performance.

## 3. Materials and Methods

This section presents an overview of the research methodology and the general context of the case study.

### 3.1. Overview

In Data Science, it is essential to define the project flow steps and the methodology to be followed. The method used in this work is the Cross-Industry Standard Process for Data Mining (CRISP-DM) [35] with minor adaptations to the application for the context of this research. Figure 1 shows the flow of the methodology model used and Figure 2 shows the proposed solution to this project.

The adapted CRISP-DM process and its six steps were applied and are presented in the sections of this paper as follows: context understanding is presented in Sections 1 and 3.2; data understanding is presented in Section 3.4; data preparation consists of the feature engineering process and is detailed in Section 3.5; the generation of models (modeling) is an iterative step that occurs in conjunction with data preparation, and this is shown in Section 3.6; evaluation of the results and its discussion are presented in Sections 4–6; and, at the end, the conclusions are shown in Section 7.

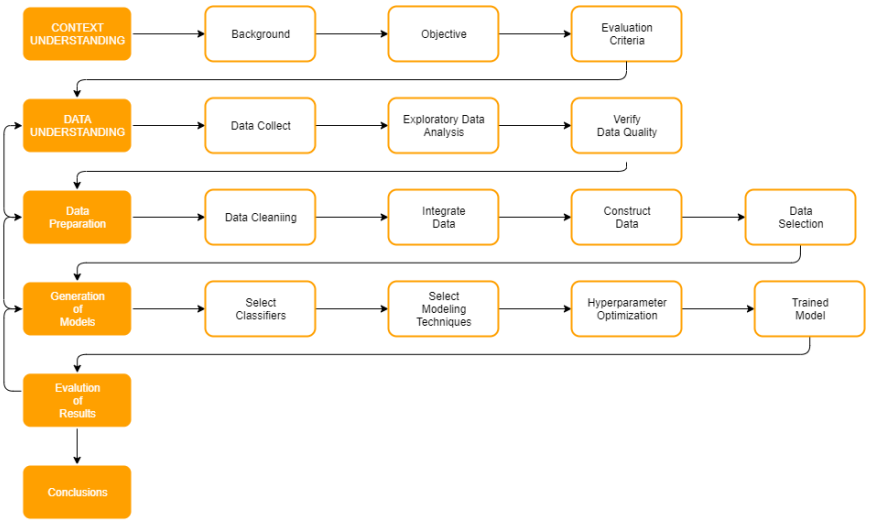


**Figure 1.** CRISP-DM with adaptation.

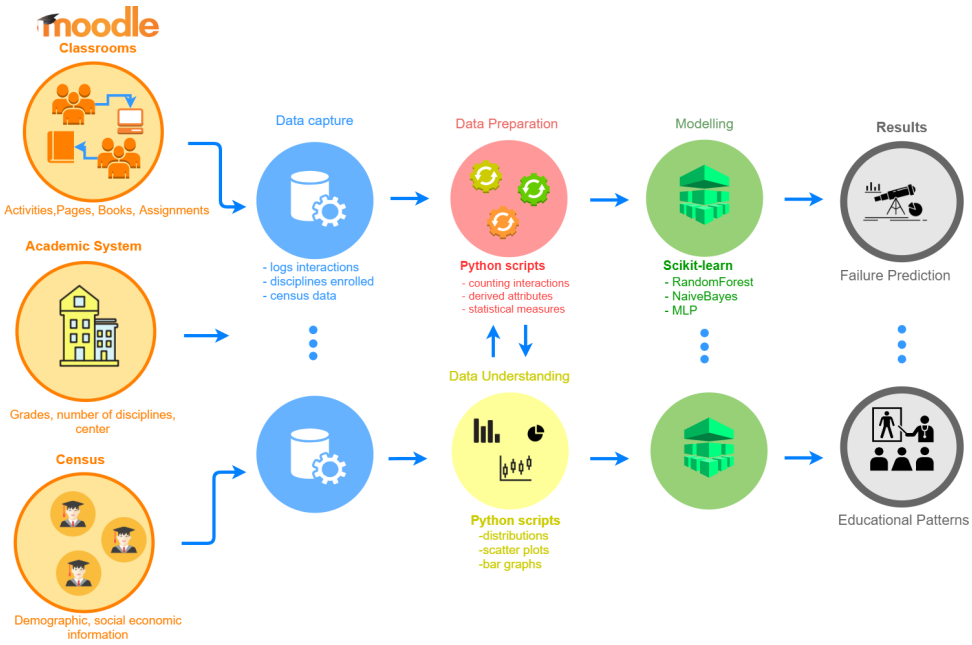


**Figure 2.** The proposed solution.

### *3.2. Contextualization: Case Study of Udelar*

In Uruguay, Udelar is the main institution of higher education, concentrating 75% of the students (public and private), 90% of the university system, and 99.5% of the public universities. It has a policy of free and unrestricted admission, with no other condition than the completion of high school. In 2020, the Udelar had 100 undergraduate courses and a few more than 200 graduate courses. In 2018, the university had more than 135 thousand undergraduate students and more than 10 thousand graduate students [36].

#### The Continuous Survey of Udelar's Students

To better understand its students, Udelar developed a set of statistical survey mechanisms to generate information about their characteristics and distribution, called "FormA-Students". The FormA-Students is a longitudinal survey that must be responded to annually by all students. The survey covers questions in the following dimensions: (a) sociodemographic, (b) pre-university education, (c) work, (d) other university and/or higher education studies outside Udelar, (e) languages, (f) academic mobility, and (g) scholarships.

In addition to these dimensions, this research also uses data of their activity and qualifications recorded in the Bedelias System, the administrative management system that collects all the official records of the students' academic career, the subjects taken and completed, the approvals and failures, and the grades received.

The present work analyzed data from the second-year students enrolled in courses from three different faculties, in the year 2017, as follows: (1) Faculty of Information and Communication (FIC), (2) Faculty of Nursing (FEnf), and (3) Faculty of Sciences (FCien). These faculties have similar number of students and represent the three macro areas in which the Udelar are organized.

### 3.3. Computational Settings

The computer used to process the data used the Operating System Ubuntu 18.04 and had an Intel i5 4th generation processor with 8 GB RAM. The environment was created using an Anaconda distribution, and the scripts were developed in Python 3.8 with scikit-learn, Pandas, and Numpy packages. The total runtime for training and testing the models was roughly 24 h. For each dataset combination, the model generation took from 2 to 4 h.

### 3.4. Data Understanding

Data from students enrolled in three bachelor programs from three different faculties of Udelar were collected. The programs are Biology (BIO), Communication (COM), and Nursing (NUR). Table 2 shows the number of subjects used in each program, the total number of interactions inside the VLE for each subject, the total number of students enrolled in, and the following: students that had success without retaking exams, students that had success after the final exams, and students that failed.

**Table 2.** Description of the student population and final status.

| College | Total of Interactions | Subjects | Students | Success | | Fail |
|---|---|---|---|---|---|---|
| | | | | Final Exam | Retake Exam | |
| BIO | 23,606 | 3 | 59 | 0 | 43 | 16 |
| COM | 150,623 | 5 | 1361 | 820 | 318 | 223 |
| NUR | 955,163 | 6 | 3109 | 914 | 901 | 1294 |
| Total | 1,129,392 | 14 | 4529 | 3089 | 1262 | 1533 |

It is important to highlight that it is not mandatory for the student to attend the classes to take or retake the final exams. This particularity affects the way students use VLE, especially during the first year when a large number of students drop out of university as this public university does not have entrance exams. This is the main reason for choosing data from second-year students as it tends to be stable in terms of dropout. In this sense, we believe that we have a clearer picture of the use of the VLE by the students, which was intended to keep them enrolled in the courses.

Two different output variables (targets) were defined for our study: the prediction of success in the course (students who passed without the need of exams) and the prediction of success in the final exams. For the first target, the models predict whether a given student will be approved directly or if they will need to take exams. For the second target, the models predict whether a given student will pass or fail after taking the exams. Together with the students' interactions inside Moodle, we also used data from the university's academic system and the FormA-Students survey database.

Students' interactions within VLE in its raw state were collected. These data were separated by students, day of interaction, and type of content. We collected, from the academic system, the subjects enrolled in by each student, the academic performance in the subjects, and the number of previous failures in each subject. The third data's source was the continuous survey called FormA-Students. This survey is completed by students annually,

and it collects 111 attributes distributed in sections referring to socio-demographic and socio-economic background, pre-university and further university studies, employment status, language proficiency, motivation and expectations about career, academic mobility, and scholarships.

According to the dean of Udelar [37], the survey data can enable the institution to think about itself in the long term and in strategies that require the prediction of the state of affairs of the different actors to achieve specific objectives. For example, it has the education and occupational category of the father and mother, marital status, family income, ethnic self-perception, disabilities, employment status, occupation classification, scholarship receptions, place of birth, and the place where they live and with whom.

The exploratory data analysis step sought to visualize the different datasets before integration to identify database sizes, become familiar with the data, and gain insights for the transformation of target features, as well as identify visible behavioral patterns.

Figure 3 shows the distribution of interactions in VLE by age. A possible observation is that the older the student, the lower their use of the VLE. This may represent an acceptance trend where younger students tend to adhere more to the use of Moodle. Still the right sidebar of the graph shows the distribution of students by age, and the top bar shows the distribution by interactions. As seen, the highest concentration of interactions was found in students between 20 and 25 years of age.

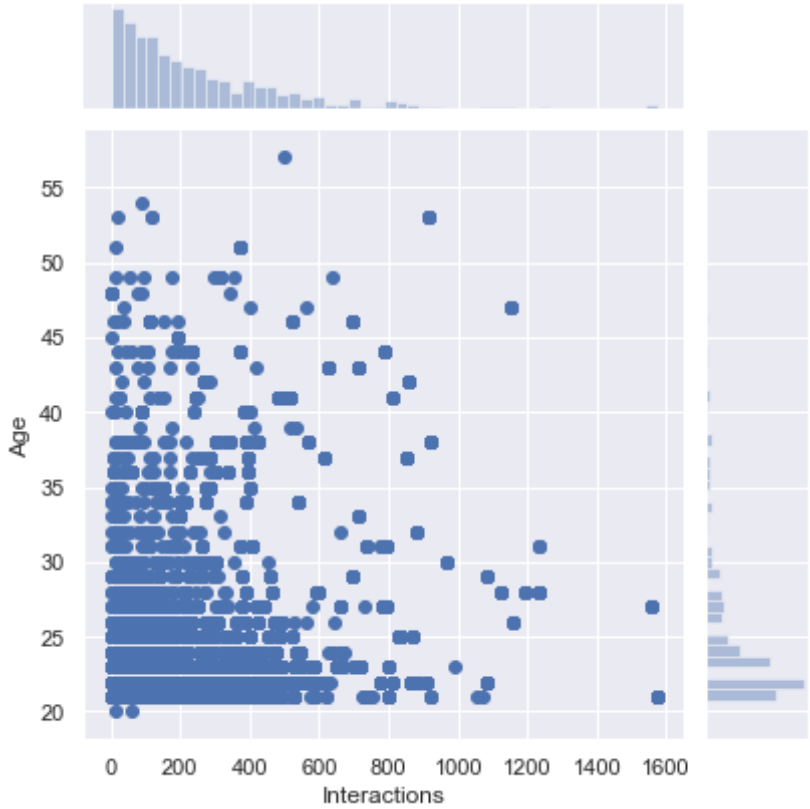


**Figure 3.** Dispersion between interactions and age.

Another important analysis of the Figure 3 is that a significant part of the dispersion was located between 0 and 200 interactions. In this range, 52 students were identified who had 0 interactions with the VLE during the courses, of which 16 passed the course (without exam), 23 passed the exam, and 13 failed. In addition to that, only two students took the course for the first time, and both failed.

Figure 4 shows the difference of interactions between students who had success versus students who failed the subjects. In the upper part of the figure, interactions are presented during the 16 weeks of the subjects, where notably the students who had success

demonstrate a higher engagement in VLE compared to those who failed. The bottom part of the figure shows the total number of interactions after the end of the semester (after the 16 weeks and the final exam) and before students retake the exams. It must be noticed that the failing students had higher engagement compared to the course progress but less than the successful students.

Figure 5 shows the distribution of interactions during the weeks of the course. The interactions grew until the partial exams (in weeks 8 and 14/16). This movement is an indication that the closer the exams/tests are in a given subject, the higher the students access to the VLE to consult the materials.

Figure 6 shows the total number of interactions per subject (upper) and the average number of interactions for each of the analyzed subjects (bellow). It is possible to analyze that, even within a program, the use of VLE was considerably different between subjects.

Analyzing the VLE subjects' didactic design, it was possible to characterize them as mainly organized as repositories of resources to support face-to-face classes, where professors upload materials, such as text, images, and videos, and provide online assessments and self-assessments. Forums are used mainly as a place for coordination and information dissemination rather than for the discussion of content-related issues.

The two main uses of a quiz are as follows: first, as a form of assessment evaluation of learning instruments, generally mandatory, by a single attempt, for all active students and carried out on a pre-established date; and second, as an interactive activity oriented to education and training over a long period and allowing multiple attempts. Rodés et al. [38] defined a typology of didactic designs according to the type of resources and activities supported by VLE. The courses analyzed here fall mainly under the repository and self-assessment types.

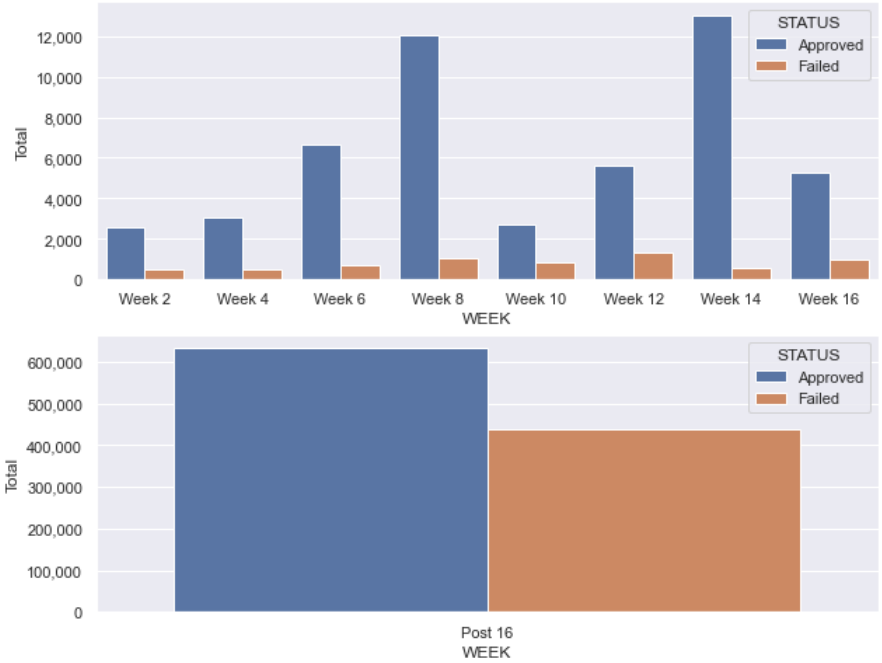


**Figure 4.** Interactions per weeks approval X failed.

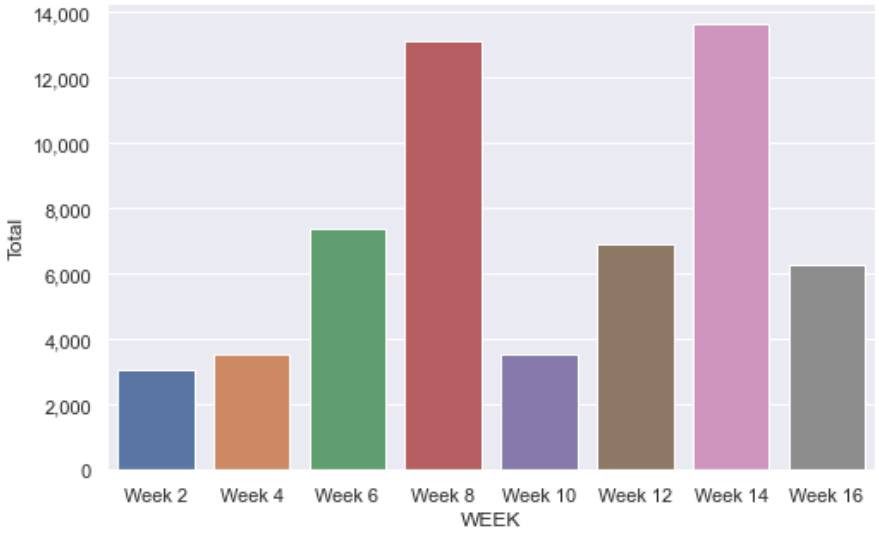


**Figure 5.** Interactions per weeks.

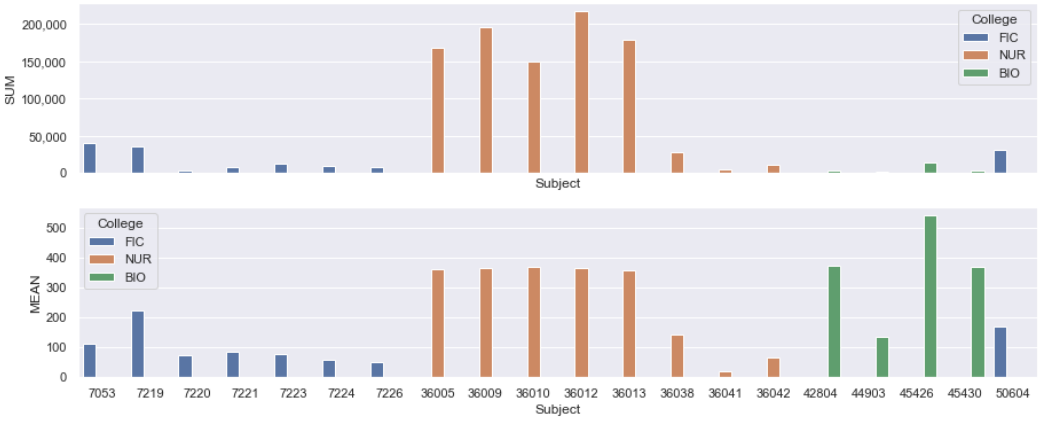


**Figure 6.** Sum and mean interactions by subject.

Figure 7 displays the frequencies of the distribution of total interactions by programs and the students' final status. In FIC and NUR, both categories have their peak of interaction near zero and do not seem to present a different distribution. On the other hand, BIO presents a different distribution of interactions between the categories, with the peak of interactions for the success category near 500 and for the failed category near 300.

To evaluate whether the VLE's students' interactions were associated with their final status in the subjects, we performed a statistical analysis. First, we used the Shapiro–Wilk test to verify whether interactions from both groups (success and failure) of each course followed a normal distribution. For the groups that follow a normal distribution, we performed a *t*-test and for the others, we applied the Mann–Whitney non-parametric test. The goal was to check whether the means/medians (depending on the test) of the groups present statistically significant differences. This analysis was performed for three different periods of the semester: week 4, week 8, and week 16 (all weeks). The results are shown in Table 3.

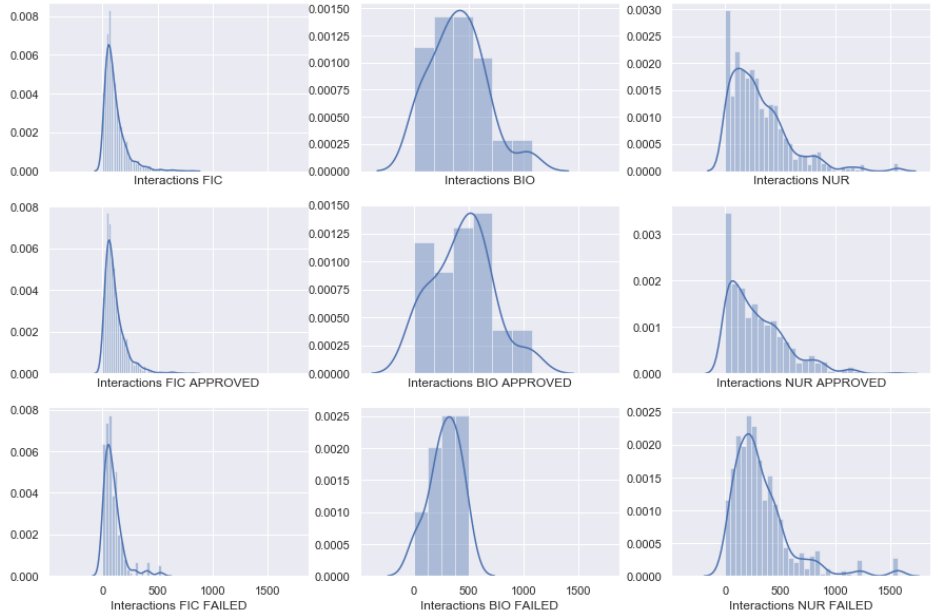


**Figure 7.** Distribution of interactions by courses.

**Table 3.** Statistical analysis.

| | Status | Shapiro | | Mann–Whitney | | Mean | Median | *t*-Test | |
|---|---|---|---|---|---|---|---|---|---|
| | | Statistic | *p*-Value | Statistic | *p*-Value | | | Statistic | *p*-Value |
| BIOAll | Success | 0.959 | 0.138 | - | - | 442.8 | 465 | −2.194 | 0.0322 |
| | Failed | 0.961 | 0.684 | | | 285.1 | 316 | | |
| BIO W4 | Success | 0.841 | 0.00 | 208.5 | 0.00 | 16.55 | 11 | - | - |
| | Failed | 0.673 | 0.00 | | | 7 | 2.5 | | |
| BIO W8 | Success | 0.897 | 0.001 | 188.5 | 0.00 | 81.18 | 92 | - | - |
| | Failed | 0.772 | 0.001 | | | 34.93 | 3.5 | | |
| FIC All | Success | 0.757 | 0.00 | 109,676.5 | 0.00 | 113.7 | 85 | - | - |
| | Failed | 0.791 | 0.00 | | | 94.72 | 73 | | |
| FIC W4 | Success | 0.433 | 0.00 | 114,591.5 | 0.00 | 3.38 | 2.5 | - | - |
| | Failed | 0.393 | 0.00 | | | 1.19 | 0 | | |
| FIC W8 | Success | 0.503 | 0.00 | 114,197.0 | 0.00 | 16.97 | 7.5 | - | - |
| | Failed | 0.278 | 0.00 | | | 4.58 | 0 | | |
| NUR All | Success | 0.814 | 0.00 | 1,061,837.5 | 0.00 | 295.1 | 235 | - | - |
| | Failed | 0.892 | 0.00 | | | 324.2 | 259.5 | | |
| NUR W4 | Success | 0.290 | 0.00 | 1,148,365.0 | 0.05 | 0.59 | 4 | - | - |
| | Failed | 0.350 | 0.00 | | | 0.43 | 0 | | |
| NUR W8 | Success | 0.279 | 0.00 | 1,135,913.0 | 0.02 | 0.87 | 13 | - | - |
| | Failed | 0.514 | 0.00 | | | 0.85 | 0 | | |

As shown in Table 3, the only case where the distribution was normal was for the Biology course considering all 16 weeks. In this case, the T-Test showed a statistical difference between the means of the two groups. For the other cases, the Mann–Whitney test showed statistical differences between the medians of the two groups. These results allowed us to conclude that the students' usage of VLE was associated with their subjects' final status: success or fail.

Another interesting observed attribute is the number of subjects the student was enrolled in and the relation with their final status. Figure 8 shows a box-plot for both groups of students versus the number of subjects enrolled. Even though the mean and median of subjects for both groups were the same (Success: median = 6.0 and mean = 5.93; Fail: median = 6.0; and mean = 6.47), a Mann–Whitney test showed a significant statistical difference between them (statistic = 2,138,630.0, $p$-value < 0.05).

Figure 8 shows that students who failed in the subjects presented a wider dispersion in the number of subjects. Students who had success, tended to enroll in five to eight subjects, while students who failed tended to enroll in three to nine subjects. One of the possible reasons for this discrepancy may be related to the fact that students may enroll in subjects that they are not necessarily interested in taking (as they are allowed to take the final exams without attending classes for those subjects).

This may contribute to the fact that some subjects have a high number of students enrolled although they are not effectively participating. For instance, one given subject from the Nursing course had 590 students enrolled. This is a relatively common practice in Udelar; however, the data seem to show that students regularly attend the subjects in which they have success. This flexibility may also reflect in the engagement of the students in the subjects with students enrolling in more subjects than they are able to attend.

The analysis of the features used in Moodle (Figure 9) showed that Folder, Forum, Page, Quiz, and URL represented around 90% of the students' environment interactions. From these, most were interactions with folders and quizzes.

Although they are all asynchronous tools, they can be separated into two categories: interactive and non-interactive. First, there are methods that do not interact with students or professors and are basically used as a content repository, such as Folder, Page, and URL. Second, there are others with interactions, such as forums and quizzes, but no synchronous communication was found, such as conferencing or chat.

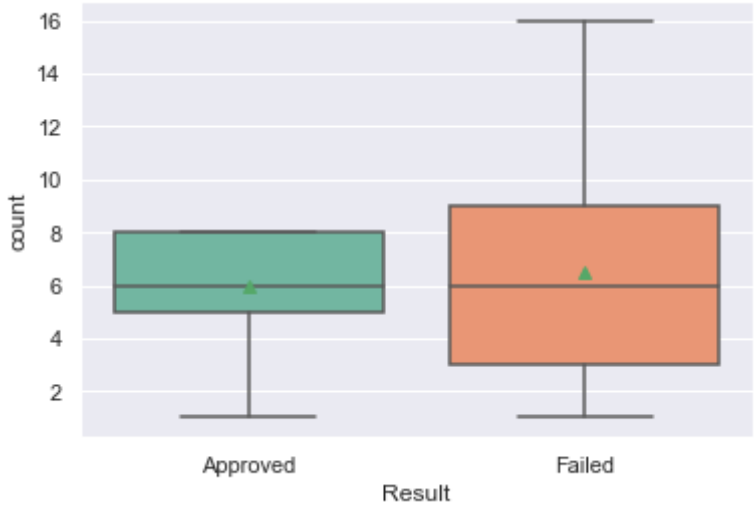


**Figure 8.** Number of subjects enrolled versus final status.

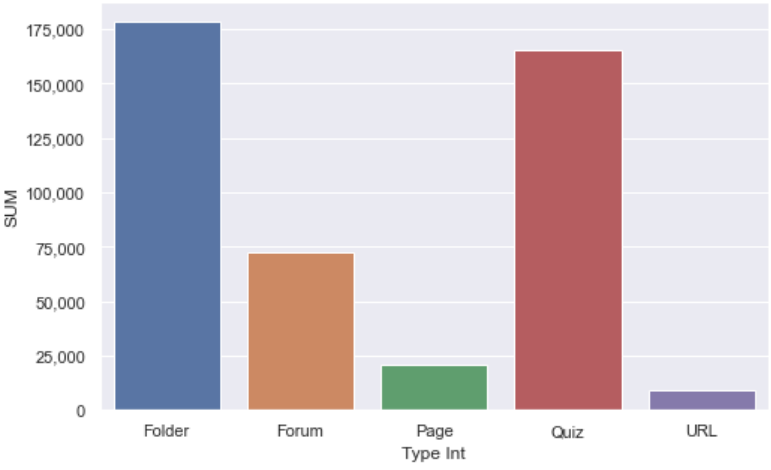


**Figure 9.** Type of interactions.

### 3.5. Data Preparation

After the exploratory data analysis, the data were cleaned, and inconsistencies were treated (such as missing values). Additionally, normalization techniques were applied wherever necessary. The next steps consisted of data integration where scripts were generated to match students enrolled in the subjects with data from survey and the academic system.

The derived attributes were generated from the student's interactions, type of interaction and subject. Subsequently, interactions were grouped every fortnight and classified according to the five main types of interactions used in Udelar's VLE (Folder, Forum, Page, Quiz, and URL). The average number of interactions per week and the standard deviation of the interactions per week (or period) were calculated. This approach is based on previous findings [20,39–41] that indicated the possibility of generating models to predict at-risk students by using the VLE's count of interactions along with the derived attributes from these counts.

The target features for both approaches were constructed based on information from the academic system. The initial academic database consisted of the student's grade in the final exam of the subject and the retaken exam (when this was the case). Thus, two variables were generated from that. In the first scenario, we classified whether the student passed the course without retaking the final exams or if they had to retake the final exams. In the second scenario, we classified whether the student who retook the final exam had success or failure.

### 3.6. Modeling

This step consists of finding the best combinations of input data to generate predictive models, as well as to fine tune the hyperparameters of the algorithms used to generate the models. Data selection and data preparation were performed together with the modeling. An essential task in data mining and predictive modeling is choosing the performance evaluation metric. For this work, we chose the Area Under the Receiver Operating Characteristic Curve (AUC) [42].

The AUC is calculated from the size of the area under the plotted curve where the Y-axis is represented by the True Positive Rate (TPR) or Sensitivity (or Recall) (A1) and the *X*-axis is the True Negative Rate (TNR) or Specificity (A2) [43]. In order to provide a general overview of the results, the following metrics are also presented for comparison: the Accuracy (A5), F1-score (A3), and Precision (A4).

Among the classifiers initially tested, AdaBoost [44], logistic regression [45], and random forest [46] obtained the best results. However, random forest exceeded the others in practically all tested scenarios, and it was chosen for the work sequence. SKlearn's GridSearchCV was chosen as the hyperparameter selection technique. GridSearchCV is a parameter selector that tests a combination of hyperparameters initially set and that returns the one that obtained the best results in the tested set. The data normalization technique with the best results was SKlearn's StandardScaler.

We generated eight different datasets to evaluate the extent to which the different configurations could help to improve the models' performance, as shown in Table 4. The main idea of these configurations is to evaluate how the combination of different datasets may interfere in the models' performance, thus, showing the importance of each database for a better prediction.

The use of DS1 seeks to assess the potential for prediction presented by the survey without any other information besides academic. DS2 is generated by adding the count of total interactions to the survey data. After the EDA, the evaluation shows that this base would be the one with the highest predictive power, being able to be considered the maximum value that can be predicted with the available data. In this way, DS2 is used to compare the gains of using information from the survey along with the information related to the count of interactions.

**Table 4.** Configuration of the different datasets.

| Dataset | Academic Data | Survey | VLE | Type of Interaction | Number of Weeks |
|---|---|---|---|---|---|
| DS1 | YES | YES | NO | - | - |
| DS2 | YES | YES | YES | YES | 16 |
| DS3 | YES | NO | YES | NO | 16 |
| DS4 | YES | NO | YES | YES | 16 |
| DS5 | YES | NO | YES | YES | 8 |
| DS6 | YES | YES | YES | YES | 8 |
| DS7 | YES | YES | YES | YES | 4 |
| DS8 | YES | NO | YES | YES | 4 |

DS3 and DS4 contain the total count of interactions within the VLE, and DS4 also contains the type of each interaction. DS5, DS6, DS7, and DS8 aim to verify the extent to which it is possible to early predict the performance of the students, so that there is time to perform pedagogical interventions. For that, the count of interactions is performed for a limited number of weeks. All datasets that used VLE data contained the derived attributes earlier described according to the number of weeks covered by the dataset and the inclusion of the type of the interaction or not.

After defining the datasets, a random forest classifier was executed in GridSearchCV to obtain the most optimized configuration for the predictive model. The 10-fold cross-validation was used to evaluate the models. The approach to deal with unbalanced data was the Synthetic Minority Oversampling Technique (SMOTE), which generated new synthesized cases on the training datasets.

## 4. Results

This section presents the results obtained by the models for each scenario evaluated and considering the different datasets.

### *4.1. Scenario 1: Predicting Success in Final Exams*

The goal here was to generate predictive models able to classify students between two groups: those who had success in the final exams and those who had to retake the exams. Table 5 presents the results for each dataset configuration and the following metrics: True Positives (TP), True Negatives (TN), Accuracy (ACC), F1-Score, Precision, and Recall. Comparing the metrics here is quite important as the AUC presented low values in some cases, as shown in Figure 10.

In the figure, True Positives (TP) represents the accuracy for classifying the successful students in the final exams and True Negatives (TN) the accuracy for classifying students who need to retake exams. This AUC low value raised the question of whether the random forest model was really learning or just classifying all students in the major category. This was the case of the classifier generated with DS8, which was able to only correctly classify a few cases of the minor category (TP = 12.58 and TN = 94.48). As evidenced in Figure 10, there was an increase in performance when using both the survey and VLE data.

The AUC shows all models with acceptable values (higher than 0.50). DS1 achieved 0.78, which can be considered excellent [43]. Moreover, DS2, DS6, and DS7 achieved values higher than 0.87 and very close to what can be considered outstanding discrimination (0.9 or higher). These are the datasets that combined information from the survey and the VLE. It is important to highlight the results obtained by using DS7, which is the dataset that used data from both the survey and the VLE's count of interactions (including the type of interactions) for the first four weeks of the courses. This model yielded excellent results (AUC = 0.864).

**Table 5.** Predicting success in the final exams versus retaking exams.

| DS | TP | TN | ACC | F1-Score | Precision | Recall |
|---|---|---|---|---|---|---|
| DS1 | 76.96 | 64.04 | 72.40 | 0.72 | 0.72 | 0.72 |
| DS2 | 88.41 | 79.00 | 85.09 | 0.85 | 0.85 | 0.85 |
| DS3 | 85.97 | 77.16 | 82.87 | 0.82 | 0.82 | 0.82 |
| DS4 | 86.98 | 76.11 | 83.00 | 0.83 | 0.83 | 0.83 |
| DS5 | 82.83 | 75.06 | 80.09 | 0.80 | 0.80 | 0.80 |
| DS6 | 87.41 | 75.06 | 83.05 | 0.83 | 0.82 | 0.83 |
| DS7 | 88.12 | 73.75 | 83.05 | 0.82 | 0.82 | 0.83 |
| DS8 | 12.58 | 94.48 | 41.48 | 0.32 | 0.65 | 0.41 |

### *4.2. Scenario 2: Predicting Approval in Retaking Exams*

The second scenario aims to predict whether students will be successful after retaking exams. Table 6 presents the results for each dataset configuration and the following metrics: Positives (TP), True Negatives (TN), Accuracy (ACC), F1-Score, Precision, and Recall. Results obtained using the AUC for the different datasets are presented in Figure 11. Here, TP is the accuracy of correctly classifying a student who failed, and TN is the accuracy of correctly classifying a student who had success in the retake exam.

As shown in Figure 11, the performance of the models for the datasets DS3, DS4, DS5, and DS8 can be classified as acceptable [43]. For DS3, DS4, and DS5, the classifiers presented low accuracy to classify successful students. This leads to the conclusion that it is not recommended to use only the data coming from VLE to predict student performance in this scenario.

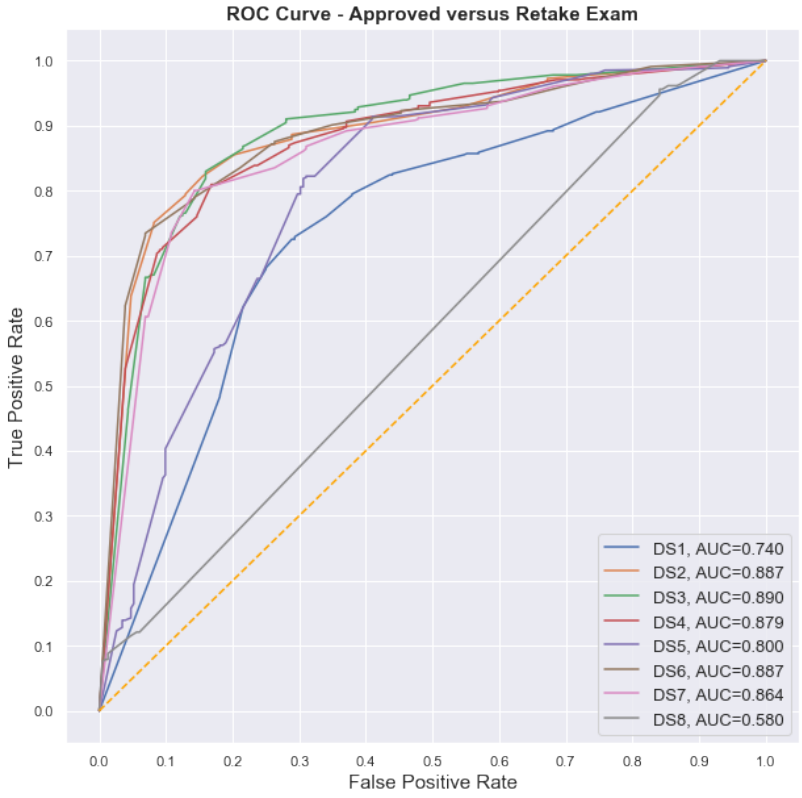


**Figure 10.** ROC Success in a course.

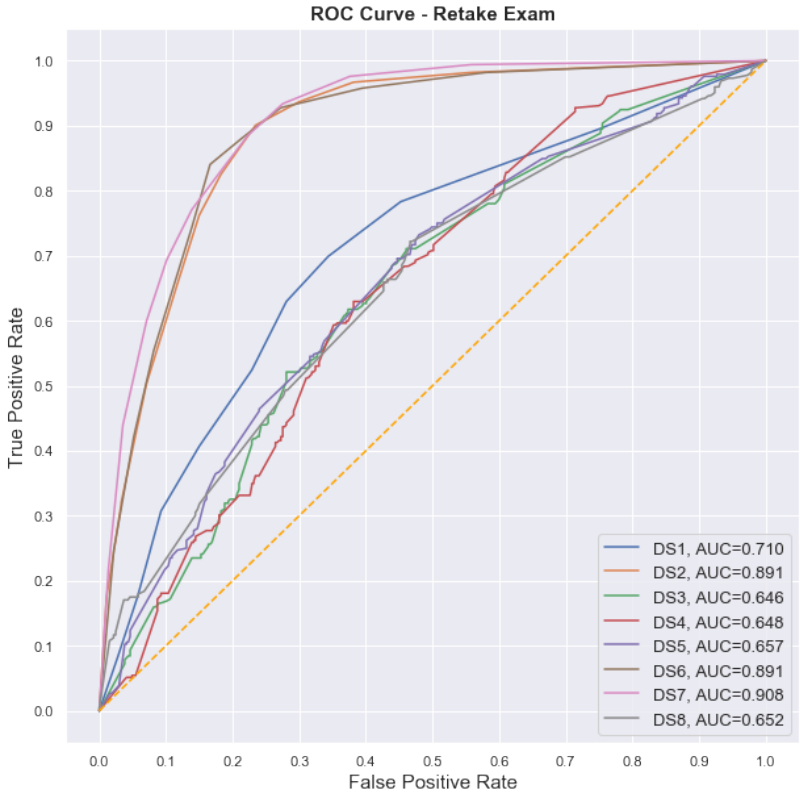


**Figure 11.** ROC Success in the exam.

**Table 6.** Predicting success in retaking the exam versus failure.

| DS | TP | TN | ACC | F1-Score | Precision | Recall |
|---|---|---|---|---|---|---|
| DS1 | 76.96 | 64.04 | 72.40 | 0.72 | 0.72 | 0.72 |
| DS2 | 88.41 | 79.00 | 85.09 | 0.85 | 0.85 | 0.85 |
| DS3 | 85.97 | 77.16 | 82.87 | 0.82 | 0.82 | 0.82 |
| DS4 | 86.98 | 76.11 | 83.00 | 0.83 | 0.83 | 0.83 |
| DS5 | 82.83 | 75.06 | 80.09 | 0.80 | 0.80 | 0.80 |
| DS6 | 87.41 | 75.06 | 83.05 | 0.83 | 0.82 | 0.83 |
| DS7 | 88.12 | 73.75 | 83.05 | 0.82 | 0.82 | 0.83 |
| DS8 | 12.58 | 94.48 | 41.48 | 0.32 | 0.65 | 0.41 |

DS1 presented an excellent AUC but low accuracy to classify students who failed (51.23%). DS2, DS6, and DS7 presented the best results, thus, confirming that using data from the survey together with the VLE's data was the best combination to generate predictive models. The performance achieved by the model trained with DS7 can be classified as outstanding (0.908), which allows us to say that it is possible to generate models to early-predict student performance using the survey and the interactions of the first four weeks for the present scenario.

## 5. Discussion

In this section, we answer the research questions proposed at the beginning of the paper.

**RQ1—Is the use of VLE associated with the students' qualifications?** Yes. We found significant statistical association between the number of student interactions within the VLE and the final status (success or fail). Moreover, after the analysis, we concluded that using only the count of the interactions inside the VLE or using only survey data to generate the predictive models led to lower model performance compared with using a combination of both.

Models trained with a combination and using the count of the first four weeks (DS7) were able to achieve excellent and outstanding performances; thus, one can say that it is

possible to predict students' final status at the beginning of the courses. These findings suggest the importance of VLE in face-to-face courses, even though its usage mainly focuses on the delivery of materials and activities without much collaboration among peers. In addition, it can also be said that different VLE's activities weigh differently inside the models, as the type of interaction also increases their performances.

The work of [20] used VLE data together with data collected from a student survey and evaluated the extent to which the use of the combination of both databases interfered in the models' performance. The authors concluded that there was no gain in using data collected from the survey.

Moreover, the authors also tested different dataset combinations considering the different types of VLE's presence (teaching, cognitive, and social presence), according to the theory of [47], and found no statistically significant difference in the performance of the models that used this differentiation. Their results contradict the present paper's findings. Based on that, one could say that the use of different combinations of databases to improve the performances of the models as well as considering different types of interactions inside the VLE are context dependent.

For the present scenario, the combination of databases and the differentiation of the types of interaction helped to improve the performance of the classifiers.

**RQ2—Which features from the different databases are the most important to early predict students' performance?** Figure 12 presents the fifteen most important attributes used by the models to predict student performance. The attribute that helped the most in predicting student performance was the number of subjects a student was enrolled in. This attribute is located in the academic system database, which was used in all possible scenarios and datasets of this study.

Moreover, attributes that belong to the VLE appeared most frequently in the list (week 2, mean week 2, week 4, and mean week 4, among others). Regarding the VLE attributes, it is important to notice that Forum Week 4 appears at the seventh place of importance. As the forum is used only by the professors to communicate operational/academic/administrative things about the subjects, this attribute may indicate the importance of students being up to date about the daily routine of their courses.

The importance of the types of interaction inside VLE becomes evident from the figure. Attributes Forum, Quiz, URL, and Page appear in the list of the most important attributes together with weeks 2 and 4. Regarding the survey data, it is important to mention that the educational level of the student's mother was the third strongest attribute used by the models. The place of residence is another attribute that played an important role in the prediction.

**RQ3—Which educational patterns can educational data mining help to unveil in the studied courses?**

The most notable finding of the present study is VLE's importance in the teaching-learning process and its association with the final status of the students. This finding can help institutions implement official policies focused on a more widespread dissemination of the VLE usage, along with the other existing faculties, departments, and courses at the university.

Such a policy could encompass different initiatives, such as offering practical training for professors in VLE, the inclusion of introductory subjects in the curriculum, focusing on VLE features and usage, and the increase of physical and personnel infrastructure to maintain new VLE services. Moreover, considering the high accuracy achieved by the predictive models developed here, it is now possible to use such models to follow up students more closely and intervene early on in situations that identify at-risk students.

The university may consider investing in the development of new tools and technologies to follow students' trajectories and improve their learning experiences, e.g., through LA dashboards [48] and e-learning recommender systems [49]. We also found that the number of subjects the students were enrolled in was the strongest attribute associated with their success.

This finding may help coordinators better plan the curriculum of their courses so that students can maintain a course load up to an ideal number. Finally, two other important attributes that influence predictive models were the "mother's education" and "place of residence". These attributes could be monitored by the university to offer assistance aimed at students in these specific categories.

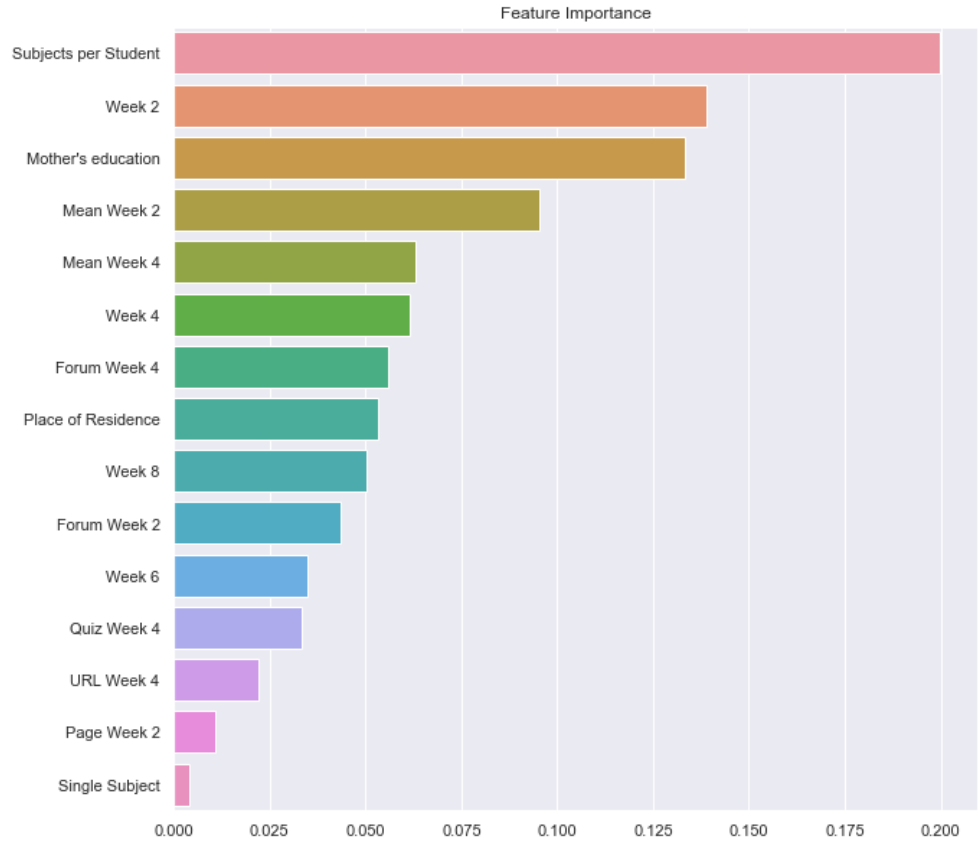


**Figure 12.** Feature importance.

## 6. Towards the Implementation of Institutional Policies

This section presents a broad discussion of the results obtained in this work in order to provide evidence for the implementation of institutional policies at Udelar.

*Institutional Polices Based on Evidence*

University policies are public policies, and the Udelar is the main higher education institution in Uruguay. Considering that, Udelar is also responsible for proposing the agenda of such policies to transform higher education in Uruguay. In this context, it is important to bring the digital inclusion process to the debate, which requires a collective action with some key actors: decision makers, researchers, students, and professors/teachers to which the findings of this study will be useful to generate such public policies.

Furthermore, this study presents conclusions based on evidence and it is fundamental to implement public policies and to treat educational problems, such as evasion, time spent in the program/course (lag), and contextual variables (both are present here). The main beneficiaries of the paper's findings are the general population and it is possible to quantify the impact of that on a regional and international scale. Based on this, we present a discussion considering some factors that are important to the creation and implementation of educational policies based on our findings.

We consider that, at any educational institution, especially at universities, there is more than one database with a diversity of features, and it is potentially possible to combine and mine these features to contribute to the understanding of the learning process. We consider

the creation of strategies to guide the educational policies based on evidence, that is, based on empirical data (information) transformed into knowledge to be very important, necessary, and not able to be postponed.

In this paper, we show evidence on predicting, starting at the beginning of a course, what the final status of the student will be based on a combination of datasets. The use of such predictors (models) allows the manager to create, early in the course, alerts and warnings to students and professors. It is also possible to help the professors to redesign their materials and pedagogical strategies in their courses.

On the one hand, we found that the expanded classrooms were the most frequent classes in the VLEs of the faculties. On the other hand, the importance of hybrid classrooms emerged from our work, since the student performance was higher in this modality. In fact, these models are proposed in a general context of educational uncertainty and changes that are necessary for digital transformation, which has been accelerated.

However, the usefulness of these models for creating educational and institutional policies not only reaches guidelines for the individuals (students, professors, researchers, and so on) but also contributes to the understanding of the educational problems associated with backwardness and evasion, thus, helping to create policies and technical teams to support and protect educational paths.

The results highlight the importance of participants' mediation within the online classrooms, which increases student performance. Additionally, due to the strict relation between the learning and teaching process, the lecturers' formation policies should point to the development of digital skills to allow them to include hybrid models in their teaching–learning process.

Our findings present evidence that the students' age is an important factor using the VLE: the younger they are, the more they use the VLE. In this context, it is good to raise different pedagogical hypotheses to attempt to understand this behavior. If students older than 25 years old do not frequently use the VLE, it is mandatory to develop pedagogical policies to guide strategies to digital literacy focusing on these different groups of students in order to mitigate evasion and to promote lifelong education development supported by educational technologies. This is particularly important in Latin America, where educational institutions tend to present retention problems.

The distribution of interactions during the courses present differences among the subjects and disciplinary areas but are roughly similar when comparing the final status of the students. Those students who had the least interactions were the the ones who tended to fail in the courses. Why did they not use the VLE? Can it be due to educational causes, where the student infers that there is no new content, material, or changes in the teaching process? Or is it an individual cause, where the student infers that it is not necessary to revisit the content and only carry out the tests and assessments to have success?

In this case, we suggest the development of actions addressed to those students who fail, stimulating them to use the VLE through instructional design specifically dedicated to this population. The other way around, successful students had more interactions within the VLE, and this raises more questions: why did they have success in the subject? Is it due to the interactions with educational technologies? These are all questions that still need to be answered and that will help with the development of institutional policies based on evidence.

The VLE interactions appear to be a very strong indicator of a student's commitment to their studies. It shows how they tackle the learning process and what their strategies are to achieve success. Furthermore, students that sustained their participation permanently within the VLE over the whole semester were more likely to pass than students who participatee only at the end, even when they used the environment intensively (but still less than the successful students). The comparison between the student trajectories in VLE shows learning strategies that resulted in better performance and allow for the development of protection policies based on evidence. One way to do this is to design teaching and learning paths considering these students and their strategies.

The incorporation of the VLE in the educational processes of undergraduate teaching at Udelar has reached a point of naturalization. The results of the present study show the relevance of understanding the relations between the behavior of the students inside the VLE and their success in their disciplines, as this allows one to define didactic strategies, pedagogical orientations, and educational policies based on that.

Even though a VLE produces, collects, and stores a large amount of data about students and teachers interactions, there are a number of challenges and difficulties to face before properly transforming this data into meaningful knowledge. For that, specific computational strategies and tools are required. The present work also presents a contribution in this matter, as it provides a methodological framework that uses both EDM and LA to better understand students behavior inside a VLE.

## 7. Conclusions, Limitations, and Future Research

The present paper analyzed different aspects of data involving students enrolled in courses at the Universidad de la República in Uruguay. Precisely, we collected data from 4529 students of three programs and through three different sources: the academic system, an academic survey, and a VLE. We applied data science techniques (visualizations, statistics, and data mining) to understand how different combinations of the datasets could help predict students' final status in the subjects and the role that different attributes played in this task.

The results presented an overview of the institutional patterns regarding the use of the VLE, and this will help pave the way for the implementation of future policies in institutions to diminish student failures and increase persistence. Among the findings was an association between the use of the VLE and the final status of the students (success and fail) and also the different types of activities inside the VLE presenting different levels of importance in this association.

Examples of institutional policies that could emerge from these findings are as follows: the allocation of extra computational resources for improving VLE infrastructure and its widespread use in the university, the development of new tools for following students' trajectory and detecting at-risk students at early stages of their courses, and the construction of more institutional policies to mitigate students' failure based on other relevant attributes (e.g., the number of subjects the student is enrolled in, the student's mother's education, and the student's neighborhood).

The proposed methodology for combining different data sources, as well as their pre-processing and feature engineering, demonstrated that the combination of data had a high predictive power. In this regard, the combination of the survey variables, academic system, and virtual environment showed a high capacity for early prediction. Thus, it was possible to achieve prediction rates with outstanding discrimination as soon as in the fourth week of the course. This characteristic satisfies the temporal factor of precocity, which is considered to be a determining factor in identifying and attempting to reverse the problem [31,50].

This proposed approach model, although initially restricted to only three university programs, can serve as a basis for future work that seeks to implement methods of online information and prediction on student behavior, such as academic dashboards. However, for these steps, it is still necessary to clarify two key points: how this approach would behave with more data and the analysis of its acceptance regarding the technology and the reliability of the methods by the stakeholders, teachers, and students.

The present work can help the university to develop user profiles based on the students practices inside the VLE, thus, allowing the future development of systems able to continuously deliver indicators related to the learning processes. In this way, it contributes to the production of primary information that can potentially help to the evaluation of quality and the definition of strategies that guide the university teaching and learning processes.

One limitation of the present work is the lack of a qualitative analysis of the scenarios. Future work could explore the opinions of students and professors regarding the usage and importance of the VLE in their teaching and learning processes. Another limitation is the restricted number of courses used in this study. As mentioned before, Udelar has 100 undergraduate courses and the number of courses studied here (only three) can not be considered representative of the whole university, even though it serves for the purpose of an initial assessment. Future work could expand the data analyzed by increasing the number of courses. Future work could also include new data covering the period of the COVID-19 pandemic and evaluate how this period influenced the behavior of the students inside the VLE.

Moreover, future work could explore a voting scheme with the learning algorithms utilized here (AdaBoost, logistic regression, and random forest) to improve the accuracy of the predictions. Finally, it would be interesting to also explore the reasoning followed by the predictors developed here, thus, assisting the stakeholders to better understand the role each feature plays in the classification.

Finally, traditional approaches to the investigation of student persistence in the teaching–learning process are normally carried out from the sociology of education and educational sciences with a fundamentally deductive perspective. The introduction of data science tools with inductive approaches challenges and empowers traditional theoretical and methodological models of educational science. The construction of this interdisciplinary exchange bridge is perhaps the most significant contribution to the academic community that may help in constructing university educational policies.

**Author Contributions:** E.M.Q.: experimental data analysis, algorithms development, experiments conduction, results description, and manuscript writing; C.C.: methodology definition, experiments setup, and writing; A.P.C.: virtual courses setup, server administration, data setup, and pre-processing; V.R.P.: writing, editing, review, and educational policies proposals; L.R.B.: writing; V.F.C.R.: writing, review, and editing; C.R.E.: writing, editing, review, and educational policy proposals. The manuscript was written and approved to submit by all authors. All authors have read and agreed to the published version of the manuscript

**Funding:** This work was supported by Faculty of Nursing of Udelar (University of the Republic) through PIMCEU project and by CNPq (Brazilian National Council for Scientific and Technological Development) [Edital Universal, proc.404369/2016-2] [DT-2 Productivity in Technological Development and Innovative Extension scholarship, proc.315445/2018-1].

**Data Availability Statement:** Please contact the authors for data requests.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| EDM | Educational Data Mining |
| LA | Learning Analytics |
| Udelar | University of the Republic |
| VLE | Virtual Learning Environments |
| LMS | Learning Management Systems |
| Moodle | Modular Object-Oriented Dynamic Learning Environment |
| LSTM | Long Short Term Memory |
| CRISP-DM | Cross-Industry Standard Process for Data Mining |
| SLR | Self-Regulated Learning |
| KDD | Knowledge Discovery in Databases |

## Appendix A. Formulas

True Positive Rate (TPR) or Sensitivity (or Recall)

$$TPR = \frac{TP}{TP + FN} \tag{A1}$$

True Negative Rate (TNR) or Specificity

$$TNR = \frac{TN}{TN + FP} \tag{A2}$$

F-Score

$$F1 - Score = 2X\frac{Precision * Recall}{Precision + Recall} \tag{A3}$$

Precision

$$Precision = \frac{TP}{TP + FP} \tag{A4}$$

Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{A5}$$

## References


1. Hilliger, I.; Ortiz-Rojas, M.; Pesántez-Cabrera, P.; Scheihing, E.; Tsai, Y.S.; Muñoz-Merino, P.J.; Broos, T.; Whitelock-Wainwright, A.; Pérez-Sanagustín, M. Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. *Internet High. Educ.* **2020**, *45*, 100726. [CrossRef]
2. Kurilovas, E. On data-driven decision-making for quality education. *Comput. Hum. Behav.* **2020**, *107*, 105774. [CrossRef]
3. McKnight, K.; O'Malley, K.; Ruzic, R.; Horsley, M.K.; Franey, J.J.; Bassett, K. Teaching in a digital age: How educators use technology to improve student learning. *J. Res. Technol. Educ.* **2016**, *48*, 194–211. [CrossRef]
4. Palacios, C.A.; Reyes-Suárez, 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* **2021**, *23*, 485. [CrossRef] [PubMed]
5. Salazar-Fernandez, J.P.; Sepúlveda, M.; Munoz-Gama, J.; Nussbaum, M. Curricular Analytics to Characterize Educational Trajectories in High-Failure Rate Courses That Lead to Late Dropout. *Appl. Sci.* **2021**, *11*, 1436. [CrossRef]
6. Gómez-Pulido, J.A.; Park, Y.; Soto, R. Advanced Techniques in the Analysis and Prediction of Students' Behaviour in Technology-Enhanced Learning Contexts. 2020. Available online: https://www.mdpi.com/2076-3417/10/18/6178 (accessed on 3 May 2021).
7. OECD. *Benchmarking Higher Education System Performance*; OECD Publishing: Paris, France, 2019; p. 644. [CrossRef]
8. Gralka, S. Persistent inefficiency in the higher education sector: Evidence from Germany. *Educ. Econ.* **2018**, *26*, 373–392. [CrossRef]
9. Aldowah, H.; Al-Samarraie, H.; Fauzy, W.M. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telemat. Inform.* **2019**, *37*, 13–49. [CrossRef]
10. Moissa, B.; Gasparini, I.; Kemczinski, A. A systematic mapping on the learning analytics field and its analysis in the massive open online courses context. *Int. J. Distance Educ. Technol. (IJDET)* **2015**, *13*, 1–24. [CrossRef]
11. Romero, C.; Ventura, S. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2010**, *40*, 601–618. [CrossRef]
12. Kabathova, J.; Drlik, M. Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Appl. Sci.* **2021**, *11*, 3130. [CrossRef]
13. Brown, M. Learning analytics: Moving from concept to practice. In *EDUCAUSE Learning Initiative, v. 7*; 2012; pp. 1–5. Available online: https://library.educause.edu/-/media/files/library/2012/7/elib1203-pdf.pdf (accessed on 24 June 2021).
14. Waheed, H.; Hassan, S.U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* **2020**, *104*, 106189. [CrossRef]
15. Gasevic, D.; Tsai, Y.; Dawson, S.; Pardo, A. How do we start? An approach to learning analytics adoption in higher education. *Int. J. Inf. Learn. Technol.* **2019**, *36*, 342–353. [CrossRef]
16. Alghamdi, A.; Karpinski, A.C.; Lepp, A.; Barkley, J. Online and face-to-face classroom multitasking and academic performance: Moderated mediation with self-efficacy for self-regulated learning and gender. *Comput. Hum. Behav.* **2020**, *102*, 214–222. [CrossRef]
17. Xia, X. Interaction recognition and intervention based on context feature fusion of learning behaviors in interactive learning environments. In *Interactive Learning Environments*; 2021; pp. 1–18. Available online: https://www.tandfonline.com/doi/full/10.1080/10494820.2021.1871632 (accessed on 24 June 2021).
18. MOODLE. Statistics. 2020. Available online: https://stats.moodle.org/ (accessed on 3 April 2020).
19. Hegazi, M.O.; Abugroon, M.A. The state of the art on educational data mining in higher education. *Int. J. Comput. Trends Technol.* **2016**, *31*, 46–56. [CrossRef]
20. Macarini, B.; Antonio, L.; Cechinel, C.; Batista Machado, M.F.; Faria Culmant Ramos, V.; Munoz, R. Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Appl. Sci.* **2019**, *9*, 5523. [CrossRef]

21. Leitner, P.; Ebner, M.; Ebner, M. Learning Analytics Challenges to Overcome in Higher Education Institutions. In *Utilizing Learning Analytics to Support Study Success*; Ifenthaler, D., Mah, D.K., Yau, J.Y.K., Eds.; Springer: Cham, Switzerland, 2019; pp. 91–104, ISBN 978-3-319-64792-0. [CrossRef]
22. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Appl. Sci.* **2020**, *10*, 1042. [CrossRef]
23. Cechinel, C.; Ochoa, X.; Lemos dos Santos, H.; Carvalho Nunes, J.B.; Rodés, V.; Marques Queiroga, E. Mapping Learning Analytics initiatives in Latin America. *Br. J. Educ. Technol.* **2020**, *51*, 892–914. [CrossRef]
24. Chui, K.T.; Fung, D.C.L.; Lytras, M.D.; Lam, T.M. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Comput. Hum. Behav.* **2020**, *107*, 105584. [CrossRef]
25. Ding, M.; Yang, K.; Yeung, D.Y.; Pong, T.C. Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*; LAK19; ACM: New York, NY, USA, 2019; pp. 135–144. [CrossRef]
26. Foster, E.; Siddle, R. The effectiveness of learning analytics for identifying at-risk students in higher education. *Assess. Eval. High. Educ.* **2020**, *45*, 842–854. [CrossRef]
27. Gutiérrez, F.; Seipp, K.; Ochoa, X.; Chiluiza, K.; De Laet, T.; Verbert, K. LADA: A learning analytics dashboard for academic advising. *Comput. Hum. Behav.* **2020**, *107*, 105826. [CrossRef]
28. Lee, S.; Chung, J. The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Appl. Sci.* **2019**, *9*, 3093. [CrossRef]
29. Herodotou, C.; Rienties, B.; Verdin, B.; Boroowa, A. Predictive learning analytics 'at scale': Towards guidelines to successful implementation in Higher Education based on the case of the Open University UK. *J. Learn. Anal.* **2019**, in press. [CrossRef]
30. Li, Q.; Baker, R.; Warschauer, M. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *Internet High. Educ.* **2020**, *45*, 100727. [CrossRef]
31. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* **2009**, *53*, 950–965. [CrossRef]
32. Tlili, A.; Denden, M.; Essalmi, F.; Jemni, M.; Chang, M.; Kinshuk; Chen, N.S. Automatic modeling learner's personality using learning analytics approach in an intelligent Moodle learning platform. In *Interactive Learning Environments*; 2019; pp. 1–15. Available online: https://www.tandfonline.com/doi/abs/10.1080/10494820.2019.1636084?journalCode=nile20 (accessed on 24 June 2021). [CrossRef]
33. Hu, Q.; Rangwala, H. Reliable Deep Grade Prediction with Uncertainty Estimation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*; LAK19; ACM: New York, NY, USA, 2019; pp. 76–85. [CrossRef]
34. Pintrich, P.R. The role of goal orientation in self-regulated learning. In *Handbook of Self-Regulation*; Elsevier: Amsterdam, The Netherlands, 2000; pp. 451–502.
35. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*; Springer-Verlag: London, UK, 2000; Volume 1, pp. 29–39.
36. Dirección General de Planeamiento. *Estadísticas Básicas 2018 de la Universidad de la República*; Technical Report; Universidad de la República: Montevideo, Uruguay, 2018.
37. Universidad de la República. Relevamiento de Estudiantes: Udelar Crece y Democratiza. 2019. Available online: http://www.universidad.edu.uy/prensa/renderItem/itemId/43652/refererPageId/12 (accessed on 3 June 2021).
38. Rodés, V.; Canuti, L.; Regina Motz, N.M. Aplicando una categorización a diseños educativos de cursos en entornos virtuales. *Calid. Y Accesibilidad De La Form. Virtual* **2012**, *1*, 425–432.
39. Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, 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. *Appl. Sci.* **2020**, *10*, 3998. [CrossRef]
40. Queiroga, E.; Cechinel, C.; Araújo, R. Predição de estudantes com risco de evasão em cursos técnicos a distância. In *Anais da XXVIII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação (SBIE 2017))*; de Menezes, C.S., Melo, J., Eds.; Sociedade Brasileira de Computação—SBC: Recife, Brazil, 2017; pp. 1547–1556.
41. Machado, M.; Cechinel, C.C.; Ramos, V. Comparação de diferentes configurações de bases de dados para a identificação precoce do risco de reprovação: O caso de uma disciplina semipresencial de Algoritmos e Programação. *Braz. Symp. Comput. Educ. (Simpósio Bras. De Inform. Na Educ. SBIE)* **2018**, *29*, 1503. [CrossRef]
42. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 26. [CrossRef]
43. Gašević, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet High. Educ.* **2016**, *28*, 68–84. [CrossRef]
44. Schapire, R.E. Explaining adaboost. In *Empirical Inference*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
45. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition*, 4th ed.; Foundations, v. 19; Pearson Deutschland GmbH: München, Germany, 2021; p. 23.
46. Liu, Y.; Wang, Y.; Zhang, J. New Machine Learning Algorithm: Random Forest. In *Information Computing and Applications*; Liu, B., Ma, M., Chang, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 246–252.
47. Garrison, D.R.; Anderson, T.; Archer, W. Critical inquiry in a text-based environment: Computer conferencing in higher education. *Internet High. Educ.* **1999**, *2*, 87–105. [CrossRef]

48. Einhardt, L.; Tavares, T.A.; Cechinel, C. Moodle analytics dashboard: A learning analytics tool to visualize users interactions in moodle. In Proceedings of the 2016 XI Latin American Conference on Learning Objects and Technology (LACLO), San Carlos, Costa Rica, 3–7 October 2016; pp. 1–6.
49. dos Santos, H.L.; Cechinel, C.; Araújo, R.M. A comparison among approaches for recommending learning objects through collaborative filtering algorithms. *Program* **2017**, *51*, 35–51. [CrossRef]
50. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Mousa Fardoun, H.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]

# APÊNDICE C – A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course

applied sciences

MDPI

Article

# A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course

**Emanuel Marques Queiroga [1,2,*], João Ladislau Lopes [2], Kristofer Kappel [1], Marilton Aguiar [1], Ricardo Matsumura Araújo [1], Roberto Munoz [3,*], Rodolfo Villarroel [4] and Cristian Cechinel [5]**

1 Centro de Desenvolvimento Tecnológico (CDTEC), Universidade Federal de Pelotas (UFPel), Pelotas 96010610, Brazil; kristofer.kappel@gmail.com (K.K.); marilton@inf.ufpel.edu.br (M.A.); ricardo@inf.ufpel.edu.br (R.M.A.)

2 Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-Grandense (IFSul), Pelotas 96015560, Brazil; joao.lblopes@gmail.com

3 Escuela de Ingeniería Informática, Universidad de Valparaíso, Valparaíso 2362735, Chile

4 Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile; rodolfo.villarroel@ucv.cl

5 Centro de Ciências, Tecnologias e Saúde (CTS), Universidade Federal de Santa Catarina (UFSC), Araranguá 88906072, Brazil; contato@cristiancechinel.pro.br

* Correspondence: emanuelmqueiroga@gmail.com (E.M.Q.); roberto.munoz@uv.cl (R.M.)

Received: 1 May 2020; Accepted: 28 May 2020; Published: 9 June 2020


**Abstract:** Contemporary education is a vast field that is concerned with the performance of education systems. In a formal e-learning context, student dropout is considered one of the main problems and has received much attention from the learning analytics research community, which has reported several approaches to the development of models for the early prediction of at-risk students. However, maximizing the results obtained by predictions is a considerable challenge. In this work, we developed a solution using only students' interactions with the virtual learning environment and its derivative features for early predict at-risk students in a Brazilian distance technical high school course that is 103 weeks in duration. To maximize results, we developed an elitist genetic algorithm based on Darwin's theory of natural selection for hyperparameter tuning. With the application of the proposed technique, we predicted the student at risk with an Area Under the Receiver Operating Characteristic Curve (AUROC) above 0.75 in the initial weeks of a course. The results demonstrate the viability of applying interaction count and derivative features to generate prediction models in contexts where access to demographic data is restricted. The application of a genetic algorithm to the tuning of hyperparameters classifiers can increase their performance in comparison with other techniques.

**Keywords:** at-risk students; genetic algorithm; learning analytics; educational data mining

## 1. Introduction

Learning analytics (LA) approaches have emerged in the context of the increasing use of digital information and communication technologies in education [1]. LA provides information and knowledge so that institutions can overcome core challenges with the qualification of their teaching and learning processes [2,3]. Student dropout is one of the main problems in e-learning that has received considerable attention from the research community. Early detection of students at risk of dropout plays an essential role in reducing the problem, enabling targeted actions aimed at specific situations [4–6].

According to OECD [7], contemporary education is vast, and there are many concerns about the performance of education systems. Among the various important challenges faced in education, one of the most difficult to tackle is the low completion rates observed in many institutions [8], being the final representation of the high dropout rates and low student performance in courses. These problems are related to many factors other than teaching methodologies, such as the profile of the students and their ability to self-manage time [7,9,10].

Dropout rates in e-learning are generally higher compared with face-to-face education [8]. According to the European Commission on Education and Culture, countries like Poland, Sweden, and Hungary have dropout rates in higher education of 38%, 47%, and 47%, respectively [11]. In Spain, the dropout rate is 50% at the Spanish National Distance Education University (UNED) [12]. In Brazil, the enrolment numbers significantly increased in the last few years, but student dropout rates simultaneously increased. The last census of distance education in Brazil [13] reported dropout rates of 50% in the distance courses offered by the Ministry of Education.

Studies have established that student success in distance courses is directly correlated with their engagement inside virtual learning environments (VLEs). Distance learning technology allows tutors to measure the engagement of students by looking into system logs and evaluating the intensity of students' interactions in the different activities available inside virtual classrooms [9,14,15].

In the educational context, access to data is a considerable challenge [16]. The distribution of institutional and academic data across numerous systems creates challenges for accessing social-demographic and previous academic data. This occurs typically because VLEs are usually unprepared for the storage of this kind of data and because several educational institutions apply different learning modalities, such as face-to-face learning, hybrid learning, and distance learning, thus requiring a central academic system. Data are usually concentrated in a central academic system that has no direct connection with the virtual environment. This situation restricts the automated retrieval of data external to VLE, for example, to use in dashboards for data visualization or in the generation of predictive models. In many institutions, the access to use this kind of data is restricted either due to internal policies or data access legislation [16–18].

One of the main advantages of distance learning courses is the large amount of data generated by interactions between students and the system, which provides new possibilities for studying and understanding the data. In e-learning courses, the interaction between students and teachers is usually mediated by a VLE. Thus, VLEs generate a large volume of data that can be consumed by machine learning models [19]. Machine learning algorithms have been used to build successful classifiers using diverse student attributes [10]. While these models showed promising results in several settings, these results are usually attained using attributes that are not immediately transferable to other courses or platforms.

In machine learning, the parameters are defined by the model generated by the algorithms, unlike regular programming, where the term *parameter* is used to refer to the entry of a given function. The final accuracy of models is directly linked to the quality of the fine-tuning of their hyperparameters on the algorithm input [20]. Thus, more adjusted hyperparameters result in more accurate models [21,22]. The control variables of the classifiers are called hyperparameters, which aim to define relevant issues regarding the model to be trained, such as the number of estimators in a random forest algorithm or the number of layers hidden in a neural network [21]. In a neural network context, parameters are adjusted during the training phase using weights. Hyperparameters are variables set before the training, such as the network topology or learning information [22].

In this context, we previously proposed exploiting students' interaction counts solely over time (and other attributes derived from the counts) to predict at-risk students [23–25]. This approach was tested and produced good results, allowing the early prediction of students at risk of dropout and achieving overall accuracies varying from 65% to 90% in the first eight weeks of a two-year distance courses. These studies produced results comparable to those in the literature. For instance,

Jayaprakash et al. [26] obtained general accuracies varying from 73% to 94% and Manhães et al. [4] reported accuracies from 62.22% to 67.77%.

Maximizing the results obtained by predictions is a considerable challenge [27], as the different algorithms commonly present a wide variation in the performance rates that depend on the combination of several characteristics (e.g., balance among classes, amount of data, input variables, and others) and algorithm hyperparameters [28]. Evolutionary computation, and especially genetic algorithms (GAs), are used for optimization problems and tuning classifiers in several areas such as medicine [20] and emotion recognition [29], producing significant results. Here, we propose the use of an evolutionary GA to tune the hyperparameters of the classifiers, thereby optimizing the performance of the models for the early detection of students at risk of dropping out.

This paper is a continuation of these previous works, now aiming to enhance the results by applying an approach that uses GAs to tune machine learning algorithms' hyperparameters. This paper contrasts the results of two methods for hyperparameter optimization applied on models to detect at-risk students in technical e-learning courses based on the counting of students' interactions inside the VLE. The first method for hyperparameter optimization is based on a GA created by the authors, and the second is the traditional widely used method called grid search [21]. During this study, we aimed to answer the following research questions:

**RQ1.** Does the GA approach to hyperparameter optimization outperform traditional techniques?
**RQ2.** Does the resulting predictive models generated by the use of the GA approach for hyperparameter optimization perform better than models with default hyperparameters?

The remainder of this paper is organized as follows: Section 2 presents the theoretical background and related work about the problem of predicting at-risk students and the use of GAs in this context. Section 3 presents the case study conducted to test the proposed solution, detailing the data gathered, the methodology, the proposed GA for fine-tuning, and the experiments. Section 4 discusses the results, and Section 5 concludes the paper and proposes future work.

## 2. Theoretical Background

This section presents works focused on predicting at-risk students in different scenarios and the use of hyperparameter techniques to improve results. Several works in the field of learning analytics and educational data mining deal with the problem of early predicting at-risk students. The works usually differ according to several aspects, such as (1) the sources of data used to generate the models for prediction (demographic, VLEs, surveys, exams); (2) the level of education of the courses (high school, secondary education); (3) the goal of the predictive models (e.g., to predict performance or evasion); (4) the scope of the prediction focused on an entire program or a specific course or discipline; (5) the modality of the course (formal or informal, face-to-face, blended, or distance learning); and (6) whether or not to use tuning techniques for classifiers.

According to Liz-Domínguez et al. [30], data analysis is the set of techniques used to transform data into information and knowledge, thus revealing correlations and hidden patterns. The data resulting from this process can be used to create early warning systems to predict future events. This process mainly aims to support learning and mitigate some of the problems, such as academic performance, retention, and dropout. The reliability of the predictions by the predictor is one of the main factors established by Liz-Domínguez et al. [30] and Herodotou et al. [31] for their application on a large scale.

According to Liz-Domínguez et al. [30], researchers have experimented with methodologies in different scenarios. However, according to Hilliger et al. [32] and Cechinel et al. [33], in Latin America, these studies are mainly concentrated in the university context, so more applications in other contexts are necessary.

González et al. [34] demonstrated that information and communication technologies have a greater impact on the teaching and education process. González et al. [34], de Pablo González [35]

demonstrated the significant impact of the use of VLEs by teachers on student learning. This impact can be maximized using intervention methods based on machine learning, as proposed by Herodotou et al. [36]. Herodotou et al. [31] demonstrated that the classes where teachers used predictive methods produced a performance at least 15% higher than the classes without that use. This improvement was also observed in comparison with classes with the same teachers but from previous years.

In the educational context, traditional research usually uses data from educational systems and virtual environments. The research by Zohair [37] proposed only using data from the academic system (e.g., extracurricular courses, grades, and age) to predict performance in graduate students. Some of the extracted data were extracurricular courses taken and the respective grades, initial training course, and descriptive data about the grades and the age of the student. This study demonstrated that for small groups of students, this is a logical approach that produces good results with few pre-processing steps and a limited set of data. The author focused on the use of algorithms that perform well with low amounts of data, such as support vector machines and multilayer perceptrons (MLP), that produce results with accuracy above 76%.

The search for methods that can be generalized and therefore replicable for other courses represents a significant portion of the research. Thus, studies such as [38] proposed an architecture that is not dependent on a single type of datum, working with the flow of clicks that academics make in a Massive Open Online Course (MOOC). To do so, data are captured from a course and different prediction models are trained and tested in other courses and environments. The experiments showed 87% accuracy when testing in different courses and 90% when tested in the same course, not varying significantly according to the environment.

In [39], several techniques for pre-processing data were compared in terms of interactions with the virtual environment Moodle in risk prediction. Data from the plugin Virtual Programming Laboratory (VPL) were used for risk prediction in algorithm and programming disciplines in undergraduate courses. Data such as weekly interaction count, an average of interactions, median, number of weeks without interactions, standard deviation, and commitment factor are generated based on a previously proposed technique [25,40]. Data added included the teacher interaction count, social count, and cognitive count based on a proposed theory Swan [41]. With naturally unbalanced data, the synthetic minority over-sampling technique (SMOTE) was applied to create balance. Several datasets were generated with different variables to compare the techniques. The results demonstrated that the use of only the interaction count as proposed in [24,25] presented results superior to the other techniques, including their union.

For instance, [5] proposed a students' dropout prediction system that combines outcomes from three different algorithms (neural network, support vector machine (SVM), and probabilistic ensemble simplified fuzzy Adaptive Resonance Theory (ARTMAP—PESFAM)). The authors gathered static demographic data, like sex and place of residence; academic data, like performance and scholar degree; and dynamic data, such as the number of interactions in the virtual environment, grades, and even delivery dates of activities. After applying the algorithms, three distinct approaches to the dropout prediction were generated: (1) A student is considered a dropout case if at least one method classified them as such, (2) a student is considered as a dropout if at least two methods indicated the student to be a dropout and, (3) the student is only presumed as a dropout if all three techniques classified them as a dropout. The accuracy of the results obtained ranged from 75% to 85%, and the best results were achieved using the less restrictive approach, the first one, which achieved accuracies up to 85% on the first section of a given course.

Jayaprakash et al. [26] proposed a warning system focused on student performance to reduce dropout and retention rates. The system provides the student with updated feedback on their potential scholarly performance. To do so, the system uses several types of data, such as demographic (sex and age), student interactions on the VLE, previous academic performance, time passed since the student entered the university, online time spent on the VLE, and outcomes from the scholastic aptitude

test (SAT) (verbal and math). Different models of prediction were produced using J48, Bayesian networks with naive Bayes, SVM with minimal sequential optimization (SMO), and logistic regression, considering data from 9938 students. These classifiers presented similar results, with the classifier based on logistic regression producing slightly superior outcomes (94.2% general accuracy and 66.7% precision for identifying students at dropout risk).

A classifier able to early predict student dropout using students' interactions inside a VLE was proposed [42]. They used information such as if the student watched all video tutorials, if the student ignored some given material or activity, if the student was delayed in following the virtual classes, and the student performance in the activities. Students were then classified according to three flags: Green (low dropout risk), yellow (medium dropout risk), and red (high dropout risk). The authors did not mention the types of machine learning algorithms used but reported performance (TP accuracy) varying from 40% to 50% to predict dropout students within two weeks in advance.

Genetic algorithms are widely used in data mining and can be implemented as the classifier or as a result of the optimizer, as proposed in this approach. One of the applications of genetic algorithms for optimization is a method combining the predictions generated by classifiers. To this end, Minaei-Bidgoli and Punch [43] proposed the application of machine learning to predicting student performance in an online physics course at Michigan State University. For this, data derived from the tasks performed by the students were used. Ten different variables were extracted, including success rate, success on the first attempt, the number of attempts, the time between task delivery and deadline, the time involved in solving, and the number of interactions with colleagues and instructors. A principal component analysis (PCA) method was applied to transform the variables, and three different sets with two, three, or nine components were generated. After this, the Bayes classifier, I-nearest neighbor (I-NN), k-nearest neighbor (k-NN), Parzen-window, multilayer perceptron (MLP), and decision tree classifiers were applied. Then, the predictions obtained by the classifiers were combined with the genetic algorithm using 200 individuals with 500 generations. The GA proposed by the author achieved optimization of 10% to 12% depending on the number of components in the input.

Márquez-Vera et al. [6] proposed the evolutionary algorithms Interpretable Classification Rule Mining Algorithm (ICRM) [27] and ICRM2 [6] based on grammar-based genetic programming (GBGP). In Márquez-Vera et al. [6], ICRM was used to predict the dropout of high school students in Mexico. The authors proposed a double-approach prediction on the same algorithm, creating two classification rules: One for identifying students who tend to complete the course and the other for students who tend to drop out. The data used included 60 attributes that range from the entrance test to research data distributed to students. As a comparison method, the algorithm proposed by the author was compared with five classifiers: Naive Bayes, decision tree, Instance-based lazy learning (IBK), Repeated Incremental Pruning (JRip), and SVM. Techniques were also used to reduce the dimensionality of the base. Using the accuracy as an evaluation metric, the results obtained by the proposed algorithm showed that it can be a valid approach, especially considering the ease of interpretation of the generated classification rules.

## 3. Proposed Approach

The proposed approach consists of the use of a GA for the classifier (hyperparameter) optimization and selection of the fittest, to predict dropout in distance learning courses. Figure 1 shows the proposed solution. The following machine learning algorithms were selected to test the solution: Classic decision tree (DT), random forest (RF), multilayer perceptron (MLP), logistic regression (LG), and the meta-algorithm AdaBoost (ADA). The proposed approach was compared against the grid search method regarding hyperparameter optimization and the regular solution without hyperparameter optimization. The proposed approach uses a classification method, where several classifiers with different hyperparameters, such as DT, RF, MLP, LG, and ADA, compete against each other. In the end, the classifier and the hyperparameters with the best results are selected by a fitness function.

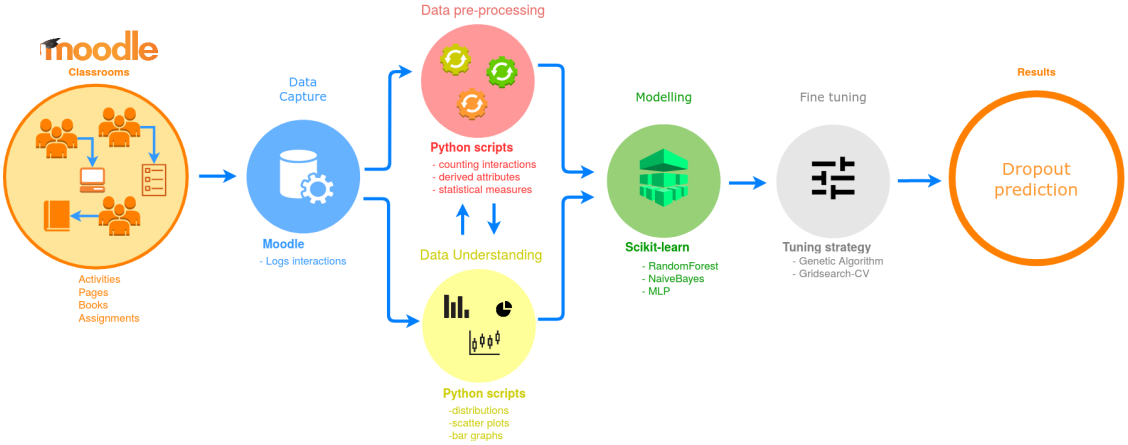


**Figure 1.** Proposed approach.

### *3.1. Case Study*

The case study consisted of the following steps: Data capture, data pre-processing, data understanding, and modelling, according to the solution proposed in the Figure 1. These steps occur in parallel, with tests, implementations, and generation of new features for developing models for the early prediction of at-risk students in a technical distance learning course. The methodology to generate the models relies on the counting of interactions of the students inside the VLE, with the use of the proposed solution described in the previous section.

Data related to the student's interactions were collected from the logs of the institutional Moodle platform of a given technical distance course of the Instituto Federal Sul Rio-grandense (IFSul) in Brazil. Table 1 shows the number of logs collected, the number of students enrolled in the course, and the percentages of dropout and success. The course is taught in 18 different cities throughout the state of Rio Grande do Sul and involves weekly activities that are posted on the VLE by the teacher. Students have one week to develop the activities with the help of tutors. The course has a maximum completion time of 103 weeks, with a total workload of 1215 h divided into disciplines. The maximum duration is 24 months, with three breaks also called vacations, and the student's final situation is determined by their performance in the evaluations and their re-enrolment every six months.

The maximum term for completion of the curriculum is four years, and the student may repeat each discipline only once and, therefore, the year. The student has the option of taking up to two subjects for the next year and taking them concurrently with the others. For approval, the student must have a grade of six or higher in each of the disciplines of the curriculum. Students who spend 365 days without interactions with the virtual environment or do not perform their annual re-enrolment are considered absent and are removed from the course. Thus, the student receives a grade from 0 to 10 at the end of a given discipline, and one of two states is associated with the student: Approved or failed. However, we aimed to predict students who drop out during the course. For this, the student will be considered dropped out if they leave, do not perform the activities during the course, and their enrolment in the following semester.

**Table 1.** Dataset summary.

| Number of Log Rows | Number of Students | Dropouts (%) | Success (%) |
|---|---|---|---|
| 1,051,012 | 752 | 354 (47%) | 398 (53%) |

The choice to only use data from the counting of interactions was motivated by previous research that achieved satisfactory results using the same approach [23,25]. This choice was also related to limitations on capturing other kinds of data for the present study. In previous works, we sought to create models that are easy to generalize so that they could be applied to other courses. To accomplish

that, we used four courses, where the model created by one was applied to the others, and the models generated with data from three courses were applied to the remaining one. In these experiments, the labeling of the type of interaction was tested and did not show significant results. When testing the models generated with data from one course on data from other courses, this type of labeling negatively impacted the results.

Studies such as Macarini et al. [39] tested the application of different types of interactions and derived data, with their labeling showing no significant differences in performance. Thus, we applied the methodology that presented the best previous results to model other courses in the same educational context, even if the model is derived from data from one course only.

The courses studied here are offered in several cities throughout the interior of Brazil and present a large demographic diversity. Nowadays, the collection of demographic data is a task manually performed by eighteen different teaching centers through a printed questionnaire that is sent to IFSul after completion. This process generates a series of problems, such as lack of data, reading and typing problems, and consequently low diversity and inconsistencies. These factors led to the lack of reliability in these data and their consequent non-use.

Data capture consisted of collecting raw data from student interactions with Moodle VLE. The data initially had the format presented in Table 2. After selection, data were validated. This stage consisted of comparing the student situation data in the VLE to the data on the institutional academic system. Both systems are independent and have no integration. Cases of inconsistency were handled manually by checking other types of internal control.

**Table 2.** Information contained in the log files.

| Column | Comment |
|---|---|
| Course | Name of the virtual classroom accessed |
| Time | Day and time of the access |
| IP Address | IP Address of the machine |
| Full name | User (student) name |
| Action Event Name | The action represents the type of interaction that the student performed in the classroom. For instance:<br>(1) Visualization and participation on chats;<br>(2) Visualization and inclusion of posts in forums;<br>(3) Visualization of resources; and<br>(4) Visualization of the course. |
| Description | Detailed description of the event. Example: Download the .pdf file. |

The course format analyzed in this project consists of 103 weeks divided over two years. As stated by [5], early identification of a risk situation is a fundamental criterion for its reversal. Thus, for this work, we chose to use the methodology based on [4], which consists of the application of data mining on the data of the first subjects of the course. Using this process, we chose to use data from the 50 weeks that compose the first year of the course. Every two weeks starting from the fourth, a prediction model was generated, so the approaches used in this work created 23 models in the period.

After validation, data were anonymized and preprocessed, and variables were generated (features extraction). Table 3 describes the variables extracted to be used as the input for training and testing the predictive models. The table shows that all variables were based on the counting of students' interactions inside the VLE. Figure 2 exemplifies the behavior of the Weekly interactions variable for some weeks of the course and according to the Student Final Status category.

Exploratory data analysis (EDA) seeks to visualize dataset information to better understand the student's behavior when using the VLE. Table 4 shows how dropout rates evolved after every 10 weeks of the course until week 50. The table also shows the dropout rates for the first and second year of

the course after week 50. We considered a student as dropped out after a period of six weeks without interactions with the VLE. The idea here was to pinpoint the period where the departure occurs.

The evasion rates between the two years of the course are practically the same (182 dropouts for year 1 and 172 dropouts for year 2). However, if we look proportionally at the number of students enrolled at the beginning of each year, the dropout rate is slightly higher in the second year, with 30.06% compared with 24.20% in the first year. These values differ from the average dropout rates known from higher institutions in Brazil [44] as well as from secondary and technical schools [45]. Unfortunately, there are no national data related to the distance learning modality to enable a more precise comparison.

A total of 86.81% of the course dropouts of the first year are concentrated in the first 20 weeks (152 dropouts of the1 82 in the first year). This shows a tendency of the students to leave at the very beginning of the course, which could be related to difficulties faced in the initial studies. This tendency is also reported in the literature in relation to face-to-face courses where difficulties in the beginning of the course are reported as the most critical factor leading students to drop out.

Figure 2 presents the bi-weekly total count, the means, and the standard deviations of the students' interactions. In the figure, students identified as dropped out in a given week are not counted in the following weeks. As shown in the figure, dropout students present a higher number of interactions than successful students until week 13. One possible explanation for this behavior is that those students are experiencing difficulties during their learning process, so they interact more with the VLE to obtain assistance. The total count of interactions per group is lower for the dropout group (considering the whole period). Figure 3 presents a boxplot of the counting of interactions for each group of students, which highlights the differences in these groups regarding the use of the VLE.

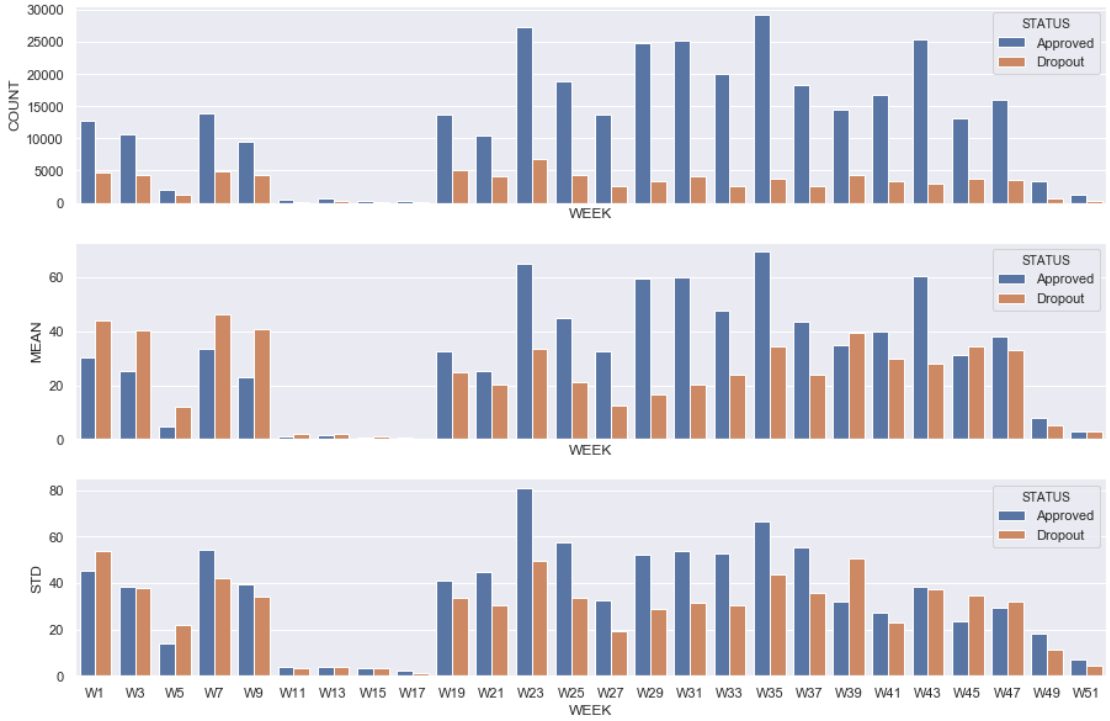


**Figure 2.** Interactions every two weeks.

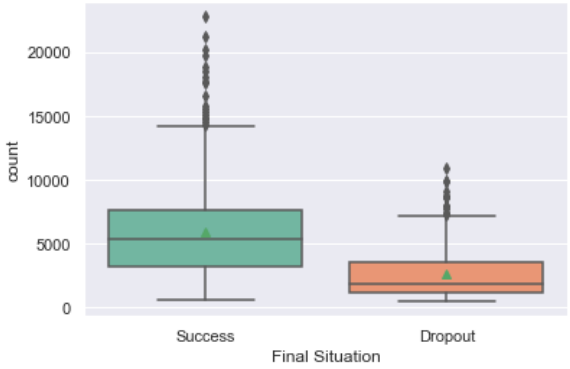


**Figure 3.** Boxplot success X dropout.

**Table 3.** Features extracted to be used as input for the models.

| Variable | Description |
|---|---|
| Daily interactions | Count of interactions of a given day (from 1 to 350 days) |
| Weekly interactions | Count of interactions of a given week (from 1 to 50 weeks) |
| Mean of the week | Average of the count of interactions of a given week |
| Standard deviation of the week | Standard deviation of the count of interactions of a given week |
| Student final status | Dependent variable representing the student final status: Dropout or success |

**Table 4.** Evolution of dropout during the course.

| Year | Period | Number of Students in Course | Number of Dropout Students (NDS) | NDS Rate | Accumulated NDS | Accmulated NDS Rate |
|---|---|---|---|---|---|---|
| Year 1 | Week 10 | 752 | 87 | 11.56 | 87 | 11.56 |
| | Week 20 | 665 | 71 | 10.67 | 158 | 21.01 |
| | Week 30 | 594 | 21 | 3.5 | 179 | 23.27 |
| | Week 40 | 573 | 1 | 0.17 | 180 | 23.4 |
| | Week 50 | 572 | 2 | 0.34 | 182 | 24.20 |
| | Total of First 50 Weeks | 752 | 182 | 24.20 | 182 | 24.20 |
| Year 2 | Total after 50 Weeks | 572 | 172 | 22.87 | 354 | 47.07 |
| Final Values | Total | 752 | 354 | 47.07 | 354 | 47.07 |

In Figure 4, the central diagonal presents the density plots of the Weekly Interactions variable for weeks 1, 10, 20, 30, and 40. The two groups of students (dropout and success) initially presented similar behavior at the beginning of the course (weeks 1 and 10), and gradually started to differ after week 20 when the number of weekly interactions of the successful students was slightly higher. The scatterplots help to better visualize the behavior of the interactions and their comparison between the weeks. The scatter plots demonstrate that there is no direct positive correlation between weeks. Students who were successful in the course tended to have more interactions, similar to that observed in Figure 2.

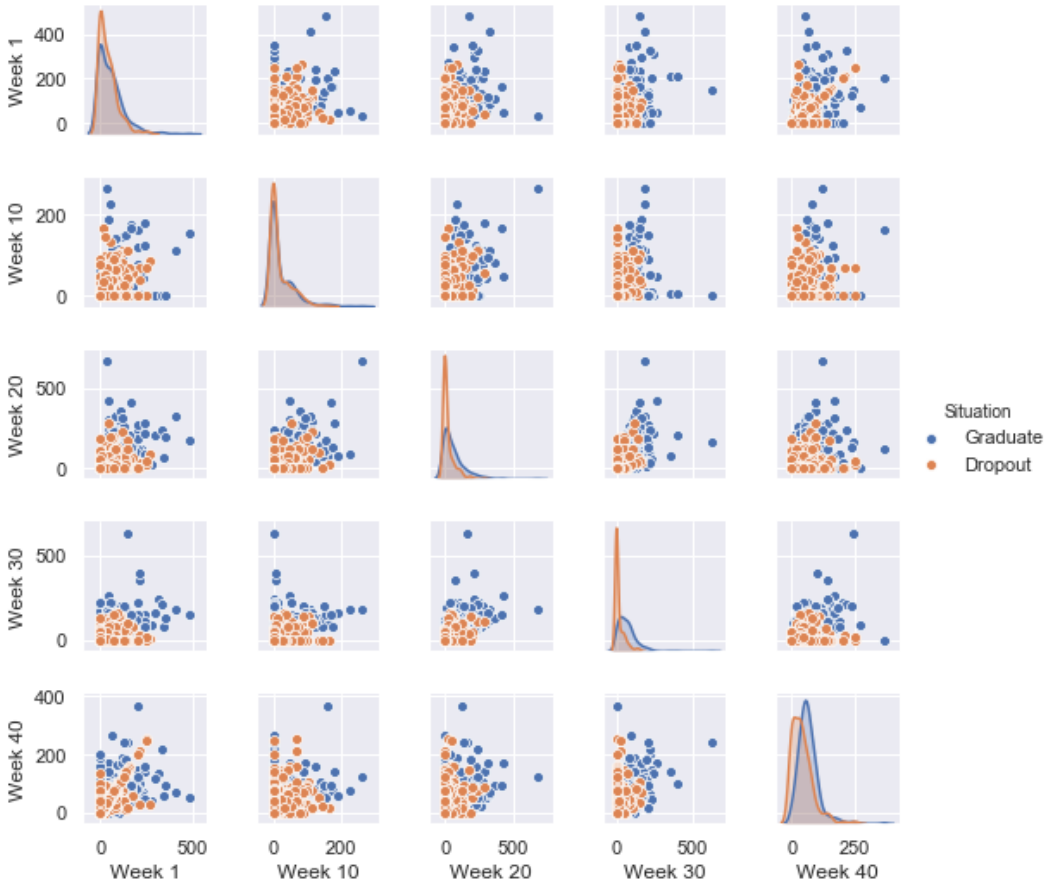


**Figure 4.** Density and scatter plots of weekly interactions.

### *3.2. Fine Tuning with Proposed Genetic Algorithm*

In the GA, the solution set is defined by a space where a search for an optimal solution occurs, which may not be the global best solution [46]. This factor is directly dependent on the problem, the time that can be spent searching, the expected result, and the input dataset, among others. These should be considered when the algorithm is designed [47]. In this work, a time-limited search approach is proposed, so the algorithm creates a number $N$ of generations, where $N$ is predefined at the time of configuration. In the end, the algorithm returns a solution with the setting that produced the best performance according to the predefined metric [48]. In this case, a learning machine model together with its hyperparameters were optimized for the prediction of students at risk in technical distance courses. As previously mentioned, this solution can be global or local. The steps of this process are presented in Figure 5.

The proposed approach is executed according to the general steps of classical GA solutions, which are: (1) Generate population, (2) fitness function, (3) selection, (4) crossover, and (5) mutation. For the context of our solution, the following definitions are provided:

(a) Epoch: One complete cycle execution of the GA (from Steps 1 to 5). The proposed approach works with 50 epochs;
(b) Individual (or candidate): A machine learning algorithm/classifier (DT, RF, MLP, LG, and ADA) together with its hyperparameters;
(c) Chromosome: A vector of hyperparameters for a given individual (machine learning algorithm). As different machine learning algorithms have different hyperparameters, the chromosomes in our study have different sizes and meaning according to the machine learning algorithm to which they are referring.

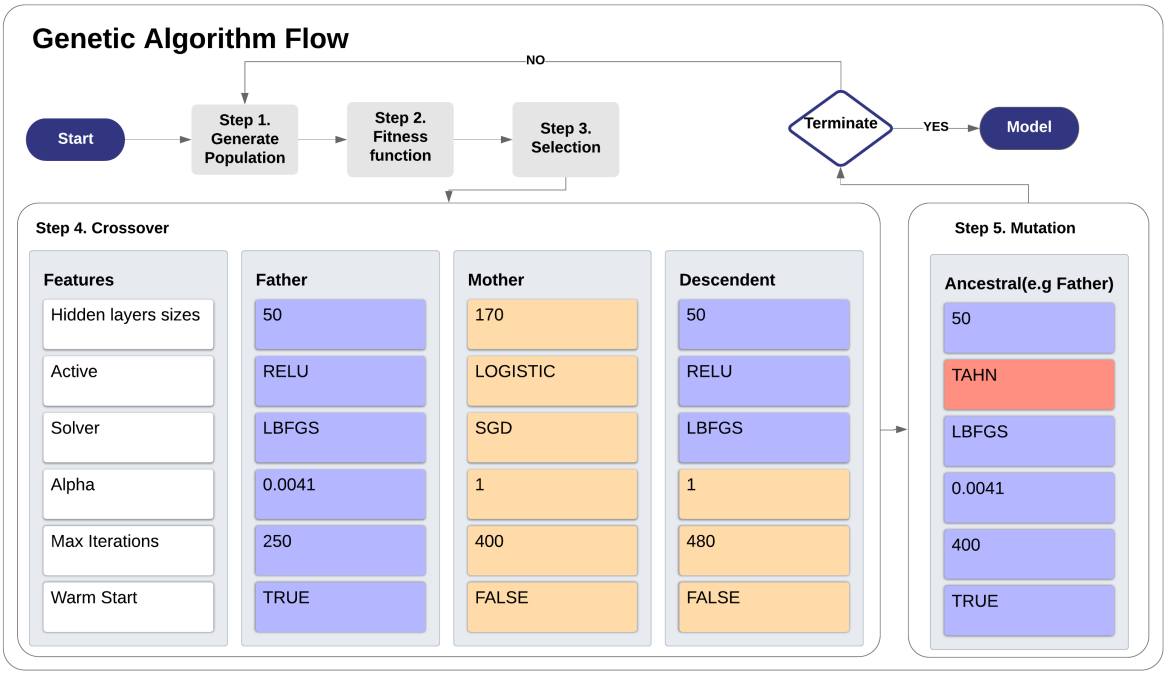


**Figure 5.** Genetic algorithm flow: An example of crossover and mutation with a multilayer perceptrons (MLP) chromosome.

Here, we outline each step of the process in the context of our proposed approach:

- Step 1 (generate population): The GA generates 100 individuals (candidates) for each machine learning algorithm (DT, RF, MLP, LG, and ADA) with hyperparameters (chromosomes) randomly defined considering the available list of options. The classifiers are trained and tested using 10-fold cross-validation and their performances are measured by using the area under the receiver operating characteristic curve metric (AUROC) [49] and as conducted by Gašević et al. [50].
- Step 2 (fitness function): The performance obtained by each of the 100 individuals of each machine learning algorithm are then compared by the fitness function.
- Step 3 (selection): The 25 individuals with the highest AUC for each machine learning algorithm are selected for the next step.
- Step 4 (crossover): The crossover is conducte using the concept based on the genetic inheritance of sexual reproductions, where each descendant receives a part of the genetic code (chromosome) of the father and part of the mother, as exemplified in Figure 5. Thus, the configurations of the fittest individuals of the last step are combined, one being the father and the other the mother. In the implemented algorithm, the individuals who will assign part of their genetic code to form a new member are chosen randomly from among the 25 best placed of that classifier in the last generation. This step results in 25 new individuals for each machine learning algorithm.
- Step 5 (mutation): This step randomly alters the chromosome (hyperparameter) of the 25 best individuals. In other words, a certain characteristic of an individual selected in the previous step receives a randomly generated configuration. As shown in Figure 5, an individual of the MLP type with hyperparameter "Active" set to "RELU" was changed to "TAHN". The mutation is set to change only one hyperparameter of the chromosome.

After Step 5, if the GA did not run the predefined number of epochs (50 for our experiment), a new population is generated in Step 1. The last important factor in generating a new population is randomness. For each generation, 25 new individuals are randomly generated again, even though they may have already been generated in earlier epochs. This seeks to ensure population diversity by narrowing the hypothesis that the solution reaches a local maximum and has no opportunity to evolve

to the global maximum. The quantitative formation of the population from the second epoch onwards for each machine learning algorithm is:

- 25 individuals selected from the previous generation from the fitness function (Steps 2 and 3);
- 25 individuals formed by crossover (Step 4);
- 25 individuals formed by mutations (Step 5); and
- 25 new individuals randomly generated (Step 1).

The process is repeated for 50 epochs. In the end, for each of the five machine learning algorithms, the individual with the highest aptitude (highest AUC) is selected. With the selection of the fittest for each machine learning algorithm, the five remaining individuals compete against themselves, and the one with the best AUC is selected.

### 3.3. Experiments

This section outlines the experiments with three different approaches to predict students at risk of dropout in the database described earlier. The first is the proposed genetic algorithm, the second is a grid search method called GridSearchCV, implemented using the Scikit-learn package. The third and last is the use of classifiers with their default hyperparameters. The machine learning techniques implemented in this study used the Python programming language with the Scikit-learn, Pandas, and Numpy libraries.

GridsearchCV allowed the testing of different combinations of hyperparameters for classifiers, facilitating choosing the best one. The hyperparameters needed to be explicitly declared and all possible combinations tested. All available combinations in Table 3 were checked with the same algorithms defined in the GA (DT, RF, MLP, LG, and ADA). For each week of the course, we selected the classifier together with its hyperparameters that achieved the best performance for the given week.

The same machine learning algorithms with their default hyperparameters were also implemented for comparison with the GA and GridsearchCV approaches. All experiments were performed with 10-fold cross-validation, and the number of combinations was approximately 5000 individuals created by the GA. Appendix A shows the quantities tested in each of the classifiers in the Evaluations column.

An essential task in machine learning is choosing the performance appraisal metric. For this work, we decided to use the area under the ROC curve, also known as AUC and AUROC. AUC is calculated from the size of the area under the plotted curve where the y-axis is represented by true positive rate (TPR) or sensitivity (Equation (1)), and the x-axis is true negative rate (TNR) or specificity (Equation (2)):

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

According to Gašević et al. [50], the AUC may be interpreted as follows:

- AUC $\leq$ 0.50: Bad discrimination;
- 0.50 < AUC $\leq$ 0.70: Acceptable discrimination;
- 0.70 < AUC $\leq$ 0.90: Excellent discrimination; and
- AUC > 0.90: Outstanding discrimination.

## 4. Results and Discussion

This section presents the results obtained by the models generated by each of the selected algorithms compared with the application of the GA. Table 5 presents the AUC results for each tested machine learning algorithm without hyperparameter optimization and for the grid search (GRID) and GA approaches.

**Table 5.** AUC for 50 weeks.

| Approach or Machine Learning Algorithm | Hyperparameter Optimization | AUC Mean | AUC Median | AUC Standard Deviation |
|---|---|---|---|---|
| GA | Yes | 0.8454 | 0.8498 | 0,0637 |
| GRID | | 0.7939 | 0.8288 | 0.1056 |
| ADA | No | 0.7509 | 0.8062 | 0.1342 |
| DT | | 0.6771 | 0.7065 | 0.1008 |
| LG | | 0.6943 | 0.7198 | 0.1110 |
| MLP | | 0.7353 | 0.7946 | 0.1277 |
| RF | | 0.7752 | 0.8243 | 0.1150 |

As can be seen from Table 5, the best AUC results were produced by the GA approach with a mean of 0.8454 and median of 0.8498. GA also produced the lowest AUC standard deviation (0.0637) among all tested approaches. Figure 6 helps visualize the performance of the models for the 50 weeks of the course.

To confirm the research hypothesis in this work (RQ1 and RQ2), two tests of statistical significance were applied. The objective of the tests was to verify if there was a significant difference in the treatments applied and, if so, which method was the most accurate. The central idea involved in the process of statistical significance is to test whether one treatment, in this study GA, presents a significant result concerning the others [51].

The results had a normal distribution, so analysis of variance (ANOVA) was chosen to verify the existence of a significant difference, and Tukey's test to determine in which treatment it occurred. For this, the *p*-value was set to 0.05; thus, values lower than this indicated that the treatment was significant and higher than not significant. In ANOVA, the *p*-value was 0.0006865, which reflects the existence of significant differences between the approaches. In Tukey's test, the results produced a *p*-value of 0.0475 for the GridSearch and 0.0003 for standard RF, indicating a statistically significant difference between the performance. Thus, statistically, the results obtained by GA were superior to the other treatments.

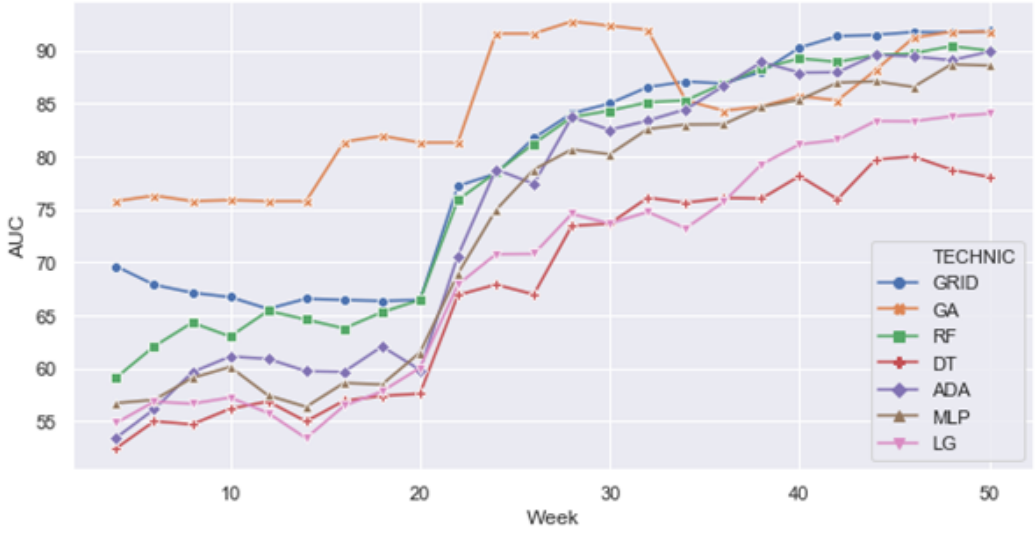


**Figure 6.** AUC results for each tested technique during the 50 weeks of the course.

The results obtained from the three approaches are presented in Figure 6. The GA achieved excellent discrimination (AUC > 0.7) as early as week 4. This held until week 24, where the GA

provides outstanding discrimination (AUC > 0.9). The other approaches still yielded acceptable discrimination results (AUC < 0.7) until week 22. However, from week 30, the performance of the GA considerably decreased, with the other approaches progressing. One of the factors determining this drop in GA performance was the increase in the number of input attributes. In this situation, the GA tends to find a local solution quickly and converge on it. However, this solution is probably a plateau and the GA getting stuck. This is a problem specific to genetic algorithms that does not occur in the other approaches tested. In the proposed algorithm, the reinsertion step tries to soften this, but as verified from weeks 32 to 42, the GA is still susceptible to this failure. However, the GA was considerably better in early prediction and with limited data. This demonstrates that the refinement of the GA is essential for tuning hyperparameters.

Table 6 presents the best configuration obtained by the GA approach for week 25 of the course (individual 37, fourth epoch), with an MLP with an AUC of 91.54, in comparison with the configuration for the same algorithm without hyperparameter optimization. From the first weeks of the courses, satisfactory results were already produced in the prediction of students at risk of dropout.

In general, the results of the models generated by GA per the AUC were satisfactory, allowing the prediction of at-risk students in the early stages of the courses. Data were naturally balanced, with similar percentages of dropout and success students. The models developed here produced similar or better results in comparison to some of the works in the literature that focused on the early prediction of dropout students.

**Table 6.** Comparison between configurations of models for week 25.

| Hyperparameter Optimization | Hidden Layer Sizes | Activation | Solver | Alpha | max Iter | Warm Start | AUC |
|---|---|---|---|---|---|---|---|
| yes | 30 | logistic | sgd | 0.2855486101 | 353 | False | 0.9154 |
| no | 100 | relu | adam | 0.0001 | 200 | False | 0.849 |

According to [31,52], one of the main factors involved in the acceptance of learning analytics by teachers and students when using prediction models is the correctness rates involved in the process. The GA proposed in this work was able to increase these rates compared to the results obtained in previous works [23,24]. However, direct comparison with these experiments is somewhat complicated, as they used the true positive (TP) and true negative (TN) of the models as metrics, and we used AUCROC.

In these previous experiments, the results obtained in scenarios similar to this experiment showed rates of TP and TN varying between 58 and 82 in the first 25 weeks of the course. However, with the approach proposed in this work, it was possible to reach an initial AUCROC above 0.75, which increased over the first 25 weeks until reaching values above 0.90.

The comparison with prominent works of predictors of educational environments is necessary to situate the results obtained. Some limitations for comparison include the various techniques used to measure the results, such as accuracy, TP, TN, AUC, and AUCROC, among others [32] Cechinel et al. [33] Liz-Domínguez et al. [30]. Still, a significant part of the works on LA are characterized by the exploration of data from disciplines of a specific course or semester, whereas the work presented in this paper is characterized by the use of data from a course of two years in duration [32]. However, even when we compare the results obtained with the related works, the rates are satisfactory. Previous studies Lykourentzou et al. [5], Zohair [37], Whitehill et al. [38] reported rates of 85% and Jayaprakash et al. [26] reported 94% overall accuracy, but only 66.7% dropout prediction.

The results obtained in the optimization with the proposed GA are close to those of the literature Minaei-Bidgoli and Punch [43], which obtained an optimization of 12%. The proposed GA was able to reach values above 10% in the experiments until the 20th week compared to the algorithms in their standard configuration. When compared to the other optimization method, Gridsearch, in that same period, GA obtained values always above 6%, sometimes reaching 15%.

The method followed here is the result of an incremental process of a series of experiments previously performed [23,24]. As such, when comparing the results achieved in this work with the results from previous actions, the hyperparameters generated by the GA allows the generation of more robust models and higher performance. This is also demonstrated in comparison to the other methods tested in this article. Thus, the methodology used both for the development of the GA and for the generation of input data from genetic algorithms demonstrated that it could be used for early prediction of students at risk of dropout. Concerning data modeling, although the use of interaction count is not unprecedented, the methodology used in this study has several attributes that produced the results.

## 5. Final Remarks

This paper presented the results of an approach for the early prediction of students at risk of dropout using the counting of their interactions inside the VLE. This approach uses genetic algorithms for the hyperparameter of classifiers. The methodology of generating a prediction model every two weeks allows every student to be followed throughout the course. This is an approach that differs from the traditional methods [6] that define models that seek to predict dropout using all available data at the end of the course. This difference and, consequently, the results obtained with smaller amounts of data contribute to the early prediction of the risk of dropout.

The proposed approach is based on the premise of allowing greater generalization when replicating the methodology in other courses and platforms, since it only uses the count of interactions within the VLE without distinguishing the types of actions performed and without using information from different data sources (demographic data, questionnaires, curriculum, etc.), the availability of which may vary between e-learning platforms. The results can be considered satisfactory since they allow the identification of students at risk of dropout with reasonable performance rates even before the end of the first semester of the course.

The prediction of academic issues, such as performance and dropout, is concentrated at the university level, with about 70% of the research destined for this purpose [10]. This trend is repeated in Latin America, with few applications considering the context of education at the secondary and technical levels [33]. While not unprecedented, the application of prediction techniques in other contexts, such as technical high school e-learning, is also relevant [10].

**RQ1.** Does the approach for hyperparameter optimization with a GA outperform traditional techniques?

The proposed GA must be evaluated to emphasize that testing different combinations of hyperparameters within the same algorithm is a complicated and time-consuming task that may require a large amount of processing time. However, the accuracy of prediction models is directly linked to the quality of hyperparameter optimization. Thus, the more adjusted they are, the more accurate the rates of the models tend to be. The alternatives to applying exhaustive search methods, such as grid-search, are computationally expensive when searching in large spaces [53]. Thus, the refinement obtained by GA with its mutation and crossover stages produces better results for model generation, surpassing the traditional techniques and grid-search. Compared to standard algorithms, the performance of the proposed method is clearly superior.

**RQ2.** Does the resulting predictive models generated by the use of the GA approach for hyperparameter optimization perform better than models with default hyperparameters?

In comparison with the classifications using the default hyperparameters, the GA produced significantly better results. In the first 20 weeks of the course, the difference between the two methods varies from 10% and 15%. Tukey's test demonstrated that the overall values obtained are significantly different. However, all techniques have advantages and limitations. The drawback of the GA is the lack of assurance that the solution is global; the positive aspect is the number of resultant

hyperparameters accepted without significantly altering the processing cost and the final results. In grid-search, the computational cost is the biggest issue, as previously reported; however, it delivers the best possible combination of hyperparameters. Concerning the standard classifiers, we highlight the cost–benefit factor as the method produces satisfactory results in short processing time, which, depending on the project, can be an essential point.

The main limitation of the proposed methodology presents is that for each course analyzed, the calendar must be studied to identify periods without classes, such as holidays. This causes extra work, which does not occur when socio-demographic data are used. Another limitation concerns generalization; although the methodology may be generalized, the models are unlikely to be suitable for courses that do not follow the same timetable as ETec. Models that seek long-term predictions are more susceptible to failures due to external situations, such as economic and epidemiological crises.

An important point to note is that the GA possibly presents slightly different results for each execution. Thus, it may be interesting to run the GA multiple times (e.g., 10). Analysis of other metrics, such as overall accuracy and true positive (TP) and true negative (TN), may provide different perspectives. The application of other hyperparameter search methods, such as random search, and algorithms, such as XGBOOST, can still be explored. These questions will possibly be studied in the future stages of this project, as well as hybrid choice methods such as the vote theory, for final classification selection.

The results obtained in this work enable the development of an early warning system using the proposed approach. Currently, the development of this system is occurring in the form of a plugin integrated with Moodle. Another future work toward improving the results is the application of survival analysis to increase student retention and consequently reduce dropout.

**Author Contributions:** Author Contributions: E.M.Q.: Experimental data analysis, algorithms development, data pre-processing, experiments conduction, results description, and manuscript writing; J.L.L.: Course lecturer, conceived and designed, and methodology definition; K.K.: Algorithms development and writing review; M.A.: manuscript writing and algorithms development; R.M.A.: Algorithms development and data pre-processing; R.V.: writing—review and editing; R.M.: writing—review and editing. C.C.: methodology definition, algorithm developed, experiments setup, manuscript writing, and writing—review and editing. The final manuscript was written and approved by all authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by CNPq (Brazilian National Council for Scientific and Technological Development) [Edital Universal, proc.404369/2016-2][DT-2 Productivity in Technological Development and Innovative Extension scholarship, proc.315445/2018-1]. R.V. and R.M. were funded by Corporación de Fomento de la Producción (CORFO) 14ENI2-26905 "Nueva Ingeniería para el 2030"—Pontificia Universidad Católica de Valparaíso, Chile.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ADA | ADABoost |
| ANOVA | Analysis of Variance |
| AUC | Area Under the Curve |
| AUROC | Area Under the Receiver Operating Characteristic Curve |
| DT | Decision Tree |
| EDA | Exploratory Data Analysis |
| EDM | Educational Data Mining |
| GA | Genetic Algorithm |
| GBGP | Grammar-Based Genetic Programming |
| GRID | Grid SearchCV |
| IBK | Instance-Based Lazy Learning |
| ICRM | Interpretable Classification Rule Mining Algorithm |
| IFSul | Instituto Federal Sul Rio-grandense |

| | |
|---|---|
| INN | I-nearest neighbor |
| kNN | k-Nearest Neighbor |
| LA | Learning Analytics |
| LG | Logistic Regression |
| LMS | Learning Management Systems |
| ML | Machine Learning |
| MAE | Mean Absolute Error |
| MLP | Multilayer Perceptron |
| NDS | Number of Dropout Students |
| PCA | Principal Component Analysis |
| RQ | Research Question |
| RF | Random Forest |
| SAT | Scholastic Aptitude Test |
| SMOTE | Synthetic Minority Over-Sampling Technique |
| SVM | Support Vector Machine |
| TNR | True Negative Rate |
| TPR | True Positive Rate |
| VLE | Virtual Learning Environment |

## Appendix A

**Table A1.** Classifiers, Hyperparameters and Number of Evaluations.

| Alg. | Hyperparameters | Possibi-Lities | Number of Ind. | Grid | Eval. |
|---|---|---|---|---|---|
| DT | criterion: [gini, entropy], max_depth: range (0, 32), min_samples_split: range (1, 15), min_samples_leaf: range (1, 20) | 19.200 | 5.100 | criterion: [gini,entropy], max_depth: [0, 1, 2, 3, 5, 7, 10, 12, 15, 17, 20, 23, 25, 30], min_samples_split: [0, 1, 2, 3, 5, 7, 10, 12, 15] min_samples_leaf: [0, 1, 2, 3, 4, 5, 7, 9, 10, 12, 15, 17, 20] | 3.726 |
| RF | n_estimators: range (1,200), criterion: [gini, entropy], max_features [1, 2, 3,4], min_samples_split: range (2, 21), min_samples_leaf: range (1, 2), bootstrap: [True, False] | 128.000 | 5.100 | n_estimators: [1, 10, 20, 30, 40, 50, 70, 100, 120, 130, 150, 170, 190, 200], criterion: [gini, entropy], max_features [1, 2, 3, 4], min_samples_split: [2, 3, 4, 5, 7, 9, 10, 12, 15, 17, 20], min_samples_leaf: [1, 2], bootstrap: [True, False] | 4.928 |
| ADA | algorithm: [SAMME, SAMME.R], n_estimators: range (1, 200), random_state: range (None, 50), learning_rate: range (1e-2,1) | 2 KK | 5.100 | algorithm: [SAMME, SAMME.R], n_estimators: [1, 10, 20, 30, 40, 50, 70, 100, 120, 130, 150, 170, 190, 200], random_state: [None, 1, 5, 10, 15, 20, 25, 30, 40, 50 ], learning_rate: [1e-2, 5e-2, 7e-2, 1e-1, 3e-1, 5e-1, 7e-1, 1] | 2.240 |
| MLP | hidden_layer_sizes: range (1,200), activation: [identity, logistic, tanh, relu], solver: [lbfgs, sgd, adam], max_iter: range (50, 200), alpha: range (1e-4, 1e-1), warm_start: [True, False] | 720 KK | 5.100 | hidden_layer_sizes: [(50,50,50), (50,100,50), (100,), (50,), (10,), (1), (5)], activation: [identity, logistic, tanh, relu], solver: [lbfgs, sgd, adam], max_iter: [1, 2, 5, 10, 30, 50], alpha: [1e-4, 1e-3, 1e-2, 5e-2, 1e-1], warm_start: [True, False] | 5.040 |
| RL | penalty: [l1, l2, elasticnet], C: [1e-4, 1e-3, 1e-2, 1e-1, 5e-1, 1, 5, 10, 15, 20, 25], dual: [True, False], solver: [newton-cg, lbfgs, lbfgs, sag, saga], multi_class: [ovr, auto], max_iter: range (50,200) | 99.000 | 5.100 | penalty: [l1, l2, elasticnet], C: [1e-4, 1e-1, 5e-1, 1, 5, 15, 25], dual: [True, False], solver: [newton-cg, lbfgs, lbfgs, sag, saga], multi_class: [ovr, auto], max_iter: [1, 10, 20, 30, 40, 50, 70, 100, 120, 130, 150, 170, 190, 200] | 5.800 |

## References


1. Chatti, M.A.; Dyckhoff, A.L.; Schroeder, U.; Thüs, H. A reference model for learning analytics. *Int. J. Technol. Enhanc. Learn.* **2013**, *4*, 318–331. [CrossRef]
2. Siemens, G. Learning analytics: The emergence of a discipline. *Am. Behav. Sci.* **2013**, *57*, 1380–1400. [CrossRef]
3. Sheehan, M.; Park, Y. pGPA: A personalized grade prediction tool to aid student success. In Proceedings of the Sixth ACM Conference on Recommender Systems, Dublin City, Ireland, 9–13 September 2012; pp. 309–310.
4. Manhães, L.M.B.; Cruz, S.d.; Costa, R.J.M.; Zavaleta, J.; Zimbrão, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In Proceedings of the Anais do XXII SBIE-XVII WIE, Aracaju, Brazil, 21–25 November 2011.
5. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* **2009**, *53*, 950–965. [CrossRef]
6. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Mousa Fardoun, H.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]
7. OECD. *Benchmarking Higher Education System Performance*; OECD: Paris, France, 2019; p. 644. [CrossRef]
8. Yukselturk, E. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *Comput. Educ.* **2014**, *17*, 118–133. [CrossRef]
9. Li, Q.; Baker, R.; Warschauer, M. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *Internet High. Educ.* **2020**, 100727. [CrossRef]
10. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Appl. Sci.* **2020**, *10*, 1042. [CrossRef]
11. Vossensteyn, J.J.; Kottmann, A.; Jongbloed, B.W.; Kaiser, F.; Cremonini, L.; Stensaker, B.; Hovdhaugen, E.; Wollscheid, S. Dropout and Completion in Higher Education in Europe: Main Report. In *European Commission*; Center for Higher Education Policy Studies and Nordic Institute for Studies in Innovation Research and Education: Enschede, The Nerthland, 2015; doi:10.2766/826962. 2015. [CrossRef]
12. Gregori, E.B.; Zhang, J.; Galván-Fernández, C.; Fernández-Navarro, F.d.A. Learner support in MOOCs: Identifying variables linked to completion. *Comput. Educ.* **2018**, *122*, 153–168. [CrossRef]
13. Censo, E. BR 2018-Relatório Analítico da Aprendizagem a Distância no Brasil. *Acesso Em* **2018**, *16*.
14. Dickson, W.P. Toward a deeper understanding of student performance in virtual high school courses: Using quantitative analyses and data visualization to inform decision making. In *A Synthesis of New Research in K–12 Online Learning*; Michigan Virtual University: Lansing, MI, USA, 2005; pp. 21–23.
15. Murray, M.; Pérez, J.; Geist, D.; Hedrick, A. Student interaction with content in online and hybrid courses: Leading horses to the proverbial water. In Proceedings of the Informing Science and Information Technology Education Conference, Santa Rosa, CA, USA, 30 June–6 July 2013; Informing Science Institute: Santa Rosa, CA, USA, 2013; pp. 99–115.
16. Leitner, P.; Ebner, M.; Ebner, M. Learning Analytics Challenges to Overcome in Higher Education Institutions. In *Utilizing Learning Analytics to Support Study Success*; Springer: Berlin, Germany, 2019; pp. 91–104.
17. Gursoy, M.E.; Inan, A.; Nergiz, M.E.; Saygin, Y. Privacy-preserving learning analytics: Challenges and techniques. *IEEE Trans. Learn. Technol.* **2016**, *10*, 68–81. [CrossRef]
18. Drachsler, H.; Greller, W. Privacy and analytics: It's a DELICATE issue a checklist for trusted learning analytics. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, Scotland, 25–29 April 2016; pp. 89–98.
19. Baker, R.S.; Inventado, P.S. Educational data mining and learning analytics. In *Learning Analytics*; Springer: Berlin, Germany, 2014; pp. 61–75.
20. Olivares, R.; Munoz, R.; Soto, R.; Crawford, B.; Cárdenas, D.; Ponce, A.; Taramasco, C. An Optimized Brain-Based Algorithm for Classifying Parkinson's Disease. *Appl. Sci.* **2020**, *10*, 1827. [CrossRef]
21. Bergstra, J.S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Granada, Spain, 2011; pp. 2546–2554, ISBN 9781618395993.

22. Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **2017**, *18*, 6765–6816.
23. Queiroga, E.; Cechinel, C.; Araújo, R. Predição de estudantes com risco de evasão em cursos técnicos a distância. In Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), Recife, Brazil, 30 October–2 November 2017; p. 1547.
24. Queiroga, E.; Cechinel, C.; Araújo, R.; da Costa Bretanha, G. Generating models to predict at-risk students in technical e-learning courses. In Proceedings of the IEEE Latin American Conference on Learning Objects and Technology (LACLO), San Carlos, CA, USA, 3–7 October 2016; pp. 1–8.
25. Detoni, D.; Cechinel, C.; Matsumura Araújo, R. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. *Revista Brasileira de Informática na Educação* **2015**, *23*, 1.
26. Jayaprakash, S.M.; Moody, E.W.; Lauria, E.J.M.; Regan, J.R.; Baron, J.D. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *J. Learn. Anal.* **2014**, *1*, 6–47. [CrossRef]
27. Márquez-Vera, C.; Cano, A.; Romero, C.; Ventura, S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl. Intell.* **2013**, *38*, 315–330. [CrossRef]
28. Xing, W.; Guo, R.; Petakovic, E.; Goggins, S. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Comput. Hum. Behav.* **2015**, *47*, 168–181. [CrossRef]
29. Munoz, R.; Olivares, R.; Taramasco, C.; Villarroel, R.; Soto, R.; Barcelos, T.S.; Merino, E.; Alonso-Sánchez, M.F. Using black hole algorithm to improve eeg-based emotion recognition. *Comput. Intell. Neurosci.* **2018**, *2018*, 22. [CrossRef]
30. Liz-Domínguez, M.; Caeiro-Rodríguez, M.; Llamas-Nistal, M.; Mikic-Fonte, F.A. Systematic Literature Review of Predictive Analysis Tools in Higher Education. *Appl. Sci.* **2019**, *9*, 5569. [CrossRef]
31. Herodotou, C.; Rienties, B.; Verdin, B.; Boroowa, A. Predictive learning analytics 'at scale': Towards guidelines to successful implementation in Higher Education based on the case of the Open University UK. *J. Learn. Anal.* **2019**. [CrossRef]
32. Hilliger, I.; Ortiz-Rojas, M.; Pesántez-Cabrera, P.; Scheihing, E.; Tsai, Y.S.; Muñoz-Merino, P.J.; Broos, T.; Whitelock-Wainwright, A.; Pérez-Sanagustín, M. Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. *Internet High. Educ.* **2020**, *45*, 100726. [CrossRef]
33. Cechinel, C.; Ochoa, X.; Lemos dos Santos, H.; Carvalho Nunes, J.B.; Rodés, V.; Marques Queiroga, E. Mapping Learning Analytics initiatives in Latin America. *Br. J. Educ. Technol.* **2020**, doi:10.1111/bjet.12941. [CrossRef]
34. González, P. Factores que favorecen las presencia docente en entornos virtuales de aprendizaje. *Tendencias Pedagógicas* **2017**, *29*, 43–58.
35. de Pablo González, G. *La Importancia de la Presencia Docente en Entornos Virtuales de Aprendizaje*; Universidad Autónoma de Madrid: Madrid, Spain, 2016.
36. Herodotou, C.; Rienties, B.; Boroowa, A.; Zdrahal, Z.; Hlosta, M.; Naydenova, G. Implementing predictive learning analytics on a large scale: The teacher's perspective. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13–17 March 2017; pp. 267–271.
37. Zohair, L.M.A. Prediction of Student's performance by modelling small dataset size. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 27. [CrossRef]
38. Whitehill, J.; Mohan, K.; Seaton, D.; Rosen, Y.; Tingley, D. Delving deeper into MOOC student dropout prediction. *arXiv* **2017**, arXiv:1702.06404.
39. Macarini, B.; Antonio, L.; Cechinel, C.; Batista Machado, M.F.; Faria Culmant Ramos, V.; Munoz, R. Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Appl. Sci.* **2019**, *9*, 5523. [CrossRef]
40. Queiroga, E.; Cechinel, C.; Araújo, R. Um Estudo do Uso de Contagem de Interações Semanais para Predição Precoce de Evasão em Educação a Distância. In Proceedings of the Anais dos Workshops do Congresso Brasileiro de Informática na Educação, Maceio, Brazil, 26–30 October 2015; p. 1074.
41. Swan, K. Learning effectiveness online: What the research tells us. *Elem. Qual. Online Educ. Pract. Dir.* **2003**, *4*, 13–47.

42. Halawa, S.; Greene, D.; Mitchell, J. Dropout Prediction in MOOCs using Learner Activity Features. *Eur. MOOC Summit EMOOCs* **2014**, *37*, 1–10.
43. Minaei-Bidgoli, B.; Punch, W.F. Using genetic algorithms for data mining optimization in an educational web-based system. In Proceedings of the Genetic and eVolutionary Computation Conference, Chicago, IL, USA, 12–16 July 2003; Springer: Berlin, Germany, 2003; pp. 2252–2263.
44. Silva Filho, R.L.L.; Motejunas, P.R.; Hipólito, O.; Lobo, M.B.d.C.M. A evasão no ensino superior brasileiro. *Cadernos de Pesquisa* **2007**, *37*, 641–659. [CrossRef]
45. Resende, M.L.d.A. *Evasão Escolar No Primeiro Ano Do Ensino médio Integrado Do Ifsuldeminas-Campus Machado*; Encontro Anual da ANPOCS: Caxambu, Brazil, 2012.
46. Fonseca, C.M.; Fleming, P.J. Genetic Algorithms for Multiobjective Optimization: Formulation Discussion and Generalization. In Proceedings of the ICGA, San Mateo, CA, USA, 17–22 June 1993; pp. 416–423.
47. Hartmann, S. A competitive genetic algorithm for resource-constrained project scheduling. *Nav. Res. Logist. (NRL)* **1998**, *45*, 733–750. [CrossRef]
48. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR)* **2002**, *34*, 1–47. [CrossRef]
49. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
50. Gašević, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet High. Educ.* **2016**, *28*, 68–84. [CrossRef]
51. Bruce, P.; Bruce, A. *Practical Statistics for Data Scientists: 50 Essential Concepts*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
52. Larrabee Sønderlund, A.; Hughes, E.; Smith, J. The efficacy of learning analytics interventions in higher education: A systematic review. *Br. J. Educ. Technol.* **2019**, *50*, 2594–2618. [CrossRef]
53. Zöller, M.A.; Huber, M.F. Survey on automated machine learning. *arXiv* **2019**, arXiv:1904.12054.