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

Tese

Análise Preditiva do Nível de Água no Canal São Gonçalo Utilizando Modelos de Aprendizado de Máquina: Uma Abordagem Integrativa para a Gestão Hídrica

Paulo Ricardo Barbieri Dutra Lima Lima

Pelotas, 2024

Paulo Ricardo Barbieri Dutra Lima Lima

Análise Preditiva do Nível de Água no Canal São Gonçalo Utilizando Modelos de Aprendizado de Máquina: Uma Abordagem Integrativa para a Gestão Hídrica

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

Orientador: Prof. Dr. Felipe Marques

Pelotas, 2024

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

L732a Lima, Paulo Ricardo Barbieri Dutra

Análise preditiva do nível de água no Canal São Gonçalo utilizando modelos de aprendizado de máquina [recurso eletrônico] : uma abordagem integrativa para a gestão hídrica / Paulo Ricardo Barbieri Dutra Lima ; Felipe Marques, orientador. — Pelotas, 2024.
145 f. : il.

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

1. Modelagem preditiva. 2. Séries temporais climatológicas. 3. Aprendizagem de máquina. 4. Modelo híbridos. I. Marques, Felipe, orient. II. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

Paulo Ricardo Barbieri Dutra Lima Lima

Análise Preditiva do Nível de Água no Canal São Gonçalo Utilizando Modelos de Aprendizado de Máquina: Uma Abordagem Integrativa para a Gestão Hídrica

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

Data da Defesa: 23 de Maio de 2024

Banca Examinadora:

Prof. Dr. Felipe Marques (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

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

Prof. Dr. Frederico Schmitt Kremer
Doutor em Biotecnologia pela Universidade Federal de Pelotas.

Prof. Dr. Gilberto Loguercio Collares
Doutor em Ciência do Solo pela Universidade Federal de Santa Maria.

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão à minha família por seu apoio inabalável ao longo desta jornada acadêmica. Seu incentivo e compreensão foram fundamentais para me manter focado e motivado durante os momentos desafiadores.

Também gostaria de expressar minha sincera gratidão ao meu orientador, Professor Felipe Marques, pela sua orientação, sabedoria e apoio ao longo deste projeto. Suas percepções, conselhos e *feedbacks* foram inestimáveis e contribuíram significativamente para o desenvolvimento deste trabalho acadêmico.

Adicionalmente, gostaria de estender meus agradecimentos ao Professor George Martino que contribuiu para com este trabalho. Suas discussões, sugestões e apoio foram valiosos e enriqueceram significativamente a qualidade deste projeto.

À medida que concluo esta etapa da jornada acadêmica, quero expressar minha profunda gratidão a todos os que estiveram ao meu lado, especialmente à minha família e ao meu orientador. Suas contribuições foram essenciais para o meu crescimento pessoal e profissional, e por isso, serei eternamente grato. Muito obrigado a todos que me ajudaram ao longo deste caminho.

RESUMO

LIMA, Paulo Ricardo Barbieri Dutra Lima. **Análise Preditiva do Nível de Água no Canal São Gonçalo Utilizando Modelos de Aprendizado de Máquina: Uma Abordagem Integrativa para a Gestão Hídrica**. Orientador: Felipe Marques. 2024. 148 f. Tese (Doutorado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

No âmbito da hidrologia, modelos de inteligência artificial têm-se destacado como ferramentas eficazes em diversas pesquisas. A previsão precisa do nível da água em bacias hidrográficas e rios desempenha um papel crucial nas estratégias de prevenção de inundações, gestão da navegação interior e garantia do abastecimento doméstico de água. Contudo, uma análise abrangente da aplicabilidade desses modelos, especialmente no contexto da previsão do nível de água em conjuntos de dados vastos, tem sido escassamente explorada. Este estudo investiga e compara o desempenho de diferentes modelos de inteligência artificial, incluindo redes neurais densas, redes neurais recorrentes, floresta aleatória e regressão de vetor de suporte, na previsão do nível de água. Além disso, são propostas arquiteturas híbridas visando otimizar a precisão preditiva. A avaliação prática desses modelos é realizada por meio de um estudo de caso no Canal São Gonçalo. O estudo abrange quatro análises distintas, cada uma baseada em conjuntos de dados elaborados com a colaboração de Engenheiros Hídricos especializados. Além da avaliação comparativa de modelos de inteligência artificial na previsão do nível de água, este trabalho apresenta uma contribuição inovadora ao introduzir arquiteturas híbridas que unem o potencial do modelo ARIMA com abordagens tradicionais de aprendizagem de máquina. A integração desses métodos busca capitalizar as vantagens intrínsecas de ambos, promovendo uma abordagem combinatória para otimizar a precisão e robustez das previsões hidrológicas. Este enfoque híbrido reflete uma perspectiva pouco explorada no desenvolvimento de estratégias avançadas para antecipação do nível de água, sugerindo novas possibilidades para aprimorar a eficiência dos sistemas de gestão hídrica. O estudo destaca, assim, uma dimensão inovadora no campo, contribuindo para o avanço do conhecimento e práticas na área de previsão hidrológica. Outra contribuição significativa desta tese foi a meticulosa coleta e análise dos dados climatológicos e de nível de água, que serviram como base para a elaboração de quatro conjuntos de dados distintos, cada um caracterizado por suas particularidades. Esse esforço não apenas ampliou a compreensão do cenário climatológico abordado, mas também proporcionou a criação de conjuntos de dados representativos e diversificados. A cuidadosa seleção e manipulação desses conjuntos de dados permitiram uma abordagem mais abrangente e aprofundada na

modelagem preditiva, enriquecendo a pesquisa com percepções valiosas sobre a dinâmica complexa entre os fatores climáticos e os níveis de água. Essa metodologia multidimensional fortalece a fundamentação teórica e prática da tese, contribuindo para a robustez e relevância dos resultados obtidos. Os resultados obtidos revelam que os modelos híbridos demonstram uma performace preditiva superior em comparação com abordagens isoladas. Esta constatação reforça a viabilidade e eficácia das técnicas de aprendizado de máquina na esfera hidrológica, destacando seu potencial como ferramentas de suporte à tomada de decisões em contextos relacionados à gestão de recursos hídricos.

Palavras-chave: Modelagem Preditiva, Séries Temporais Climatológicas, Aprendizagem de Máquina, Modelo Híbridos

ABSTRACT

LIMA, Paulo Ricardo Barbieri Dutra Lima. **Predictive Analysis of Water Level in São Gonçalo Channel Using Machine Learning Models: An Integrative Approach to Water Management**. Advisor: Felipe Marques. 2024. 148 f. Thesis (Doctorate in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2024.

In the field of hydrology, artificial intelligence models have stood out as effective tools in various research efforts. Accurate prediction of water levels in watersheds and rivers plays a crucial role in flood prevention strategies, inland navigation management, and ensuring domestic water supply. However, a comprehensive analysis of the applicability of these models, especially in the context of predicting water levels in vast datasets, has been scantily explored.

This study investigates and compares the performance of various artificial intelligence models, including dense neural networks, recurrent neural networks, random forest, and support vector regression, in predicting water levels. Additionally, hybrid architectures are proposed to optimize predictive accuracy. The practical evaluation of these models is conducted through a case study in the São Gonçalo Channel. The study encompasses four distinct analyses, each based on datasets developed in collaboration with specialized water engineers.

Beyond the comparative evaluation of artificial intelligence models in water level prediction, this work presents an innovative contribution by introducing hybrid architectures that combine the potential of the ARIMA model with traditional machine learning approaches. The integration of these methods aims to capitalize on the intrinsic advantages of both, promoting a combinatorial approach to optimize the accuracy and robustness of hydrological predictions. This hybrid approach reflects a less-explored perspective in the development of advanced strategies for anticipating water levels, suggesting new possibilities to enhance the efficiency of water management systems. The study thus highlights an innovative dimension in the field, contributing to the advancement of knowledge and practices in hydrological forecasting.

Another significant contribution of this thesis was the meticulous collection and analysis of climatological and water level data, serving as the basis for the development of four distinct datasets, each characterized by its specificities. This effort not only expanded the understanding of the addressed climatological scenario but also facilitated the creation of representative and diversified datasets. The careful selection and manipulation of these datasets allowed for a more comprehensive and in-depth approach to predictive modeling, enriching the research with valuable insights into the complex

dynamics between climatic factors and water levels. This multidimensional approach strengthens the theoretical and practical foundation of the thesis, contributing to the robustness and relevance of the obtained results.

The results obtained reveal that hybrid models demonstrate superior predictive performance compared to isolated approaches. This finding reinforces the viability and effectiveness of machine learning techniques in the hydrological sphere, highlighting their potential as decision support tools in contexts related to water resource management.

Keywords: Predictive Modeling, Climatological Time Series, Machine Learning, Hybrid Models

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	*Artificial Intelligence*
ALM	Agência da Lagoa Mirim
ANN	*Artificial Neural Network*
ARIMA	*AutoRegressive Integrated Moving Average*
BHMSG	Bacia Hidrográfica Mirim-São Gonçalo
BP	*Back Propagation*
DNN	*Deep Neural Network*
DT	*Decision Tree*
FAONU	Organização das Nações Unidas para Agricultura e Alimentação
FFN	*Feed Forward Network*
INMET	Instituto Nacional de Meteorologia
GP	*Process Gaussian*
INIA	Instituto Nacional de Investigación Agropecuária
KNN	*K-nearest neighbors*
LSTM	*Long short-term memory*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
MLP	*Multi Layer Perceptron*
MLR	*Regression Linear Multiple*
MSE	*Mean Squeared Error*
PNUD	Programa das Nações Unidas para o Desenvolvimento
RMSE	*Root Mean Squeared Error*
RNN	*Recurrent neural network*
RF	*Random Forest*
SGD	*Stochastic Gradient Descent*
SI	Sistema Internacional de Unidades

SVM *Suport Vector Machine*

SVR *Suport Vector Regression*

SUMÁRIO

1 INTRODUÇÃO

O uso dos recursos hídricos de rios por meio de reservatórios e barragens é de fundamental importância para a geração de energia, abastecimento de água, navegação e controle de inundações. Mais da metade dos principais sistemas fluviais do mundo possuem reservatórios represados, que controlam ou afetam o fluxo de um determinado rio (Joo; Kim, 2015). A gestão de reservatórios de água em rios e lagoas é, portanto, um problema crítico. Existem diversas tarefas em que a previsão do nível de água represada no reservatório é um fator preocupante, tais como: para avaliar problemas estruturais em barragens, abastecimento de água e disponibilidade de recursos, qualidade da água, bio conservação da diversidade, gestão da navegação, prevenção de desastres e otimização da produção de energia hidrelétrica.

A previsão do nível da água desempenha um papel importante no bem-estar e na subsistência econômica de uma comunidade. Mudanças no nível de água podem impulsionar a ocorrência de processos físicos em lagos, resultando em mudanças na mistura de água, portanto, podem afetar ainda mais a qualidade da água e dos ecossistemas aquáticos. A previsão do nível de água tem atraído cada vez mais a atenção de pesquisadores.

A natureza não-linear e não-estacionária das séries temporais pode resultar em incertezas em certas aplicações. Todavia, alguns pesquisadores tentam combinar diferentes tipos de modelos para melhorar o desempenho preditivo. A combinação de diferentes modelos é chamado de *Ensembles*. Um exemplo típico da aplicação é o algoritmo *Random Forest*.(Joo; Kim, 2015)

Os dados coletados neste trabalho são referentes a barragem Eclusa de São Gonçalo, conforme a Figura 1, que representa uma estrutura hidráulica disposta numa seção do canal São Gonçalo, construída entre 1974 e 1977, com o intuito de impedir a intrusão salina advinda da Laguna dos Patos. Trata-se de uma estrutura indispensável para o desenvolvimento regional, garantindo assim atividades consolidadas na região, tais como capitação da água doce para consumo humano e uso desse recurso para a irrigação, importante para o cenário econômico regional. Associada a esta estrutura, esta uma eclusa, que através de sua operação permite a navegação neste corpo

hídrico.

Figura 1 – Barragem Eclusa do São Gonçalo (Collares, 2022)

O Canal São Gonçalo atua como escoadouro natural das águas da Lagoa Mirim, que através do mesmo e da Laguna dos Patos atingem o Oceano Atlântico. Pelo canal, escoam as águas drenadas pela Bacia Hidrográfica Mirim-São Gonçalo, cuja área superficial é de cerca de 62 mil quilômetros quadrados, sendo 47% desta área localizada em território brasileiro e 53% em território uruguaio.

Como consequência das características do Canal São Gonçalo, sendo este, uma via fluvial responsável pela ligação de dois corpos hídricos de grandes volumes, o canal apresenta um regime de escoamento extremamente complexo, invertendo periodicamente o sentido de sua corrente fluvial, a que lhe vale a designação de canal de comunicação fluvial. O Canal ainda possui como principais características sua extensão da ordem de 75 quilômetros, com larguras variáveis de aproximadamente 200 metros e profundidades também variáveis, oscilando em torno de 5 metros.

Diante desse contexto, a previsão do nível de água na barragem de São Gonçalo revela-se crucial. Tal previsão oferece uma série de vantagens práticas e estratégicas. Primeiramente, proporciona uma gestão mais eficiente dos recursos hídricos, permitindo antecipar variações significativas nos níveis de água. Isso possibilita a implementação de medidas preventivas, como alertas de enchentes ou ações de controle de fluxo, minimizando potenciais impactos ambientais e sociais.

Além disso, a capacidade de prever os níveis hidrométricos contribui para o planejamento adequado das atividades que dependem diretamente dessas condições, como a navegação pelo canal. Com informações antecipadas sobre os níveis de água, é possível otimizar as operações de transporte fluvial, reduzindo riscos e custos associados a variações imprevistas.

Em síntese, a previsão do nível de água na barragem de São Gonçalo não apenas se revela como uma ferramenta valiosa para a gestão sustentável dos recursos hídri-

cos, mas também como um instrumento essencial para a segurança e eficiência das atividades ligadas ao Canal São Gonçalo e seus arredores.

Logo, o objetivo deste estudo é apresentar diversos modelos de aprendizado de máquina para prever os níveis hidrométricos da barragem de São Gonçalo. Analisando as características únicas de cada modelo e sua capacidade de previsão, buscamos desenvolver uma metodologia preditiva que nos permita antecipar os futuros níveis de água, com base no estudo de caso proposto.

A previsão do nível de água é uma área crucial de pesquisa que tem importantes implicações na gestão dos recursos hídricos. Neste contexto, é essencial compreender os diversos aspectos relacionados ao comportamento hidrológico, hidráulico e ambiental dos canais, lagos ou rios. Na próxima seção, exploraremos em detalhes esses aspectos, destacando a influência de variáveis como precipitação, evaporação, escoamento superficial e características geomorfológicas na dinâmica do nível de água. Ao compreendermos melhor esses aspectos hídricos, estaremos mais embasados para entender modelos precisos de previsão do nível de água, contribuindo assim para uma gestão mais eficiente e sustentável dos recursos hídricos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Informações Hidrológicas

A água é vital para existência da vida na Terra. Considerada um solvente universal, é a substância mais reciclável da natureza, e pode-se apresentar dependendo da faixa de temperatura nos três estados físicos: sólido, líquido e gasoso.

Aproximadamente dois terços da superfície do planeta é água. Ela é encontrada na superfície, acima da superfície e abaixo da superfície terrestre, em rios, oceanos, calotas polares, plantas, animais, cursos d'água subterrâneos, ar e nuvens. O maior percentual de água é encontrado nos oceanos (97,5%), seguido de uma pequena fração nos pólos (1,5%) e na água doce disponível para consumo (1%) (Santos; Ricciardi, 2013).

Com exceção da água salgada, a água doce em superfícies de água livre (lagos, rios, reservatórios, entre outros), segundo (Van beek; Wada; Bierkens, 2011)é o recurso com maior acessibilidade para suprir a demanda hídrica da agricultura, principalmente irrigação e pecuária, bem como na indústria e residências. A demanda por água doce aumenta com o crescimento populacional, poluição de mananciais e elevadas taxas evaporativas, fazendo com que a escassez da água seja, de acordo com (Van beek; Wada; Bierkens, 2011), um dos maiores desafios ambientais do mundo, visto que há um desequilíbrio entre oferta e demanda de água.

Uma estratégia de mitigação da escassez em várias regiões do mundo consiste na construção de reservatórios para armazenar água com diversas finalidades, onde fluxos naturais dos rios não são suficientes, para garantir a segurança alimentar. Assim, os reservatórios são imprescindíveis para o planejamento dos recursos hídricos em todo o mundo (Solander et al., 2016). Entretanto, como há grandes perdas de água por evaporação e outros fatores ambientais em reservatórios, a compreensão da dinâmica da água armazenada requer um amplo leque de informações, não somente relacionadas ao corpo hídrico em si, mas também à dinâmica da vegetação e atmosfera no entorno, logo para compreender os fatores que influenciam a alteração do nível de água, é necessário entender estes parâmetros, os quais serão descritos abaixo.

2.1.0.1 Temperatura

A temperatura desempenha um papel fundamental no aumento ou diminuição do nível de água em rios e lagoas, influenciando diversos processos físicos e biológicos que afetam a hidrologia dos corpos d'água. Em primeiro lugar, a temperatura atmosférica influencia a evaporação da água da superfície dos rios e lagoas. À medida que a temperatura aumenta, a taxa de evaporação também se eleva, resultando em uma redução no nível de água.A evaporação pode responder de forma sensível a pequenas variações na temperatura, contribuindo para flutuações nos níveis de água.

Outro efeito da temperatura é sobre o derretimento de geleiras e neve nas bacias hidrográficas. À medida que a temperatura aumenta, a taxa de derretimento também se intensifica, adicionando mais água aos rios e lagos. Isso é especialmente relevante em regiões montanhosas com geleiras, onde o aquecimento global tem sido associado ao aumento do fluxo de água dos rios. Esse fenômeno pode levar a enchentes sazonais e alterações no regime hidrológico(Huss; Hock, 2015).

Além disso, a temperatura afeta a densidade da água, influenciando a sua estratificação térmica em lagos, ou seja, na presença de camadas com diferentes temperaturas. Em estações mais quentes, camadas de água se formam de acordo com a temperatura, com a água mais quente na superfície e a mais fria nas camadas mais profundas. Essa estratificação térmica pode influenciar a mistura de nutrientes e oxigênio no lago, afetando os ecossistemas aquáticos (Levine; Berenson; Stephan, 2000).

Em resumo, a temperatura exerce uma influência significativa no nível de água em rios e lagoas, sendo responsável por regular a evaporação, o derretimento de geleiras, a estratificação térmica, bem como os processos bioquímicos dos ecossistemas aquáticos. Entender como as mudanças climáticas e variações sazonais afetam a temperatura é essencial para a gestão sustentável dos recursos hídricos e a conservação dos ecossistemas aquáticos.

2.1.0.2 Evaporação

O ciclo da água é um dos mecanismos naturais de interação entre continentes, oceanos e atmosfera que ocorre por meio da conversão do balanço de radiação em calor sensível, latente e calor armazenado, ou seja, é o movimento contínuo da água presente nos oceanos, continentes, superfície, solo e rocha e na atmosfera. Esse movimento é alimentado pela força da gravidade e pela energia do Sol, que provocam a evaporação das águas dos oceanos e dos continentes (Trenberth; Fasullo; Kiehl, 2009). Componentes do ciclo hidrológico incluem: escoamento superficial, infiltração, água subterrânea, precipitação, evapotranspiração, condensação, armazenamento de água na atmosfera, neve, armazenamento de água nos oceanos, rios e lagos. O balanço hidrológico representa a contabilização desses componentes no tempo e no

espaço.

O número de moléculas que deixam a superfície está relacionado com a temperatura da superfície da água (Finch, 2001), ou seja, o grau de agitação das moléculas aumenta com a elevação da temperatura. Quando o número de moléculas que deixa a superfície supera o número das que chegam, há um saldo positivo na evaporação. À medida que aumenta a quantidade de vapor d'água acima da superfície da água, maior a pressão, já que o choque das moléculas contra a superfície da água é que determina a pressão de vapor d'água, que será máxima quando o ar estiver saturado.

A evaporação varia de um local para o outro, devido às diferenças climáticas (temperatura do ar, velocidade do vento, radiação solar, umidade relativa, precipitação), características e práticas de manejo do uso da água (Wurbs; Ayala, 2014). A evaporação ocorre sempre que houver um déficit de pressão de vapor d'água entre a superfície da água e a atmosfera acima dessa superfície, e da disponibilidade de energia necessária para o processo, em outras palavras, caso a quantidade de moléculas que deixam o líquido seja maior do que a quantidade de moléculas que entram, teremos como resultado a evaporação.

2.1.0.3 Radiação Solar

No ciclo hidrológico, a principal fonte de energia para iniciar e manter a evaporação no período diurno é a radiação solar (onda curta). Ao interagir com a superfície da água, uma fração da radiação é refletida e outra é absorvida nas camadas superficiais ou irá penetrar na massa hídrica.

A profundidade alcançada pela propagação da radiação solar também é determinada pela turbidez da água, sendo turbidez a condição da água com quantidade excessiva de partículas em suspensão. Visto que existe uma relação direta entre a turbidez e a concentração de sedimentos suspensos (Tananaev; Debolskiy, 2014). Isto demonstra que ao elevar a carga de sedimentos na água, maior a turbidez, consequentemente, maior a dificuldade da penetração da radiação solar no meio.

O balanço de radiação é a contabilidade entre os fluxos radiativos de onda curta (radiação solar) e de onda longa (radiação terrestre). A interação de tais fluxos a superfície desempenha papel fundamental nos processos físicos, químicos e biológicos que ocorrem na interface superfície-atmosfera, especificamente no nível inferior da camada limite planetária. A quantidade de água evaporada da superfície de corpos hídricos depende do balanço de radiação.

Para (Box; Box, 2015) A radiação solar desempenha um papel crucial no aumento do nível de água em rios e lagoas, influenciando diretamente os ciclos hidrológicos e climáticos. Quando a luz solar atinge a superfície da Terra, parte dela é refletida, enquanto outra parte é absorvida pelos corpos d'água. Essa radiação solar absorvida é convertida em calor, aquecendo a água e aumentando sua temperatura. O calor

gerado impulsiona a evaporação, transformando a água líquida em vapor de água.

Com base nos princípios da Termodinâmica, o vapor de água, menos denso que o ar, tende a ascender na atmosfera. A umidade acumulada nessa ascensão resulta na formação de nuvens, onde a água se condensa novamente, formando gotículas que compõem as nuvens de chuva. A precipitação ocorre quando a capacidade de retenção do ar é excedida, liberando a água na forma de chuva ou neve (Box; Box, 2015).

Esse ciclo de evaporação, condensação e precipitação é essencial para a manutenção do nível de água nos corpos hídricos. Sem a radiação solar, esse ciclo seria interrompido, resultando em alterações significativas nos níveis de água em rios e lagoas, afetando ecossistemas, atividades humanas e o suprimento de água potável (Box; Box, 2015).

2.1.0.4 Nível de Água

Em sistemas hidrológicos fluviais, eventos meteorológicos podem causar rápidas alterações hidrodinâmicas. Esses ambientes, foram qualificados por (Barros, 2011) como os de maior vulnerabilidade física e socioeconômica em razão da fragilidade dos seus ecossistemas, sua importância turística e da forte concentração populacional.

A observação dos níveis assumidos pelas águas e dos fenômenos meteorológicos e astronômicos atuantes nesses ambientes contribuem para estabelecer relações de causa e efeito. O conhecimento dessas relações tem especial importância em sistemas de alerta de inundações e monitoramento de cheias, principalmente para previsões de curto e longo prazo.

A partir da previsão dos níveis, poderiam ser beneficiadas atividades como a pesca, recreação náutica, irrigação, atividades portuárias e da indústria naval, bem como auxiliar na gestão de recursos minerais, no planejamento de operações de manutenção necessárias à segurança das vias aquaviárias e na navegação.

O aumento ou diminuição do nível de água em rios e lagos pode trazer uma série de benefícios para os ecossistemas aquáticos e as comunidades humanas que dependem dessas fontes de água. Primeiramente, o aumento do nível de água durante períodos de chuva intensa ou derretimento de geleiras pode recarregar os aquíferos subterrâneos, aumentar a disponibilidade de água para uso humano e irrigação agrícola. Isso é especialmente importante em regiões propensas a secas sazonais ou estiagens prolongadas. Além disso, o aumento temporário do nível de água em rios pode contribuir para a renovação dos leitos e a formação de novos habitats para a vida aquática, promovendo a biodiversidade e favorecendo a pesca sustentável (Vörösmarty et al., 2010).

Por outro lado, a diminuição do nível de água em rios e lagos pode trazer benefícios específicos para alguns ecossistemas e atividades humanas. Por exemplo, em áreas

propensas a inundações sazonais, a redução do nível de água pode minimizar os impactos negativos sobre infraestruturas urbanas e agrícolas, evitando danos materiais e econômicos significativos. Ademais, a diminuição do nível de água em lagos e reservatórios pode expor áreas anteriormente submersas, revelando novos leitos e atraindo a atenção de pesquisadores e arqueólogos, que podem descobrir e estudar artefatos históricos e culturais anteriormente ocultos. Em resumo, o aumento ou diminuição do nível de água em rios e lagos pode oferecer vantagens distintas, ressaltando a importância da gestão cuidadosa dos recursos hídricos para garantir um equilíbrio saudável entre os benefícios ambientais e as necessidades humanas (Vörösmarty et al., 2010).

2.1.0.5 Ventos

De acordo com (Karsburg, 2016), os ventos são definidos como deslocamentos de ar no sentido horizontal, e são originários de gradientes de pressão. Seu deslocamento ocorre no sentido de áreas de maior pressão (mais frias) para áreas de menor pressão (mais quentes), e quanto maior a diferença entre as pressões dessas áreas, maior será a velocidade de seu deslocamento. O vento é responsável pela geração de ondas locais e correntes marinhas, que afetam a deriva litorânea de sedimentos e a configuração das praias, o controle da morfologia dos corpos aquosos costeiros (lagos e lagunas), pela criação de lagos rasos, influenciando na sedimentação dos corpos aquosos (Tomazelli, 1993).

(Cavalcante; Mendes, 2014), demonstram que a utilização de dados dos ventos, contribui para a construção de melhores modelos hidráulicos e hidrológico-hidráulicos, os quais podem ser de base física, conceitual ou empírica.

Os ventos desempenham um papel significativo no aumento do nível de água em rios e lagoas, influenciando diretamente os processos de transporte e acumulação de água. A interação entre a atmosfera e a superfície da água é o que impulsiona essa relação. Quando os ventos sopram sobre a superfície de um corpo d'água, ocorre um fenômeno conhecido como arraste do vento. Esse arraste cria tensão superficial e arrasta as camadas superficiais da água na direção do vento. Essa água deslocada é substituída por água mais profunda, o que acaba elevando o nível do rio ou lago temporariamente (Chanson, 2006).

Além disso, os ventos também podem provocar uma elevação do nível do mar, fenômeno conhecido como "maré meteorológica"ou "maré de tempestade". Em regiões costeiras, ventos fortes associados a tempestades ou sistemas meteorológicos podem empurrar grandes volumes de água para o litoral, causando uma elevação significativa do nível do mar em áreas costeiras baixas e estreitas, como estuários, rios e lagoas costeiras(Chanson, 2006).

2.1.0.6 Precipitação

A precipitação é entendida como toda a água advinda do meio atmosférico que atinge a superfície terrestre, seja ela na forma de neblina, chuva, granizo, saraiva, orvalho, geada e neve. A disponibilidade de dados de precipitação em uma bacia hidrográfica durante um período de tempo é um fator determinante para quantificar diversas atividades, como o abastecimento doméstico e industrial, necessidade de irrigação de culturas e dessedentação animal, dentre outros, para o controle de inundação e da erosão do solo, a determinação de sua intensidade torna-se extremamente importante (Karsburg, 2016).

Este fator tem grande relevância na caracterização do clima de uma região. Períodos de seca muito longos afetam o nível de água dos mananciais e dos reservatórios das usinas hidrelétricas, trazendo problemas e consequências para o abastecimento urbano e também para a geração de energia elétrica (Karsburg, 2016).

A função da precipitação é essencial para o aumento do nível de água em rios e lagoas, desempenhando um papel fundamental nos ciclos hidrológicos naturais. A precipitação refere-se à queda de água em suas diversas formas, como chuva, neve, granizo ou saraiva, sobre a superfície terrestre. Quando a água da precipitação atinge o solo, ela pode seguir três caminhos principais: pode infiltrar-se no solo e recarregar os aquíferos subterrâneos, pode evaporar novamente para a atmosfera, ou pode escoar superficialmente em direção aos corpos d'água, como rios e lagoas (Wescoat jr, 1994).

As chuvas são especialmente cruciais para a manutenção do fluxo dos rios e o aumento do nível das lagoas. A água escoa dos terrenos inclinados para as regiões mais baixas, formando os cursos d'água que compõem os rios. Esses rios, por sua vez, alimentam as lagoas e outros corpos d'água. A precipitação é o principal fator que recarrega o suprimento de água desses sistemas, mantendo seu fluxo e volume (Wescoat jr, 1994)..

2.2 Informações Hidrológicas sobre o Canal São Gonçalo

2.2.1 Estudo de Caso em Questão

Representando uma estrutura hidráulica situada em um trecho do canal de São Gonçalo, localizado no sul do estado do Rio Grande do Sul, Brasil. Construída entre 1974 e 1977, a barragem tem o propósito crucial de impedir a intrusão salina proveniente da Lagoa dos Patos, que está ligada ao Oceano Atlântico. Apresenta-se como uma estrutura indispensável ao desenvolvimento regional, assegurando atividades vitais na zona, como o fornecimento de água doce para consumo humano e o apoio à irrigação, essencial para o panorama econômico regional. Acompanha esta estru-

tura uma eclusa e uma barragem, facilitando a navegação dentro deste corpo hídrico. Atualmente é administrado pela Agência Lagoa Mirim (ALM) da Universidade Federal de Pelotas, marcando 45 anos de valiosos serviços prestados ao desenvolvimento regional das comunidades Brasil-Uruguai.

Considerando um regime pluviométrico de alta irregularidade e de características regionais de evapotranspiração, o Canal São Gonçalo consegue apresentar descargas máximas da ordem de 3.000 metros cúbicos por segundos durante a ocorrência de inundações. Por outro lado, nas estiagens prolongadas, chega até mesmo reduzir tal descarga a zero, quando geralmente se verifica a inversão de sentido em sua corrente, que ocorre também por conta do baixo nível das águas da Lagoa Mirim, aliado ao efeito dos ventos, fazendo as águas escoarem no sentido da Laguna dos Patos para a Lagoa Mirim.

A barragem eclusa consta com uma estrutura transversal conforme a Figura 2 (lado esquerdo), com extensão de 245 metros, de margem a margem, e com 12 metros de profundidade, dos quais nove estão abaixo do fundo regularizado do Canal (cota -5,00 metros em relação ao nível médio do mar - NMM). Em seu trecho central, possui uma extensão de 217 metros e 18 comportas basculantes com vão livre de 11,80 metros por 3,20 metros de altura. O coroamento da parte fixa da barragem localiza-se na cota -2,00 metros e o topo das comportas fechadas atinge a cota 1,20 metros.

Figura 2 – Comportas da barragem e Canal da Eclusagem (Gouvêa, 2009)

Os portões basculantes e as comportas localizam-se nas duas cabeceiras, os quais equalizam/regulam os níveis dentro da eclusa Figura 2 (lado direito), permitindo a passagem de embarcações (Gouvêa, 2009).

O fluxo de escoamento do canal São Gonçalo, normalmente é no sentido da lagoa Mirim para a laguna dos Patos, entretanto, este fluxo pode inverter-se, o que geralmente ocorre em períodos de estiagem (Hartmann; Harkot, 1990). Em virtude de seus usos múltiplos, e como consequência dessa dinâmica de inversão de fluxo, uma barragem eclusa foi construída em sua extensão.

O canal São Gonçalo localiza-se entre as coordenadas 31°45’ e 32°20’ de latitude sul e 52°05’ e 52°50’ de longitude oeste 3. O clima da região, conforme a classificação

climática de *Köppen*, em estudo realizado por (Kuinchtner; Buriol, 2001), é caracterizado como subtropical chuvoso.

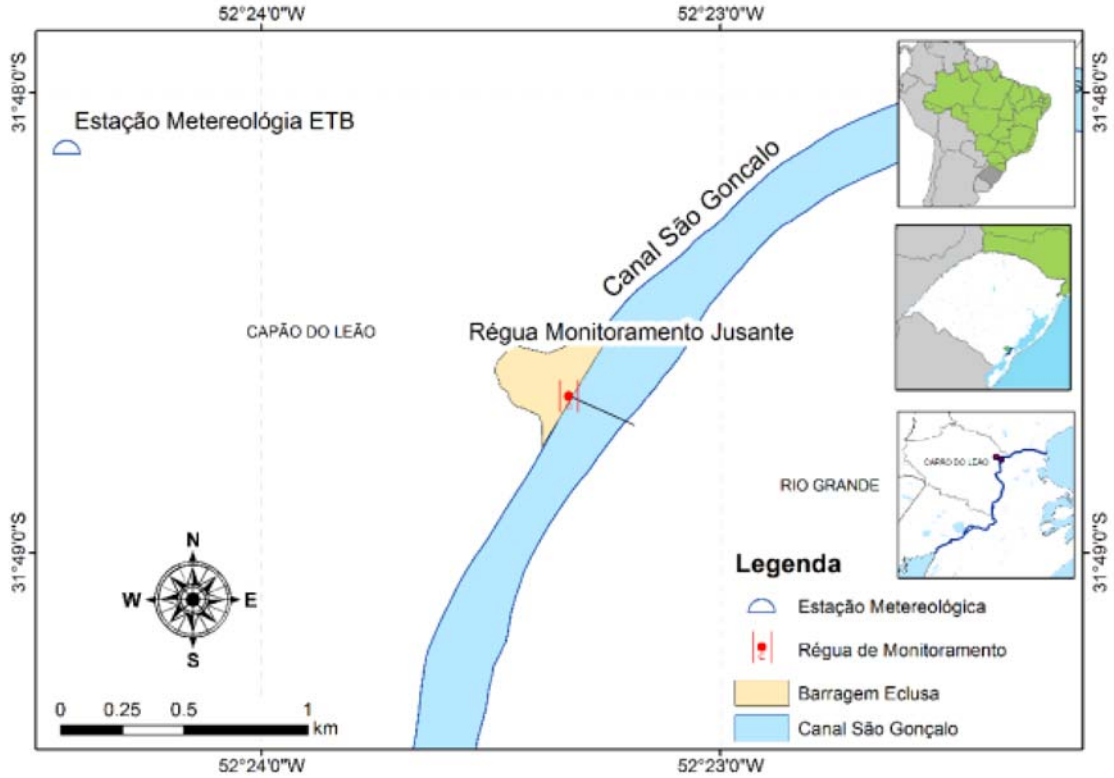

Figura 3 – Localização geográfica do canal São Gonçalo, da barragem em sua extensão (Kuinchtner; Buriol, 2001)

A compreensão da influência hidrológica no Canal São Gonçalo é essencial para uma análise abrangente do comportamento do nível de água neste importante curso d'água. Ao considerar os aspectos hidrológicos específicos deste canal, como a precipitação local, a capacidade de escoamento, a sazonalidade das chuvas e a interação com os sistemas aquáticos adjacentes, podemos elucidar de forma mais precisa os padrões de variação do nível de água ao longo do tempo. Esta próxima seção explorará em detalhes como esses fatores hidrológicos moldam o regime hídrico do Canal São Gonçalo, fornecendo percepções para o desenvolvimento de modelos de previsão do nível de água e para uma gestão mais eficiente e sustentável dos recursos hídricos nesta região específica.

2.2.2 Influência Hidrológica

Conforme o relatório da (Clm, 1970) descrito por (Ambrosi, 2018) os ventos podem ser os principais influenciadores do sentido da corrente do canal São Gonçalo, principalmente quando o nível d'água da Lagoa Mirim apresenta a subida crítica de até 0,70 metros. Segundo (Ambrosi, 2018), geralmente alterações de níveis de água de aproximadamente 0,50 metros ou mais, registrados de um dia para o outro, poderiam

ocorrer no canal.

No verão, o vento de direção dominante do quadrante nordeste, possibilita a ocorrência do fluxo das águas para o sentido da Lagoa Mirim, enquanto o vento proveniente da direção sudoeste, faz com que o sentido de escoamento das águas seja para o sentido da Lagoa dos Patos. Portanto, ocorre um equilíbrio de massa de água, onde o vento acentuado de direção sudoeste, resulta na remoção de água da Lagoa Mirim e o vento predominante de nordeste, provoca a recolocação de água (Ambrosi, 2018).

Conforme (Clm, 1970) no canal São Gonçalo, o avanço da frente salina é provocado pelos ventos e seus efeitos influenciam diretamente nos níveis d'água nas regiões extremas do corpo hídrico. Segundo o mesmo relatório, o Rio Piratini em épocas marcadas pela entrada de água salgada, apresenta uma diminuição de sua descarga, proporcionando sentido de fluxo das águas para a Lagoa Mirim.

Os fenômenos de El Niño e La Niña são variações anômalas das águas superficiais do oceano Pacífico Tropical, o evento de El Niño é identificado como a elevação da temperatura destas águas, enquanto o fenômeno La Niña, é representado pelo resfriamento destas águas superficiais. Na Lagoa dos Patos, o fenômeno El Niño resulta em uma elevada descarga fluvial, por outro lado, durante o fenômeno de La Niña, registra-se a ocorrência de vazões menores (Vaz; Möller junior; Almeida, 2006).

No canal São Gonçalo, conforme o relatório (Ambrosi, 2018) as descargas e os níveis de água são influenciados diretamente pelos ventos conforme sua permanência, direção e amplitude. Os ventos provenientes de nordeste favorecem a intrusão de água salgada no sentido da Lagoa Mirim, entretanto os ventos oriundos de sudoeste, proporcionam o escoamento das águas para a Lagoa dos Patos (Ambrosi, 2018).

Outro fator a ser considerado, com influente na previsão do nível de água, é a precipitação. (Blain et al., 2009) analisaram a variabilidade amostral das séries mensais de precipitação pluvial em Pelotas-RS e Campinas-SP. Os autores verificaram que na região de Pelotas-RS, as chuvas são distribuídas de forma semelhante ao longo do ano, ao contrário do que foi observado para a região de Campinas-SP, que possui uma estação seca definida.

Na região do canal São Gonçalo, a precipitação média anual é de 1.366,9mm, sendo os meses de janeiro, fevereiro e julho são os mais chuvosos, proporcionando diferenciados fenômenos como alagamentos das áreas mais baixas, ligados aos elevados volumes de precipitação e consequente deficiência no escoamento das águas (Simon; Silva, 2015).

No que se refere a influência do nível de água, os valores de níveis representam a quantidade de água que um recurso hídrico possui. É altamente variável, afetado em relação às estações do ano, períodos de estiagem e cheias, dentre outros.

Conforme exposto anteriormente, nos itens pregressos (precipitação e ventos), a dinâmica do ambiente em estudo é relativamente complexo. Outro fato que merece

importante destaque e atenção, é o sentido do fluxo de escoamento do canal São Gonçalo. Normalmente é no sentido da lagoa Mirim para a laguna dos Patos, cerca de 70% do tempo, entretanto, este fluxo pode inverter-se, o que geralmente ocorre em períodos de estiagem (Hartmann; Harkot, 1990). Exatamente por esta inversão de fluxo, que a barragem eclusa foi construída, objetivando impedir que as águas salinas adentrem à montante da barragem. Somando-se a este cenário, a baixa declividade do canal, com a atuação dos ventos e com os níveis de água da laguna dos Patos e da lagoa Mirim, são os fatores que mais influenciam no processo de inundação da planície aluvial do canal São Gonçalo. Pelo fato de conectar as duas lagoas, esse canal se torna o curso da água que mais sofre com o efeito de eventos de inundações.

Como consequência, a vazão do canal depende do nível da água na laguna dos Patos, das condições de fluxo do Rio Piratini (seu principal contribuinte), do nível do canal na sua desembocadura e das condições impostas na barragem do canal São Gonçalo. Seu nível está relacionado, além dos fatores acima mencionados, com a ação dos ventos que, na maior parte das vezes, é o fator mais importante (Hartmann; Harkot, 1990).

As precipitações pluviométricas também são um fator crítico que afeta o nível de água no canal São Gonçalo. Chuvas intensas podem resultar em um rápido aumento do nível de água, levando a enchentes e transbordamentos. Além disso, o fluxo de água no canal São Gonçalo também é afetado pela vazão de outros rios e cursos d'água que desembocam nele. A disponibilidade de água proveniente de bacias hidrográficas adjacentes pode variar significativamente ao longo do ano, influenciando diretamente o nível de água no canal.

Além disso, a temperatura também se mostrou relevante. A elevação da temperatura pode afetar a evaporação da água do canal, contribuindo para uma diminuição do nível de água durante períodos mais quentes. A relação entre a temperatura e a evaporação também pode influenciar a qualidade da água, com potencial aumento da concentração de poluentes durante secas prolongadas.

2.2.2.1 Dados de Nível

Os dados de nível utilizados neste estudo, são os das seguintes estações: Santa Isabel do Sul, Santa Vitória do Palmar e os níveis da Eclusa Jusante e Motante. O local de monitoramento em Santa Isabel do Sul fica a cerca de 5 km da entrada da Lagoa Mirim para o Canal São Gonçalo, marcando a região mais ao norte da lagoa. Enquanto isso, o posto de monitoramento no Porto de Santa Vitória do Palmar está situado aproximadamente 170 km ao sul de Santa Isabel, representando as condições na parte sul da lagoa. Ambos esses pontos de observação são fundamentais para entender como a água se move nessa área. Por isso, eles são monitorados por um longo período, e os dados coletados nesses lugares nos fornecem várias informações

sobre a Lagoa Mirim e o sistema que a envolve.

Na Figura 4, são apresentadas as cotas máximas e mínimas diárias e as medianas baseadas na série histórica de 1978 até 2016 para o posto de monitoramento de Santa Isabel do Sul, e de 1978 até 2013 para o posto de monitoramento do Porto Santa Vitória do Palmar. A região marcada em lilás caracteriza os dados entre 10 e 90% de permanência para os dados diários de cotas. Na prática, isto indica, o intervalo de cotas ao longo do ano que se encontram 80% dos dados diários. As variações diárias de cotas, indicam que historicamente as medianas mínimas ocorrem no mês de março (1,02 e 1,18 m) e as máximas nos meses de outubro e setembro (2,37 e 2,46 m),para as estações de Santa Isabel do Sul e Santa Vitória do Palmar, respectivamente (Collares, 2022).

Os máximos valores de cotas ocorreram no dia 21 de julho de 1984 no posto de monitoramento de Santa Isabel do Sul (5,13 m) e em 17 de julho de 1984 em Santa Vitória do Palmar (5,22 m). No ano de 1984, a precipitação acumulada até o mês de julho foi de 1201 e 1378 mm para as estações da Embrapa e INIA(Instituto Nacional de Investigación Agropecuária), respectivamente. Lâminas de precipitação muito superiores, ao valor mediano acumulado até o referido mês (799 e 825 mm) para as estações da Embrapa e INIA, respectivamente(Collares, 2022).

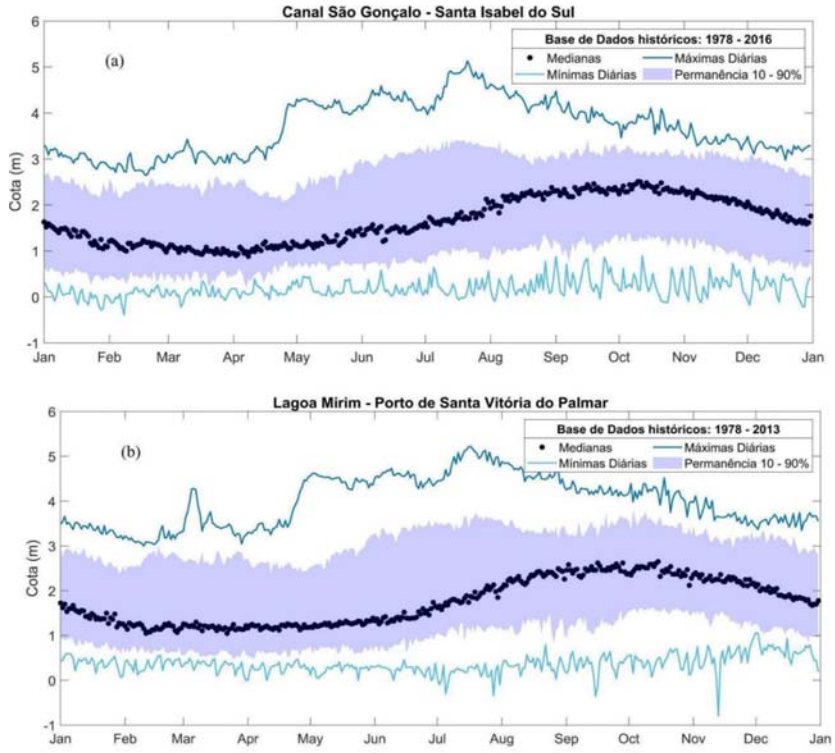

Figura 4 – Série histórica de níveis para dois postos de monitoramento: (a) Santa Isabel do Sul - Canal São Gonçalo, entre 1978-2016; (b) Porto de Santa Vitória do Palmar - Lagoa Mirim, entre 1978-2013.(Collares, 2022)

Segundo os dados coletados entre 1971 e 2020 na estação agroclimatológica da

Embrapa Clima Temperado, observa-se que os ventos predominantes na BHMSG ao longo do ano são aqueles provenientes do nordeste. No entanto, no outono e inverno, há uma maior prevalência de ventos provenientes do quadrante sul, em comparação com a primavera e o verão. As velocidades médias dos ventos permanecem relativamente estáveis durante as quatro estações, variando de 4,0 m/s na primavera a 2,8 m/s no outono.

A predominância dos ventos do nordeste é resultado da circulação atmosférica anticiclônica sobre o Oceano Atlântico Sul. Enquanto isso, os ventos do quadrante sul são mais frequentes durante a passagem de massas de ar frio e ciclones extratropicais, ocorrendo com maior frequência durante o inverno devido ao avanço para o norte da frente polar.

De acordo com (Beltrame et al., 1998) a velocidade do vento tem um impacto significativo no nível de água na bacia da Lagoa Mirim. Devido à orientação natural da lagoa no sentido Sudoeste-Nordeste e aos ventos predominantes do quadrante Nordeste, ocorre um deslocamento da massa de água na lagoa. Isso resulta em variações constantes nos níveis de água.

Quando os ventos são mais intensos, a velocidade de propagação das ondas gravitacionais aumenta. Isso pode levar a um aumento nos níveis de água em certas áreas da lagoa. Por outro lado, quando os ventos são mais fracos, a velocidade de propagação das ondas diminui, o que pode resultar em níveis de água mais baixos. Essas variações nos níveis de água devido à velocidade do vento podem ter impactos significativos nas áreas inundadas ciclicamente pela lagoa, especialmente nos banhados ribeirinhos.

De acordo com (Collares, 2022) é relatado que ao analisar os dados de nível dos postos de monitoramento em Santa Isabel do Sul e Porto de Santa Vitória do Palmar, juntamente com informações sobre a velocidade do fluxo no Canal São Gonçalo durante um curto período, de 13 a 24 de maio de 2022, e correlacioná-los com os registros de vento em Santa Isabel do Sul, observa-se padrões interessantes. Durante o período de 15 a 19 de maio, quando os ventos predominaram no quadrante sul, houve uma diminuição nos níveis de água na parte sul da Lagoa Mirim e um aumento correspondente nos níveis na porção norte da lagoa e no próprio Canal São Gonçalo. Como resultado, a velocidade do fluxo de água aumentou significativamente, chegando a atingir 1,2 metros por segundo, fluindo da Lagoa Mirim em direção à Laguna dos Patos.

Em resumo, o nível de água da Lagoa Mirim oscila devido a períodos de estiagem e cheia em diferentes estações do ano, apresentando um avanço significativo das margens durante a primavera, recuando durante o verão, e mantendo-se durante o outono e parte do inverno. Além disso, a variação do nível da Lagoa também está associada à disposição dos materiais e à incidência de ventos, que podem causar

ondas com grande potencial erosivo. Outro fator que contribui para o aumento ou redução do nível é o efeito da precipitação na Lagoa Mirim, as condições de velocidade e direção dos ventos também influenciam na dinâmica dos níveis na lagoa.

A compreensão da influência hidrológica no Canal São Gonçalo não apenas nos permite entender os padrões de variação do nível de água, mas também nos oferece uma base para explorar a aplicação de conceitos de séries temporais na análise desses dados. A análise de séries temporais é uma ferramenta fundamental para identificar tendências, padrões sazonais e variações cíclicas nos dados hidrológicos coletados ao longo do tempo. Nesta próxima seção, iremos abordar a parte conceitual das séries temporais, explorando métodos e técnicas que nos permitem modelar e prever o comportamento do nível de água com base em dados históricos e padrões identificados. Ao integrar o conhecimento sobre a influência hidrológica específica do Canal São Gonçalo com as técnicas analíticas das séries temporais, estaremos melhor equipados para desenvolver modelos precisos e robustos de previsão do nível de água, contribuindo assim para uma gestão mais eficaz dos recursos hídricos nesta região.

2.3 Séries Temporais

Aspectos relacionados ao tempo são elementos globais na experiência humana. Quase todas as atividades que podemos realizar e os fenômenos que podemos observar trazem consigo algum componente de natureza temporal. Em muitas das situações, o estudo desse componente é essencial para ser possível aprimorar como uma atividade é realizada ou para que se possa melhor compreender algum aspecto da natureza

Uma forma de representar dados relativos ao tempo é através de séries temporais, que são, resumidamente, um conjunto de observações sequenciais que expressam o desencadeamento de algum fenômeno durante um período de tempo. Séries temporais são um tipo de dado que é explorado em diversas áreas do conhecimento, como financeiro, saúde etc. Uma aplicação particularmente relevante para esta Tese é a regressão de séries temporais, na qual o objetivo é estimar uma variável numéria (contínua) a partir de dados temporais,

Podemos conceituar uma determinada série temporal como um conjunto de observações quantitativas ordenadas cronologicamente (Kirchgässner; Wolters; Hassler, 2012). Séries temporais podem implicar em representar conceitos muito mais complexos do que a simples coleção de suas observações e, logo, são naturalmente empregadas para descrever fenômenos que variam com o tempo. Em algumas situações o estudo de séries temporais permite compreender melhor a natureza dos fenômenos por elas descritos.

Em outra abordagem, uma série temporal pode ser definida como a realização de um processo estocástico. Um processo estocástico é um processo aleatório que se desenvolve com o passar do tempo. Mais precisamente, um processo estocástico é um conjunto ordenado de variáveis aleatórias $X = X_t$. O índice $_t$ é contido em $N = 1, 2, 3, ...$ (Lawler, 2018), em outras palavras, um processo estocástico de tempo discreto é um conjunto contável de variáveis aleatórias

A ideia básica que norteia a análise de séries temporais é que há um fator relativamente causal por vezes constante, relacionado com o tempo, que exerceu influência sobre os dados no passado e pode continuar a fazê-lo no futuro. Este fator causal costuma atuar criando padrões não aleatórios que podem ser detectados em um gráfico da série temporal, ou mediante algum outro processo estatístico. Essas influências presentes nas séries temporais serão explorados na próxima seção.

2.3.1 Aspectos de uma Série Temporal

A maneira tradicional de analisar uma série temporal é através da sua decomposição nos componentes de tendência, ciclo e sazonalidade, onde a tendência de uma série indica o seu comportamento “de longo prazo”, em outras palavras, se ela cresce, decresce ou permanece estável, e qual a velocidade destas mudanças. Nos casos mais comuns trabalha-se com tendência constante, linear ou quadrática, como ilustrado na figura abaixo.

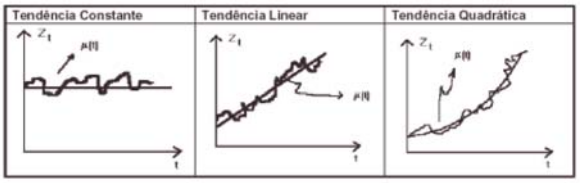

Figura 5 – Tendências de uma Série Temporal (Lawler, 2018)

Outro parâmetro são os ciclos caracterizados pelas oscilações de subida e de queda nas séries, de forma suave e repetida, ao longo da componente de tendência. por exemplo, ciclos relacionados à atividade econômica ou ciclos meteorológicos.

A sazonalidade em uma determinada série de dados corresponde às oscilações de subida e de queda que ocorrem sempre em um determinado período do ano, do mês, da semana ou do dia. A diferença essencial entre os componentes sazonal e cíclica é que a primeira possui de certa forma movimentos facilmente previsíveis, ocorrendo em intervalos regulares de tempo, enquanto movimentos cíclicos tendem a ser irregulares.

Os resíduos de uma série temporal representam os valores restantes após a remoção dos outros componentes e é resultante de flutuações de curto prazo da série que não são sistemáticas nem previsíveis.

Um dos principais objetivos da análise desse componente é identificar o grau de irregularidade da série. Uma vez que a série possui um grau de irregularidade muito

alto, os resíduos tendem a influenciar no movimento da série, camuflando os valores de tendência e sazonalidade, o que irá implicar em estimativas não sustentáveis.

Uma característica importante a ser observada é a estacionariedade. Para uma série ser estacionária, a média não pode estar em função do tempo, isto é, a média da série não deve alterar com o passar do tempo, ou seja, os dados oscilam sobre uma média constante, independente do tempo, com a variância das flutuações permanecendo essencialmente a mesma.

(Diniz et al., 1998) conceitua uma série temporal relatando que sua estacionalidade é um processo aleatório que oscila em torno de um nível médio constante. Séries temporais sazonais ou com tendência linear, ou exponencial são exemplos de séries temporais com comportamento não estacionário. A condição de estacionariedade de 2ª ordem implica em:

a) Média do processo é constante;

b) Variância do processo é constante.

De certa forma, ao analisarmos uma série temporal estamos interessados em:

a) Análise e modelagem da série temporal - descrever a série, verificar suas características mais relevantes e suas possíveis relações com outras séries;

b) Previsão na série temporal - a partir de valores históricos da série (e possivelmente de outras séries também) procura-se estimar previsões de curto ou de médio prazo *forecast*. O número de tempos à frente para o qual é feita a previsão é chamado de horizonte de previsão.

Para que a escolha de um determinado modelo para uma série temporal, seja considerando sendo por métodos estatísticos ou por redes neurais artificiais, é necessário conhecer a relação entre as observações atuais e as anteriores. Uma forma de avaliá-la é através das funções de autocorrelação.

A função de autocorrelação é de certa forma muito utilizada para estimar parâmetros de modelos estatísticos como ARIMA (Tiao; Tsay, 1985). Embora a função de autocorrelação pode ser estimada para qualquer série temporal. Este processo de autocorrelação pode ser utilizada para detectar se o processo é estacionário. Os coeficientes de autocorrelação de um processo estacionário rapidamente se aproximam de zero.

Na Figura 6, observa-se uma série temporal de uma determinada Companhia Aérea onde o eixo do x representa os meses e y o número de passageiros. Neste caso podemos fazer a seguinte pergunta que padrões não aleatórios essa série apresenta? Observe que há uma tendência crescente no número de passageiros transportados e verifica-se que há uma sucessão regular de “picos e caídas” no número de passageiros transportados, isso deve ser causado pelas oscilações devido a feriados, períodos de férias escolares, etc., geralmente relacionados às estações do ano, e que se repetem todo ano (com maior ou menor intensidade), característica da sazonalidade. Em

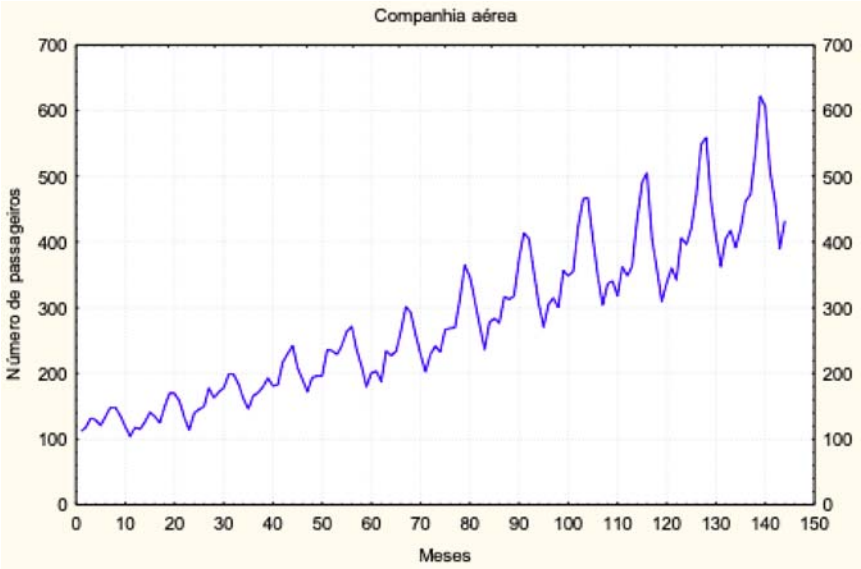

Figura 6 – Número de Passageiros Transportados (Lawler, 2018)

outras palavras, observa-se dois padrões que podem tornar a ocorrer no futuro: crescimento no número de passageiros transportados e flutuações sazonais. Tais padrões poderiam ser incorporados a um modelo estatístico, possibilitando fazer previsões que auxiliarão na tomada de decisões.

Verificando agora outra série temporal conforme a Figura 7, onde temos um gráfico de produção veicular no Brasil de janeiro de 1997 a dezembro de 2014. Então podemos realizar o seguinte questionamento, quais padrões podem ser identificados?

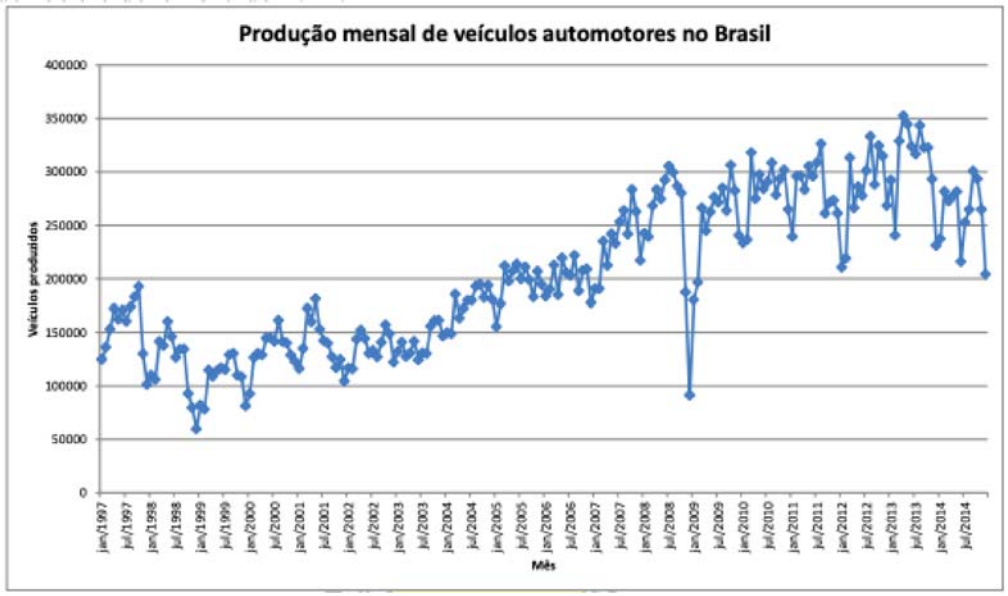

Figura 7 – Série mensal da produção de veículos automotores - (Adaptado pelo Autor, Fonte: https://anfavea.com.br/site/)

Observe na Figura 7, que há uma tendência crescente no número de veículos produzidos (começando em cerca de 125000 em janeiro de 1997 e terminando em 200000 em dezembro de 2014). Outra análise é, que as flutuações (picos e caídas)

não são tão regulares quanto as identificadas na Figura 6.

Verifica-se também uma queda na produção no mês de janeiro de 2009, em fins de 2008 a produção mensal estava em torno de 300000 veículos, e caiu para menos de 100000 naquele mês, provavelmente por algum fator externo na economia.

A análise deste gráfico 7 indica uma linha de tendência crescente por haver alteração na média e na variância, revelando que a série é não-estacionária, sugerindo assim uma transformação para estabilizar a série em estudo.

2.3.1.1 Análise Séries Temporais

Como já mencionado, a análise de séries temporais pode ter variados propósitos, com isso, as ferramentas utilizadas nessa tarefa também variam, conforme a série a ser analisada. Há quatro classes básicas de objetivos que geralmente se buscam atingir com a análise de uma ST (Morettin; Toloi, 2018), podem ser eles:

- Controle: Esse tipo de análise é feito quando há a necessidade de se acompanhar a qualidade de um processo por meio dos índices que compõem a série temporal de interesse. Por exemplo, em uma usina nuclear, em que a ST é originada a partir das medidas de temperatura de um tanque de resfriamento e faz-se necessário o controle dessa temperatura para evitar acidentes.

- Descrição: Essa análise descreve a maneira pela qual a série temporal se comporta, identificando componentes como tendência, sazonalidade e índices que diferem do comportamento padrão dessa ST. Como exemplo, podemos tomar uma série que representa as temperaturas mensais em determinada cidade nos últimos dez anos, e o objetivo seria determinar a sazonalidade nessas temperaturas e identificar com isso, a possível existência de valores discrepantes.

- Explicação: Esse objetivo busca explicar o comportamento de uma série em detrimento do comportamento de outra, ou seja, há duas ou mais séries temporais e busca-se identificar a influência das variações de uma ou mais séries sobre outra(s). Um exemplo seriam duas séries, uma representando o índice de chuvas mensal de uma região, e outra o nível de água em uma represa, no mesmo período e na mesma região, desse modo, o objetivo seria explicar o nível dessa represa em detrimento do índice de chuva.

- Previsão: Nessa análise os valores passados de uma ST são utilizados para a identificação dos valores futuros dessa. Um exemplo seria uma série temporal representando a quantidade de pacotes que trafegam nas interfaces de conexão com as LANs em um roteador, o objetivo dessa análise seria identificar a quantidade de pacotes que trafegarão em cada uma dessas interfaces, com o intuito realizar a realocação da largura de banda máxima admissível para essas interfaces.

2.3.1.2 Estacionariedade em séries temporais

Estacionariedade em uma série temporal diz respeito ao seu desenvolvimento no tempo ao redor de uma média constante, denotando, com isso, uma forma de equilíbrio constante. Entretanto, geralmente as séries temporais são não-estacionárias e essa característica pode seguir, por exemplo, uma tendência linear descendente ou ascendente, ou então pode apresentar um comportamento explosivo (Morettin; Toloi, 2018). A Figura 8, apresenta essa não-estacionariedade explosiva.

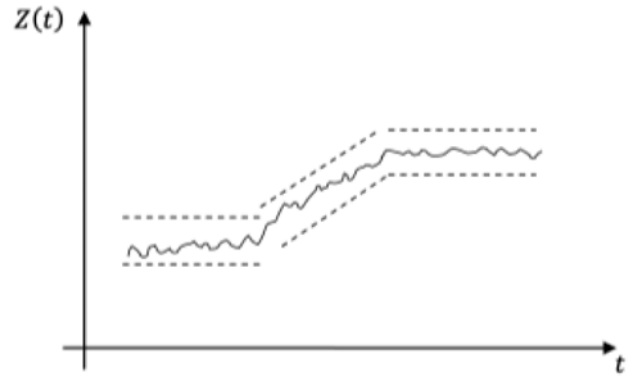

Figura 8 – Série temporal não-estacionária quanto ao nível de inclinação (Morettin; Toloi, 2018)

A não-estacionariedade é um comportamento presente na maioria das séries temporais existentes, entretanto os métodos de análise estatística supõem que elas sejam estacionárias. Desse modo, é necessário transformar os dados originais, caso não sejam estacionários. A transformação normalmente utilizada é a realização de diferenças sucessivas entre os coeficientes da ST, até que ela se torne estacionária. Geralmente, uma ou duas diferenças são suficientes para tornar a série estacionária (Morettin; Toloi, 2018).

2.3.1.3 Previsão em séries temporais

A previsão de uma série temporal é a estimação de seus valores futuros. Para isso, são utilizados os dados passados e presentes da série para projetar os próximos índices da série (Sorjamaa et al., 2007). Assim se podem predizer os novos índices de uma ST usando os do passado e o atual, são descritos pela Equação abaixo.

$$Z_{t+1} = f(Z_t, Z_{t-1}, Z_{t-2}, ...) \tag{1}$$

onde $f(Z_t, Z_{t-1}, Z_{t-2}, .)$ é uma função que usa os índices passados $Z_t, Z_{t-1}, Z_{t-2}, ..$ para predizer o valor no instante t+1.

O principal interesse, neste trabalho, é a análise das séries temporais com o objetivo da previsão do nível de água em um canal. Do ponto de vista da disponibilidade dos dados, há duas maneiras de se analisar séries temporais para previsão: qualitativa ou quantitativa (Levine; Berenson; Stephan, 2000).

A análise qualitativa, geralmente, é empregada nos casos em que existem poucos dados para compor a ST e requer, portanto, maior conhecimento sobre ela. Métodos qualitativos de previsão são essenciais em situações em que os dados históricos da série não estão disponíveis, entretanto são, geralmente, arbitrários e subjetivos (Levine; Berenson; Stephan, 2000).

A análise quantitativa usa modelos matemáticos para se chegar a valores preditos. Os métodos quantitativos de previsão empregam valores passados da ST para se chegar a um valor futuro. Esses métodos são divididos em dois grupos: i) Métodos de previsão de séries temporais, em que há a extrapolação de valores futuros da série a partir de índices passados e presentes; ii) Métodos causais de previsão, utilizando outros fatores externos à ST e até mesmo outras séries para explicar o comportamento futuro da série temporal de interesse (Levine; Berenson; Stephan, 2000).

Os métodos de previsão podem ser divididos em dois grupos: métodos não paramétricos, que se adaptam a diversos comportamentos assumidos por determinada ST com o passar do tempo. Esses métodos não possuem parâmetros para uma distribuição em particular. Alguns deles são os que combinam técnicas como lógica nebulosa e algoritmo genético os baseados no algoritmo kNN *k-Nearest Neighbor*, os que utilizam ANN *Artificial Neural Network*, além de outros; métodos paramétricos, o qual são modelados a partir de parâmetros predeterminados e seguem uma distribuição específica.

Um método de predição utilizado é o *Naïve*. Esse modelo é simples, em relação aos demais métodos de predição, entretanto, pode apresentar acuracidade satisfatória em determinadas séries temporais, ao ser comparado com outros modelos de predição. Esse modelo é dado pela equação abaixo, em que o valor predito é o último observado (Hanke; Wichern, 2005).

$$\hat{Y_t} = Y_{t-1} \tag{2}$$

A média móvel simples, outra técnica utilizada na predição de séries temporais, utiliza a média de uma amostra da série temporal, para expressar o próximo valor daquela série (Morettin; Toloi, 2018).A média móvel é dada pela Equação abaixo:

$$MM = \frac{V_1 + V_2 + ...V_N}{N} \tag{3}$$

Onde, V é o valor observado na série e N é a quantidade de valores observados.

A Suavização Exponencial é outro modelo de predição, que atribui pesos às médias móveis. Esses pesos decrescem à medida que os índices observados se distanciam em relação ao índice presente, observado (Hanke; Wichern, 2005). A abordagem mais comum, expressa na Equação abaixo é:

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y - \hat{Y}_t) \tag{4}$$

onde, α é a constante de suavização.

O modelo que mais se adapta a uma diversidade maior de séries é o modelo ARIMA *Autoregressive Integrated Moving Average*, originado a partir da combinação de outros modelos. Essa característica é possível, pois, geralmente, as séries são não estacionárias e, para adequá-las à análise estatística, o modelo ARIMA realiza a diferenciação entre os índices da série, de modo a torná-la estacionária. Esse modelo é derivado do modelo ARMA *Autoregressive Moving Average*, que por sua vez é a combinação dos modelos AR *Autoregressive*, MA *Moving Average* (Morettin; Toloi, 2018).

Após a análise dos conceitos fundamentais das séries temporais, vamos integrar esses conhecimentos com os avanços na área de aprendizado de máquina para aprimorar a capacidade de previsão do nível de água no Canal São Gonçalo. Na próxima seção, será abordado os princípios e modelos de aprendizagem de máquina aplicados à previsão hidrológica, destacando técnicas como redes neurais, árvores aleatórias, modelos de regressão e algoritmos de séries temporais. Ao combinar o entendimento das séries temporais com as capacidades preditivas da aprendizagem de máquina, pode-se desenvolver abordagens mais sofisticadas e precisas para prever o comportamento do nível de água neste canal específico, contribuindo assim para uma gestão mais dos recursos hídricos.

2.4 Aprendizado de Máquina

O Aprendizado de Máquina, do inglês *Machine Learning* (ML), pode ser definido em sua essência como a extração de conhecimento a partir de bases de dados. O ML é um tópico composto pela intersecção de diversas outras áreas, como estatística, inteligência artificial e ciência da computação, abrangendo campos como a análise preditiva e o aprendizado estatístico (Bakshi; Bakshi, 2018). Nos últimos anos, o ML tem possibilitado o desenvolvimento de aplicações como: carros autônomos, processamento de linguagem natural, mecanismos eficientes de busca, sistemas inteligentes de recomendação, dentre outros.

A aplicação do ML tem por objetivo construir novos algoritmos e aprimorar os existentes de forma que esses sejam capazes de aprender a partir de bases de dados, construindo modelos generalizáveis capazes de realizar predições precisas e/ou encontrar padrões, a partir de dados desconhecidos (Bakshi; Bakshi, 2018).

O processo de aprendizado consiste em melhorar um sistema de conhecimento através da ampliação ou rearranjo da base de conhecimento/motor de inferência (Russell, 2010). O aprendizado de máquina compreende métodos computacionais para

adquirir novos conhecimentos, novas habilidades e novas maneiras de organizar o conhecimento existente (Tyugu, 2011).

Os problemas de aprendizado podem variar bastante em termos de complexidade, desde a aprendizagem paramétrica (que visa o aprendizado de valores para determinados parâmetros) até formas complicadas de aprendizado simbólico (que visa o aprendizado de conceitos, gramáticas, funções e até mesmo comportamentos) (Tyugu, 2011).

A compreensão e o tratamento adequado das características das séries temporais, como tendências e sazonalidades, são fundamentais para muitas aplicações analíticas. No entanto, em muitos casos, esses dados podem ser ainda mais explorados e utilizados em conjunto com técnicas de aprendizagem de máquina. Ao unir as percepções obtidas da análise de séries temporais com os algoritmos de aprendizagem de máquina, é possível aprimorar ainda mais a capacidade de previsão e entendimento de padrões, permitindo compreensões mais profundas e precisas em uma ampla gama de contextos.

2.4.1 Métodos de Aprendizado

Independentemente da técnica de ML escolhida para resolver um determinado problema, o aprendizado dela está intrinsecamente ligado a sua capacidade em reconhecer e classificar padrões *(pattern recognition)* e *(pattern classification)* respectivamente. A criação de classificadores ou regressores envolve a definição de um modelo geral e o uso de dados de treinamento para aprender ou estimar os parâmetros desconhecidos desse modelo (Duda; Hart et al., 2006). Na prática, o aprendizado pode ocorrer de quatro maneiras distintas, detalhadas a seguir.

2.4.2 Aprendizado Supervisionado

No aprendizado supervisionado *supervised learning*, um supervisor fornece um rótulo categórico (ou custo) para cada padrão do conjunto de treinamento, de modo que o objetivo do problema passa a ser a redução do somatório de custos para estes padrões(Duda; Hart et al., 2006).

Na prática, os algoritmos de aprendizado supervisionado já são bem compreendidos na área e fáceis de avaliar o desempenho. De modo geral, se uma aplicação é modelada como um problema de aprendizado supervisionado e são fornecidas as saídas necessárias no conjunto de treinamento, é provável que o ML supervisionado resolva o problema (Bakshi; Bakshi, 2018).

Os algoritmos de aprendizado supervisionado podem ser interpretados como classificadores ou regressores, dependendo da natureza do problema (Bakshi; Bakshi, 2018). Na classificação, o objetivo é prever um rótulo de classe, sendo uma escolha de uma lista predefinida de possibilidades. A classificação pode, ainda, ser binária ou

multi classe, isto é, lidar com apenas duas classes pré-definidas ou até mais de duas classes.

Na regressão, o objetivo é prever uma ou mais variáveis contínuas. Um exemplo de problema de regressão seria a predição do salário anual de um indivíduo a partir do seu nível de educação, idade e local de residência. Neste caso, o salário, pode ser qualquer valor em um determinado intervalo e, por existir esta continuidade, trata-se de um problema de regressão e não de classificação.

Na etapa de aprendizado, entende-se por ideal a criação do modelo mais simples possível capaz de realizar predições com uma satisfatória taxa de acurácia. Quando são criados modelos muito complexos para uma determinada base de dados, diz-se que houve um super-treinamento *overfitting*. O super-treinamento ocorre quando os modelos são criados demasiadamente próximos da natureza da própria base de dados de treinamento. Neste cenário, apesar do modelo funcionar bem para os dados do conjunto de treinamento, ele é incapaz de extrapolar boas predições para dados desconhecidos (Bakshi; Bakshi, 2018).

Por outro lado, quando são criados modelos muito simples, dificilmente eles conseguirão capturar todos os aspectos relevantes das respectivas bases de dados e, neste caso, diz-se que houve um sub-treinamento *underfitting* (Bakshi; Bakshi, 2018).

2.4.3 Aprendizado Não Supervisionado

Na aprendizagem não supervisionada, os algoritmos assumem que não se conhece a classe à qual os exemplos pertencem e procuram encontrar nos valores de atributos similaridades ou diferenças que possam, respectivamente, agrupar os exemplos pertencentes à mesma classe ou dispersar os exemplos de classes distintas. (Russell, 2010)

Dentre estas aplicações do aprendizado não-supervisionado, a visualização é, potencialmente, a mais utilizada em problemas de natureza não-supervisionada. Destacam-se as técnicas: *Principal Component Analysis (PCA)*, *Non-negative Matrix Factorization (NMF)*, *Singular Value Decomposition (SVD)*, L*inear Discriminant Analysis (LDA)* e *t-SNE* (Bakshi; Bakshi, 2018).

2.4.4 Aprendizado Semi-Supervisionado

O aprendizado semi-supervisionado combina o modo de aprendizagem supervisionado e não supervisionado, utilizando um pequeno conjunto de exemplos rotulados e um conjunto de exemplos não rotulados.

Destacam-se as técnicas: Algoritmo de Maximização de Expectativas *Expectation Maximization – EM* com Modelos de Misturas Generativas, *Self-training*, *Co-training*, *Transductive Support Vector Machines (TSVM)* e métodos baseados em grafos, como o *Particle Competition and Cooperation (PCC)*(Zhu, 2005).

2.4.5 Aprendizado por Reforço

Na aprendizagem por reforço *Reinforcement Learning*, existe uma interação semelhante ao do aprendizado supervisionado no que diz respeito ao modo como um agente influência no aprendizado do ambiente. Todavia, nesta abordagem, o único *feedback* dado em relação à predição do classificador é se a tentativa foi correta ou errada, não; revelando, portanto, a verdadeira categoria do dado analisado (Duda; Hart et al., 2006). Deste modo, diferentemente do aprendizado supervisionado, as saídas ótimas do problema são descobertas pelo classificador/regressor através de um processo de tentativa e erro (Bishop; Nasrabadi, 2006).

Tendo em vista os conceitos básicos de aprendizagem de máquina, estabelecemos uma base sólida para explorar os principais modelos preditivos aplicados à previsão do nível de água. Estes conhecimentos, nos proporciona as ferramentas necessárias para abordar problemas complexos de previsão. Ao avançarmos para a próxima seção, adaptaremos esses conhecimentos teóricos às percepções específicas do contexto da previsão do nível de água, explorando como esses modelos podem ser ajustados e otimizados para lidar com os desafios únicos apresentados por esse tipo de série temporal.

2.4.6 Modelos

Nesta seção serão apresentados detalhes dos principais modelos utilizados para a previsão do nível de água.

2.4.6.1 Modelagem de Séries Temporais

Existem diversos modelos utilizados para realizar a previsão em séries temporais. Os modelos propostos por *Box-Jenkins* (Box et al., 2015) são os mais utilizados, pois conseguem caracterizar, de maneira adequada, uma grande variedade de séries temporais. Esses modelos são classificados como paramétricos lineares univariados e envolvem apenas uma série de tempo. Essa metodologia explora a correlação temporal existente entre os valores da série.

Uma questão importante, que auxilia no processo de identificação de um modelo adequado de série temporal e, consequentemente, proporciona maior acurácia na predição, é a análise da FAC (Função de Autocorrelação). É por meio da FAC que é possível descobrir a correlação existente entre as observações atuais e passadas da série. Desse modo, uma FAC estabelece uma relação linear entre o presente e seu passado. A FAC é representada por meio de um conjunto de medidas de autocorrelação entre Z_t e Z_{t-1} , denominados de coeficientes α. Essas medidas são resultantes da divisão da covariância da série temporal nas defasagens, pela variância da população(Morettin; Toloi, 2018).

2.4.6.2 Modelo AR - Autoregressive

No modelo AR, os valores passados da variável aleatória são usados para explicá-la, além de um erro aleatório. Quando há autocorrelação entre as observações da série, é que se dá origem ao nome do modelo, autorregressive (autorregressivo). Esse modelo é paramétrico, ou seja, deve-se estabelecer sua ordem, que indica o número de defasagens a serem utilizadas como preditores, ou seja, a quantidade de valores passados são usados para realizar a predição do atual (Box et al., 2015).

Dada uma série temporal Z_t no instante t, então se pode chamar Z_t de um processo $AR(p)$, ou seja, autorregressivo de ordem p . Neste caso, p é a quantidade de parâmetros a um erro aleatório (Morettin; Toloi, 2018). Desse modo, a modelagem de Z_t por meio de um modelo $AR(p)$é dada por:

$$Z_t = \alpha Z_{t-1} + \alpha Z_{t-2} + ... + \alpha Z_{t-p} + \sigma_t \tag{5}$$

onde: Z_t é a observação no instante t, com t=1,2...,n; α é o parâmetro do modelo, com i=1,2...,n; Z_{t-1} é a observação no instante $t - i$ i=1,2...,n; σ_t erro aleatório no instante t, com média 0;

2.4.6.3 Modelo MA(Moving Average)

No modelo MA, o termo *Moving Average* (média móvel) representa as defasagens dos erros, passado e presente, também denotado por ruído branco. Ele é aplicado quando existe autocorrelação entre os valores da série. As médias móveis são interpretadas como mudanças de valores não previstos, ou seja, eventos externos ou choques no sistema ao qual a série temporal representa. Desse modo, o componente de erro no instante t, se relaciona ao valor da série no instante t + 1 (Box et al., 2015).

Um modelo $MA(q)$ indica o grau da defasagem entre o erro e a variável da série, assim, um modelo $MA(1)$ indica que a série é influenciada pelo erro defasado de um período. O modelo $MA(q)$, é dado por(Morettin; Toloi, 2018):

$$Z_t = \sigma_t - \theta_1 \sigma_{t-1} - \theta_2 \sigma_{t-2} - ... - \theta_q \sigma_{t-q} \tag{6}$$

onde: Z_t é a observação no instante t, com t=1,2...,n; θ_j é observação do modelo, com i=1,2...,q; σ_{t-j} erro aleatório no instante $t - j$, com j=1,2,...,q;

2.4.6.4 Modelo ARMA(Autoregressive Moving Average)

O modelo $ARMA(p, q)$ é a combinação dos modelos $AR(p)$ e $MA(q)$. Esse modelo é empregado em séries originadas a partir de sistemas que possuem características autorregressivas e de médias móveis (Box et al., 2015). Um modelo $ARMA(p, q)$ indica a ordem p de um processo autorregressivo e o grau q de defasagem entre o erro e a variável da série (Morettin; Toloi, 2018). A Equação que representa o modelo

$ARMA(p, q)$:

$$Z_t = \alpha Z_{t-1} + \alpha Z_{t-2} + ... + \alpha Z_{t-p} + \sigma_t - \sigma_t - \theta_1 \sigma_{t-1} - \theta_2 \sigma_{t-2} - ... - \theta_q \sigma_{t-q} \quad (7)$$

onde: Z_t é a observação no instante t, com t=1,2...,n; θ_j é observação do modelo, com i=1,2...,q; σ_{t-j} erro aleatório no instante $t - j$, com j=1,2,...,q; α é o coeficiente autoregressivo, com i=1,2...,n; σ_t erro aleatório no instante t, com média 0;

2.4.6.5 ARIMA(Autoregressive Integrated Moving Average)

Um dos modelos mais utilizados para a previsão do nível da água é o modelo ARIMA, este modelo combina características auto-regressivas onde o objetivo é explicar o efeito do *momentum* em uma série, em outras palavras, significa quantos valores no passado que se deve considerar para uma determinada previsão. O modelo ARIMA é representado também pelo componente de médias móveis que tenta capturar o efeito do ruído em uma série, que pode ser entendido como efeitos inesperados em uma série. A integração é tirar uma diferença da série temporal, subtraindo o valor anterior de cada valor, tendendo a tornar os dados mais estacionários.

Em outra perspectiva, pode-se afirmar que o modelo ARIMA é aplicado quando as séries apresentam não estacionariedade, ou seja, não se desenvolvem no tempo em torno da média constante. Essa característica é presente em séries geradas por muitos sistemas. Para remover o componente não estacionário das séries, o modelo $ARIMA(p, d, q)$ possui um parâmetro de diferenciação d, que é a quantidade de vezes que a série tem de ser diferenciada até atingir a estacionariedade. Os parâmetros p e q vem dos modelos $AR(p)$ e $MA(q)$ (Morettin; Toloi, 2018). Assim, pode-se representar o modelo $ARIMA(p, d, q)$ por:

$$F_t = \alpha F_{t-1} + \alpha F_{t-2} + ... + \alpha F_{t-p} + \sigma_t - \theta_1 \sigma_{t-1} - \theta_2 \sigma_{t-2} - ... - \theta_q \sigma_{t-q} \quad (8)$$

onde: F_t é o valor da série original θ_j é o coeficiente associado aos termos de média móvel; σ_{t-j} erro aleatório no instante $t - j$, com j=1,2,...,q; α é o coeficiente associado aos termos autorregressivos, com i=1,2...,n;

No que se refere à precisão do modelo de previsão Arima este é significativamente afetada pelo ruído do conjunto de dados de uma série temporal. A natureza não linear e não estacionária das séries temporais pode resultar em incertezas em aplicações diretas do modelo ARIMA.

2.4.6.6 Metodologia de Box-Jekins

A metodologia desenvolvida por *Box-Jenkins* foi concebida com a finalidade de ajustar um modelo ARIMA(p,d,q) a uma série temporal, de modo que esse modelo

represente essa série de maneira adequada (Box et al., 2015). Após encontrar o modelo mais adequado, é realizada a predição, utilizando-o. Em resumo, a metodologia de *BoxJenkins* consiste em identificar os três parâmetros do modelo ARIMA(p,d,q) que melhor representem uma série temporal. O diagrama dessa metodologia pode ser visto na 9.

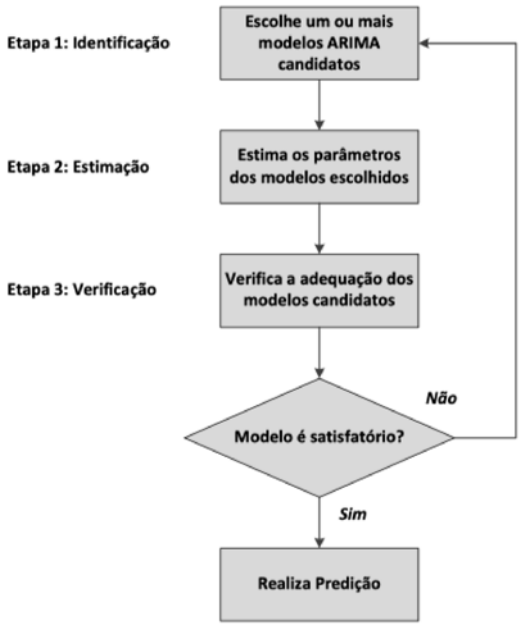

Figura 9 – Etapas da abordagem interativa para a construção do modelo ARIMA, proposta por Box-Jenkins.

Essa metodologia possui um ciclo procedimentos (Morettin; Toloi, 2018): Identificação: nesta fase, são encontrados os valores adequados para p, d,q, com base na FAC (Função de Autocorrelação Simples) e na FACP (Função de Autocorrelação Parcial);Estimação: nesta fase, são estimados o parâmetro autorregressivo e o parâmetro de médias móveis, por meio do método de máxima verossimilhança;Verificação ou diagnóstico: nesta fase, são analisados os resíduos e as funções FAC e FACP com o intuito de verificar se o modelo encontrado se ajusta adequadamente à série temporal.

2.4.6.7 Arima com Variáveis Exógenas

O modelo ARIMA *AutoRegressive Integrated Moving Average* com variáveis exógenas é uma técnica poderosa na análise de séries temporais, pois permite incorporar fatores externos que podem afetar a série em análise. Essas variáveis exógenas fornecem informações adicionais que podem melhorar a precisão das previsões e capturar relações complexas nos dados (Hyndman; Athanasopoulos, 2018).

Uma das principais vantagens do modelo ARIMA com variáveis exógenas é sua aplicabilidade em diversos campos, como economia, finanças e meteorologia. Por exemplo, em economia, pode ser usado para prever o impacto de políticas governa-

mentais nas vendas de um produto, incorporando variáveis como taxas de juros ou políticas fiscais nas previsões.

A incorporação de variáveis exógenas no modelo ARIMA também é valiosa na previsão de demanda em cadeias de suprimentos. Por exemplo, em uma empresa de varejo, pode-se incluir informações sobre promoções sazonais ou eventos especiais como variáveis exógenas para melhorar as estimativas de demanda e otimizar os níveis de estoque.

Além disso, a pesquisa acadêmica tem explorado extensivamente o uso do ARIMA com variáveis exógenas em previsões climáticas, incorporando dados meteorológicos como temperatura e precipitação para melhorar as previsões de eventos climáticos extremos, como tempestades e secas (Hyndman; Athanasopoulos, 2018).

Em resumo, o modelo ARIMA com variáveis exógenas é uma ferramenta versátil que pode ser aplicada em uma ampla gama de setores para melhorar a precisão das previsões, incorporando fatores externos relevantes nos modelos de séries temporais. Sua capacidade de adaptação a diferentes domínios torna-o uma técnica valiosa para analisar e prever dados temporais complexos.

2.4.6.8 Random Forest

Métodos baseados em árvores representam uma alternativa interessante para a construção de modelos preditivos quando a relação entre os preditores e a resposta de interesse é não linear e complexa. As árvores de regressão, para respostas contínuas, e as de classificação, para respostas categóricas, representam as abordagens mais simples (Hastie et al., 2009).

Para ambos os casos descritos acima, o algoritmo pode ser resumido em duas etapas: 1) partição do espaço dos preditores, ou seja, do conjunto de valores possíveis para $x_1, x_2, ..., x_p$, em j regiões distintas e disjuntas $r_1, r_2, ..., r_j$. 2) ajuste de um modelo simples (usualmente uma constante) para predição da resposta de interesse em cada região. Como o conjunto de regras utilizado para o particionamento do espaço dos preditores pode ser completamente descrito por uma árvore, esta abordagem é conhecida como árvore de decisão (Hastie et al., 2009).

Árvores de decisão, geram um conjunto de condições compreensíveis, ou seja, relativamente fáceis de se compreender. Devido à sua estrutura de construção, podem ser aplicadas a conjuntos de treinamento que apresentam dados faltantes ou diferentes tipos de preditores (esparsos, assimétricos, contínuos, categóricos, entre outros) sem a necessidade de pré-processamento, não requerem a especificação de uma forma para a relação entre os preditores e a resposta, e podem, ainda, conduzir à seleção de variáveis. (Kuhn; Johnson et al., 2013)

Todavia, os modelos de predição resultantes de árvores de decisão são, em geral, instáveis, ou seja, pequenas modificações na parte de treinamento podem implicar em

alterações na estrutura da árvore ou em suas regras e, consequentemente, alterar a interpretação do modelo ajustado. Logo, diversos métodos foram desenvolvidos a fim de melhorar o desempenho preditivo de árvores de decisão. Uma dessas formas é a construção de múltiplas árvores,a fim de combinar suas predições em uma predição final, mais acurada. (James et al., 2013)

Ensemble Learning tem como fundamento a utilização de vários modelos, treinados sobre os mesmos dados, calculando a média dos resultados de cada modelo para problemas de regressão e número de votos para problemas de classificação, encontrando um resultado preditivo com o maior desempenho. O requisito, para o *ensemble learning* é que os erros de cada modelo (neste caso, árvore de decisão) sejam independentes e diferentes de árvore para árvore.

Random Forest, é uma técnica de aprendizado de máquina usada para resolver problemas de regressão e classificação. Ela utiliza uma abordagem de aprendizagem por conjunto, que é uma técnica que combina muitos classificadores ou regressores para fornecer soluções para problemas complexos. Este processo consiste em muitas árvores de decisão. A floresta gerada pelo algoritmo é treinada por meio de *Bagging* (Breiman, 1996) ou *Bootstrap Agregation* .*Bagging* é um algoritmo de conjunto que melhora a precisão dos algoritmos de aprendizado de máquina.

Bootstrap ou *Bagging*, é uma funcionalidade que envolve a retirada de amostras com repetição para estimar um parâmetro da população como média. É um modelo relativamente preciso e eficiente para se estimar a distribuição de uma população. No *Bagging*, são utilizadas amostras, esses subconjuntos incluem todos os preditores e uma árvore de decisão é processada em cada uma desses subconjuntos, previsões de cada modelo são combinadas para obter um resultado, esses modelos aprendem independentemente uns dos outros e de forma paralela.

Em um problema de regressão, o *Bagging* consiste em treinar diversas vezes uma mesma árvore de regressão com versões de amostras bootstrap dos dados de treinamento e calcular a média do resultado. Já em um problema de classificação, a técnica consiste em utilizar um comitê de árvores onde cada uma lança um voto para a classe predita. Independentemente do problema em questão, a ideia é a mesma: calcular a média de modelos com pouco viés e muito ruído para reduzir a variância (Hastie et al., 2009)

O modelo *Random Forest*, combina árvores de decisão de forma aleatória da mesma forma que o *Bagging*, entretanto, no RF, é utilizado em um conjunto de preditores, não todos, razão para fazer isto é diminuir a correlação entre as árvores em uma amostra comum. Se um ou mais preditores fortes forem selecionando em uma, ou mais árvores, isto pode gerar correlação entre elas.

Um atributo importante para o *Random Florest* é a entropia que representa o quanto um parâmetro possui de ganho de informação, em outras palavras, dado um

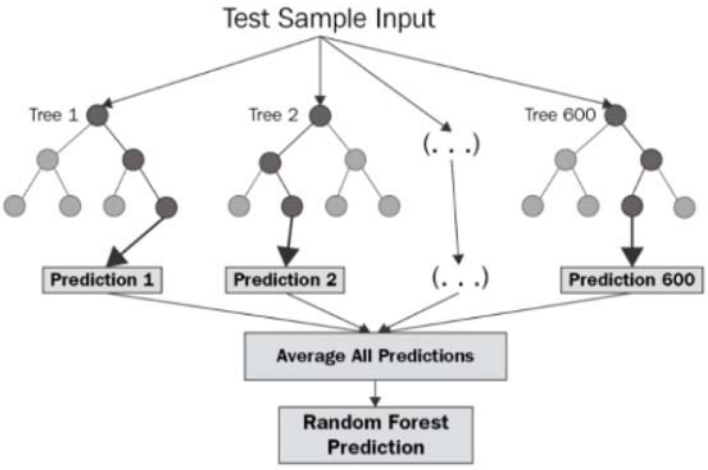

Figura 10 – Visão Geral do Modelo Random Forest (Hastie et al., 2009)

conjunto, define-se o grau de pureza deste conjunto, entretanto, a entropia é usada para medir a impureza em problemas de classificação, em problemas de regressão, o *Random Forest* utiliza métricas como o MSE ou a variância para avaliar e otimizar a qualidade das divisões nos nós das árvores.

A Figura 10, relata a estrutura do *Random Forest*. Observa-se que as árvores seguem em paralelo sem interação entre elas. Uma RF opera construindo várias árvores de decisão durante o tempo de treinamento e produzindo a média das classes como a previsão de todas as árvores. Para entender melhor o algoritmo RF, verificam-se as seguintes etapas: É definido aleatoriamente k pontos de dados do conjunto de treinamento. Elabora-se uma árvore de decisão associada a esses k pontos de dados, logo após escolha-se o número N de árvores que deseja construir e repita os passos anteriores. Para um novo ponto de dados, faça com que cada uma de suas árvores $N - tree$ preveja o valor de y para o ponto de dados em questão e atribua o novo ponto de dados à média de todos os valores de y previstos.

Dadas essas definições, entende-se que RF é uma modificação substancial do *Bagging* que constrói uma grande coleção de árvores não correlacionadas e, então, calcula a média delas (Hastie et al., 2009). Considerada uma das principais ferramentas para a predição e a análise de dados, a técnica é interessante por apresentar uma alta acurácia, não ser suscetível ao super-treinamento e, também, poder lidar com amostras pequenas, espaços de características de alta dimensão e estruturas de dados complexas.

No que se refere a hiperparâmetros, o RF pode ser ajustado através de sua profundidade máxima, que se for um valor muito pequeno pode acontecer de fornecer para o algoritmo pouca flexibilidade para a captura de padrões. Entretanto, se este valor for alto pode gerar *overfitting* sendo este termo relacionado ao ajuste excessivo aos dados de treino. Outro parâmetro, é número máximo de preditores que a cada vez que

ocorre uma divisão, o algoritmo pega dois parâmetros que tem menor impureza e cria dois ramos.

Outros hiperparâmetros são: número de árvores, que controle o número de árvores que um regressor pode conter. Outro fator a ser ajustado é a quantidade mínima de amostras que um nó interno pode conter.

2.4.6.9 Máquina de Vetores de Suporte

A máquina de vetor de suporte do inglês *Suport Vector Machine* (SVM), é atualmente um tópico importante na área de aprendizado estatístico. Desde o final da década de 1990, as abordagens tradicionais de redes neurais sofreram sérias dificuldades com generalização que é a principal razão para a o desenvolvimento das SVMs. No aprendizado estatístico, as máquinas de vetor de suporte são um método de aprendizado supervisionado com algoritmos que analisam o conjunto de dados. Ele foi introduzido pela primeira vez como um método para resolver problemas de classificação. No entanto, devido a muitas características interessantes, foi estendida à área de análise de regressão, foco desta tese.

As SVMs representam uma generalização de um algoritmo denominado classificador de margem máxima $(maximalmarginclassifier)$, que acomoda fronteiras não lineares em problemas de classificação. Adicionalmente, o SVM difere dos demais algoritmos utilizados para classificação por não estimar, diretamente, probabilidades, mas sim a classe da resposta de interesse para uma nova observação.

Em geral, se os dados podem ser separados usando um hiperplano, então existirão infinitamente muitos desses hiperplanos. Para decidir qual separação do hiperplano pode-se utilizar, necessita-se usar o classificador de margem máxima, que também é conhecido como o hiperplano de separação ideal. O classificador de margem máxima está mais distante da observação de treino. Isto é, quando calculamos a distância de cada observação ao separar o hiperplano, o mínimo é o que chamamos de 'margem'. O classificador máximo da margem é o hiperplano para o qual a margem é a maior. (James et al., 2013)

A premissa básica desse algoritmo é aprender hiperplano de separação de margem máxima, que é de certa forma parecido com o algoritmo de regressão linear, pois é traçado uma linha de separação. Quanto mais a linha mencionada estiver distante dessas margens, mais o algoritmo irá conseguir fazer a generalização dos parâmetros.

Na Figura 11, observa-se a reconstrução do hiperplano, onde SV representa os vetores de suporte. A aprendizagem deste algoritmo é encontrar esses vetores, tendo esses elementos é possível realizar a construção desta linha, sendo que cada SV são registros na base de dados.

Uma das formas de se criar esta margem separadora é através da envoltória convexa, que consiste em basicamente fazer uma ligação desses vetores, em outras pa-

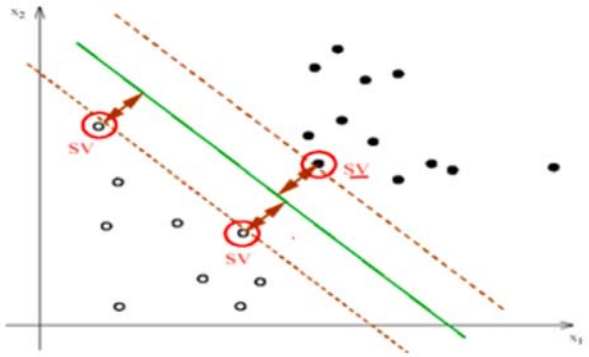

Figura 11 – Margem Separador SVM (Fonte:https://scikit-learn.org/stable/modules/svm.html acesso em:15/08/2023)

lavras uma região convexa é uma região onde, para cada par de pontos dentro da região, cada ponto no segmento de reta que une o par também está dentro da região.

As principais vantagens de se utilizar SVM são: é pouco influenciado por ruído nos dados, aprende conceitos não presentes nos dados originais e é utilizado para classificação e regressão. O fato de aprender novos conceitos é realizado por funções matemáticas chamadas de *kernel*, onde basicamente um conjunto não linear torna-se linear.

Embora esse classificador/regressor seja um algoritmo simples para problemas de classificação, sua implementação só é possível se existir um hiperplano que separe perfeitamente os dados de treinamento conforme a classe a que os mesmos pertencem. No entanto, para a maioria dos problemas reais, não é possível definir esse hiperplano, pois algumas observações de diferentes categorias da variável resposta podem estar sobrepostas (James et al., 2013).

A princípio, o algoritmo de SVMs foi desenvolvido para solucionar problemas envolvendo apenas duas classes. No entanto, com o passar dos anos, extensões permitiram que o algoritmo solucionasse, também, problemas multiclasse e problemas de regressão neste caso, usa-se *Support Vector Regression* - SVR (Brereton; Lloyd, 2010).

Segundo (Verdério et al., 2015) as máquinas de vetores suporte para regressão diferem da técnica de classificação no sentido de que, enquanto a segunda busca dividir os dados em diferentes classes e classificar os dados futuros, a primeira busca encontrar um preditor que aproxime bem os dados de amostra.

O algoritmo de regressão de vetores de suporte utiliza uma função de perda chamada de ϵ *loss function* que ignora erros que estão além de uma certa distância dos valores considerados válidos, ou seja, erros são permitidos somente se forem menores do que ele.

Os modelos baseados em SVM podem ser utilizados em problemas lineares e não lineares. Para o último, onde os parâmetros não são linearmente separados, deve-se

utilizar Kernels: linear, gaussiano, polinomial ou tangente hiperbólica.

As funcionalidades de *kernel* têm como fundamento projetar os vetores de características de entrada em um espaço de características de alta dimensão para classificação de problemas que se encontram em espaços não linearmente separáveis. Esse processo é realizado visto que a medida que se aumenta o espaço da dimensão do problema, aumenta também a probabilidade desse problema se tornar linearmente separável em relação a um espaço de baixa dimensão. Entretanto, para obter uma boa distribuição para esse tipo de problema é necessário um conjunto de treinamento com um elevado número de instâncias (Gonçalves, 2010). A Figura 12 mostra o processo de transformação de um domínio não linearmente separável, em um problema linearmente separável através do aumento da dimensão, onde é feito um mapeamento por uma função de *kernel* $F(x)$.

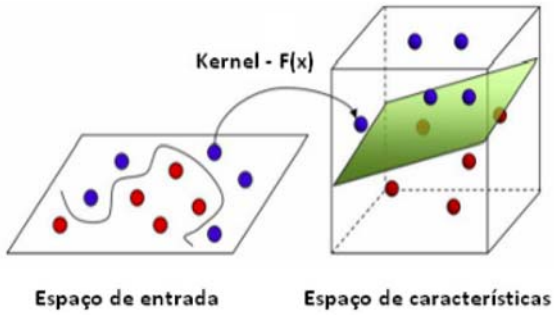

Figura 12 – Transformação: problema não linearmente separável em um problema linearmente separável (Gonçalves, 2010)

O *kernel Radial Base Function* (RBF) (Chen; Cowan; Grant, 1991) é muito utilizado para resolução de problemas de aprendizagem, inclusive é usado computacionalmente como padrão em muitas bibliotecas de linguagens de programação que utilizam o algoritmo SVM. No RBF, diferentemente do *kernel* linear, é possível resolver problemas, originalmente, não linearmente separáveis, através do mapeamento para um espaço de maior dimensão. Existem dois parâmetros que podem ser alterados em busca de um melhor resultado para o aprendizado do classificador/regressor, são eles: γ e C. A constante C representa um hiperparâmetro referente ao a severidade de violações, ou seja, classificações incorretas com relação à margem e ao hiperplano. γ, quanto maior o seu valor, mais tentará ajustar-se exatamente ao conjunto de dados de treinamento, isto é, erro de generalização e causar problemas de ajuste excessivo

Dessa forma, C controla o limiar ou viés/variância do algoritmo de aprendizado: conforme seu valor aumenta, maior a margem e, portanto, mais tolerante a classificações incorretas será o modelo ajustado, situação equivalente a obter um classificador mais viesado, porém com menor variância. Por outro lado, a medida que C diminui, a margem torna-se mais estreita, de modo que a mesma será raramente violada, o que conduz a um modelo com viés baixo, porém com variância alta (James et al., 2013).

O *kernel* polinomial é uma representação mais generalizada do linear, não é tão preterido quando as outras funções de *kernel*, pois é menos eficiente e preciso. Neste tipo de *kernel* simplesmente calcula-se o produto escalar aumentando a potência do *kernel*.

2.4.6.10 Redes Neurais

Desde a criação dos primeiros estudos em redes neurais, a criação do modelo perceptron (Rosenblatt, 1958) consistia na imitação do que seria um neurônio do cérebro humano. Neste modelo, vários sinais ou valores (análogos aos sinais elétricos das sinapses dos neurônios do cérebro) são ponderados para gerar uma saída calculada. Na Equação abaixo, xi corresponde cada um dos sinais de entrada, wi corresponde aos pesos treinados e y à saída calculada. As entradas são aplicadas à rede neural. A partir das entradas, as saídas da rede neural são calculadas. As diferenças entre as saídas calculadas e as saídas esperadas compõem um vetor de erros. O vetor de erros é utilizado para ajustar os pesos.

$$y = \sum_{i=1}^{n} (x_i * w_i) \tag{9}$$

As propriedades básicas de qualquer rede neural são as entradas, os pesos e os neurônios com suas funções de ativação. As entradas simulam um sinal de estímulo, os pesos são treinados para alguma finalidade específica e as funções de ativação moderam o valor das entradas com os pesos. Dependendo da função de ativação o comportamento da saída pode variar e dentro das possíveis funções estão a tangente hiperbólica, ReLU, linear, sigmóide, entre outras.

Vários métodos foram desenvolvidos para chegarmos ao estágio atual como o algoritmo *backpropagation* (Kelley, 1960) (Rumelhart; Hinton; Williams, 1986), a regra da cadeia, para calcular as derivadas e conseguirmos o treinamento em redes neurais, os algoritmos de treinamento como o ADAM (nome derivado de *Adaptative Moment Estimation* ou Estimativa Adaptativa usando Momento) para treinamento adaptativo, técnicas para evitar *overfitting* como *dropout* e normalização. A normalização consiste em dividir todos os valores pelo valor máximo do conjunto de valores, fazendo com que todos os valores fiquem entre zero e um. Estes avanços vêm conseguindo atingir bons patamares em treinamentos e resultados de redes.

Os três tipos distintos de treinamentos de redes neurais mais utilizados são: o supervisionado, onde temos exemplos com resultados conhecidos para o treinamento da rede. Estes exemplos são então calculados pela rede neural e a saída comparada com as saídas conhecidas. A diferença encontrada é então utilizada para ajuste dos pesos. No treinamento não supervisionado, os alvos dos dados do treinamento não são conhecidos. Então são aplicadas técnicas como, por exemplo, de agrupamentos

para determinar grupos de classes. E no aprendizado por reforço, a rede aprende baseada em reforços positivos ou negativos. Neste trabalho estamos avaliando redes com treinamento supervisionado somente.

2.4.6.11 Aprendizagem Profunda

Com a evolução das redes neurais surgiram as redes neurais profundas do inglês *deep learning*, que empregam técnicas para análise de problemas não lineares, de ordem superior e de tamanho superior. Problemas como processamento de imagem são particularmente mais fáceis de processar utilizando técnicas como convolução. Esses problemas não lineares necessitam de mais camadas nas redes neurais para tratar a complexidade do que está sendo processado. Assim, as redes neurais profundas são assim denominadas pela grande quantidade de camadas. As camadas ocultas, uma ou mais das seguintes camadas:

a) Camada de normalização: normaliza os dados antes de entrar em outra camada. Pode ser colocado no início ou no interior da rede. Normalização é um passo importante, que ajuda a remover a diferenças dimensionais nos dados de entrada e melhorar a capacidade de previsão.

b) Camada de codificação/decodificação: Esta camada faz parte da rede *Autoencoders*. Esta rede representa um conjunto de camadas usadas principalmente para redução de dimensionalidade e extração de recursos.

c) Camada de convolução: esta camada é usada como redução de dimensionalidade como parte da sub-rede normalmente chamada de Rede Convolucional.

d) Camada de rede neural recorrente: Esta camada pode ser de *Long short-term Memory* (LSTM), Gated Recurrent Unit (GRU) ou camada de rede neural clássica com base em redes neurais recorrentes.

De modo geral, o *deep learning* alcança alta flexibilidade e performance por meio da representação dos dados observados como uma hierarquia aninhada de funções matemáticas relativamente simples, cada uma associada a uma camada oculta diferente do modelo. Assim, a cada aplicação de uma função diferente, uma nova representação dos dados de entrada *(input)* é obtida e, quanto mais profundas as camadas, mais abstratos são os recursos extraídos dos dados para obter uma dada representação (Goodfellow; Bengio; Courville, 2016).

Um exemplo interessante relatado por (Goodfellow; Bengio; Courville, 2016) sobre *deep learning* em uma análise de imagem, que consiste em mapear diretamente um conjunto de *pixels* para a identificação de um objeto este representa um problema extremamente desafiador. O *deep learning* analisa esse problema por meio da divisão de um mapeamento único e complexo em uma sequência de mapeamentos aninhados mais simples, cada um descrito por uma camada oculta diferente. Dessa forma, assim como em redes neurais simples, a primeira camada é representada por dados

observados e, na sequência, as camadas ocultas extraem características cada vez mais abstratas da imagem: diante dos *pixels*, a primeira camada oculta pode identificar as bordas da imagem comparando o brilho de *pixels* vizinhos. Em seguida, dada a descrição das bordas, a segunda camada oculta pode procurar por cantos e contornos, que são identificáveis por coleções de arestas. Por fim, a partir da descrição dos cantos e contornos, a terceira camada oculta pode detectar partes inteiras de objetos específicos ao identificar coleções específicas de cantos e contornos. Tal processo, de descrição da imagem por partes, ao final, possibilita o reconhecimento do objeto que a mesma contém.

No início do treinamento, são definidos valores aleatórios para os parâmetros. Portanto, é comum que o resultado do primeiro processo de treinamento não seja o esperado, como exemplo em uma rede de classificação, por exemplo, a rede pode identificar como sendo cachorro uma imagem rotulada como um gato. Portanto, o novo valor de um peso será calculado com base no valor de custo retro-propagado pela camada posterior na rede, através do cálculo do gradiente descendente. Com isso, os pesos serão ajustados para mais ou para menos na direção que o valor de custo for minimizado, de forma que nas próximas atualizações o valor de custo siga convergindo para zero. Para treinamento de uma *Dense Neural Network (DNN)* é geralmente utilizado o gradiente descendente estocástico *Stocastic Gradient Descendent — SGD*, que permite realizar o treinamento em um grande conjunto de dados, visto que ao invés de processar todos os dados do conjunto de treinamento, ele processa um mini-lote de exemplos aleatórios do conjunto em cada iteração. Esse tipo de técnica é interessante visto que geralmente o tamanho das bases de dados para problemas de séries temporais são consideráveis e podem ultrapassar os limites de processamento. Além disso, comumente vários exemplos possuem bastante semelhança, portanto, através desse mini-lote é feita uma estimativa da média do gradiente sobre todos os exemplos.(Goodfellow; Bengio; Courville, 2016).

Quando todos os exemplos utilizados para treinamento são atualizados, ocorre a passagem de uma época, sendo que esta passagem é realizada uma série de vezes para que o modelo obtenha um valor de custo cada vez menor para todos os exemplos de treinamento. Por fim, um novo conjunto de dados (conjunto de testes) é inserido no modelo, com dados que não foram utilizados no treinamento, para verificar se ele atinge resultados semelhantes em exemplos do mesmo domínio do problema não observados anteriormente. Dessa forma, pode ser medida a eficácia do modelo para a resolução do problema para qual foi treinado. Comumente é realizada a chamada etapa de validação durante o treinamento, em que uma pequena parte do conjunto de treinamento é separado, para que no fim de cada época seja realizado uma avaliação, indicando como o modelo está evoluindo para dados que não estão no conjunto de treinamento.

2.4.6.12 Redes Neurais Recorrentes

Existe um tipo de Rede Neural, a qual chama-se de Rede Neural Recorrente RNN, onde são inúmeras as suas aplicações, como, por exemplo: na previsão de um novo *frame* em um filme, em processamento de linguagem natural, na previsão de partes de um poema, não geração de legendas automatizadas em sistemas de *streams* bem como em séries temporais, onde a partir de datas anteriores pode-se estimar valores futuros. Basicamente, isto é possível, pela capacidade dos neurônios em persistir as informações.

Em resumo, Redes Neurais Recorrentes são uma classe de redes neurais que recebe entrada de duas fontes: uma do presente, e outra de um ponto passado. A informação das duas fontes são utilizadas para decidir como reagir a uma nova entrada de dados, e isso é feito através de um circuito de *feedback*, onde a saída de cada instante é uma entrada para o instante seguinte. Devido a essa característica, podemos dizer que elas possuem memória, fazendo das RNNs mais semelhantes da forma como humanos processam informação, possibilitando o reconhecimento de um contexto através da memória (Zaremba; Sutskever; Vinyals, 2014).

As Redes Neurais Recorrentes, são constituídas principalmente por três tipos de camadas: camada de entrada, camadas intermediárias (também chamadas de camadas ocultas) e camada de saída. Cada neurônio na camada de entrada está conectado aos neurônios da primeira camada intermediária, que por sua vez se conectam à próxima camada, e assim sucessivamente até a camada de saída. Em RNNs, é comum encontrar múltiplas camadas intermediárias, permitindo uma maior capacidade de representação e aprendizado de padrões complexos ao longo de sequências temporais.

Em outras palavras, uma rede que possui recorrência é chamada de rede neural recorrente, caracterizada pela realimentação dos neurônios de uma determinada camada pelas suas saídas de um estado ou época anterior.

A RNN tem o problema do *vanishing gradient* e *exploding gradient*, sendo o primeiro conceituado pelo fato dos valores tendem a ficar cada vez menores. Com isso, os neurônios no início da rede neural tendem a levar mais tempo para calcular os pesos. Já o *Exploding Gradient*, ocorre quando o gradiente fica muito maior nas camadas anteriores. Nestes casos, acabam impactando na acurácia e no tempo de processamento da rede. Problema de *vanish gradient* podem ser evitados, optando em utilizar a função de ativação que seja mais adequada para a situação em questão. O LSTM possui uma estrutura semelhante à RNN, mas, em vez de ter uma rede única, cada célula consiste em armazenar e remover memória. As células diferentes são conectadas umas às outras e carregam informações relevantes durante o processamento da sequência.

Existem algumas arquiteturas de RNN, como Elman e Vanilla (Elman, 1990) que são modelos que originaram a arquiteura utilizada neste trabalho que são *Long short-*

term Memory (LSTM).

A arquitetura de uma *Long Short-term Memory (LSTM)* se assemelha à de uma Vanilla, todavia no caso da LSTM, o seu estado interno consiste em um conjunto de sub-redes conectadas recorrentemente chamadas bloco de memória. Esses blocos contêm células de memórias com autoconexões capazes de armazenar o estado temporal da rede, além de unidades especiais chamadas portas (gates) que são responsáveis por controlar o fluxo de informações.

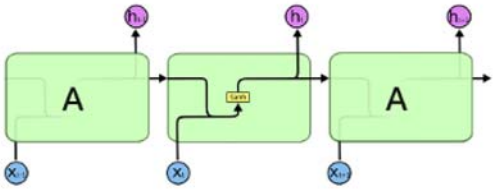

Figura 13 – Neurônio Recorrente (Staudemeyer; Morris, 2019)

Na Figura 13 verifica-se um neurônio básico de uma RNN, onde irá armazenar quais as informações são as mais importantes, para que isto aconteça é aplicada uma função de ativação chamada de Tangente Hiperbólica em cada entrada. Existe outro tipo de RNN, as LTSM que diferentemente da anterior que apontam para elas mesmas, a LTSM, adiciona células e manipulas elas, conforme se observa na Figura 14 .

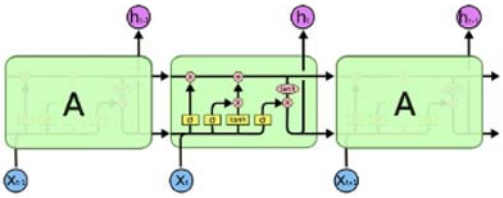

Figura 14 – Estrutura Interna - LSTM (Staudemeyer; Morris, 2019)

Uma LTSM, conforme a Figura 14, consegue aprender dependências de logo prazo. Para isso utiliza-se de uma estrutura de um *pepiline* que flui durante o tempo a informação onde o x apaga a informação e o $+$ adiciona a informação neste meio. O neurônio tem duas entradas, a primeira é a saída do tempo anterior ht_1 e X_t o valor atual e passa para a função sigmoide, que tem como saída 0 ou 1, se a saída for 0 a informação não é importante e é apagado da memória e se for 1 ela continuará no *pepiline*. A função tangente hiperbólica cria um vetor de informações mais utilizadas.

Em outras palavras, LSTM é um tipo especial de rede neural, que pode fornecer uma propagação de erro constante através da rede, isto devido a um projeto diferente da sua rede. O LSTM consiste em blocos de memória com auto-conexões definidas na camada oculta, que tem a capacidade de armazenar o estado temporal da rede. Além da memorização, a célula LSTM possui unidades multiplicativas especiais, chamadas de portas, que controlam o fluxo de informações. Cada bloco de memória consiste na

porta de entrada, que controla o fluxo da ativação de entrada na célula de memória, e na porta de saída, que controla o fluxo de saída da ativação da célula. Além disso, a célula LSTM também contém a porta de esquecimento, que filtra as informações da entrada e da saída anterior e decide qual deve ser lembrada ou descartada. Com essa filtragem seletiva de informações, a porta de esquecimento dimensiona o estado interno da célula LSTM, que é auto-recorrentemente.

Após a exploração dos aspectos teóricos envolvidos nos modelos preditivos aplicados à previsão do nível de água, é crucial avançar para a compreensão das métricas de avaliação desses modelos. Esta próxima seção abordará de forma objetiva as métricas utilizadas para avaliar a eficácia e a precisão dos modelos desenvolvidos, fornecendo uma visão abrangente sobre como determinar a qualidade das previsões e identificar possíveis áreas de melhoria. Ao compreendermos as métricas de avaliação de forma mais detalhada, pode-se otimizar os modelos preditivos e garantir resultados confiáveis na gestão do nível de água neste contexto específico.

2.5 Indicadores de Erro

No método de Aprendizado Supervisionado, o conjunto de dados contém variáveis dependentes ou de destino, junto com as variáveis independentes. Neste método são contruídos modelos usando variáveis independentes, sendo previsto variáveis dependentes ou alvo. Se a variável dependente for numérica, os modelos de regressão são usados para predizê-la. Nesse caso, o *Mean Absolute Error*(MAE) e *Mean Squeare Error*(MSE) pode ser usados para avaliar modelos.

O *Mean Absolute Error* e o *Mean Square Error* são medidas básicas de análise de desempenho em problemas de regressão. O MAE é calculado subtraindo-se todos os valores em relação à média no final somando esses resultados. A palavra absoluto se refere ao módulo que desconsidera o sinal negativo. O MSE é calculado pegando o resultado do MAE e elevando ao quadrado. O MSE, penaliza os valores que estão mais distantes da média.

O MAE mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção. Ele mede a precisão para variáveis contínuas. Em outras palavras, o MAE é a média sobre a amostra de verificação dos valores absolutos das diferenças entre a previsão e a observação correspondente. O MAE é uma pontuação linear, o que significa que todas as diferenças individuais são ponderadas igualmente na média.

O MSE mede a quantidade de erros em modelos estatísticos avaliando a diferença quadrática média entre os valores observados e previstos. Quando um modelo não tem erro, o MSE é igual a zero. À medida que o erro do modelo aumenta, seu valor aumenta. Em outras palavras, o MSE é a distância média ao quadrado entre os valores

observados e previstos. Por usar unidades quadradas em vez de unidades de dados naturais, a interpretação é menos intuitiva.

Elevar ao quadrado as diferenças elimina valores negativos para as diferenças e garante que o erro quadrático médio seja sempre maior ou igual a zero. É sempre um valor positivo. Apenas um modelo perfeito sem erro produz um MSE de zero. E isso não ocorre na prática.

Outra métrica é o RMSE, que ao envolver a raiz quadrada, é menos sensível a outliers em comparação com o MSE, pois a raiz quadrada tende a "penalizar"erros grandes de forma proporcionalmente menor.

$$MSE = \frac{\sum_{i=1}^{D}(x_i - y_i)^2}{n} \tag{10}$$

$$RMSE = \sqrt{MSE} \tag{11}$$

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{12}$$

Observando-se as fórmulas acima, temos: x representando o valor real e y o valor predito, a distância entre esses valores elevado ao quadrado é o MSE e o módulo o MAE. A raiz quadradas do MSE é o RMSE.

Outra medida a ser considerada em problemas de regressão é o desvio padrão que mede o grau de dispersão de um conjunto de dados, sendo x valor em uma posição no conjunto de dados, M media aritmética dos dados e n a quantidade de dados.

$$Dp = \sqrt{\sum \frac{(x_i - \bar{M})^2}{n}} \tag{13}$$

O coeficiente de determinação $R\hat{}2$, é uma métrica de avaliação que, analisa o quanto um determinado modelo é melhor que a pior situação possível, por exemplo, a média, em outras palavras, o quanto um determinado modelo explica da variância dos dados. Quanto temos um valor de $R\hat{}2$ de 0,87, por exemplo, pode-se dizer que o modelo é 87% melhor que a média.

Quando estamos avaliando um problema de regressão, teoricamente, quando maior a quantidade de atributos previsores ele possuir mais preciso será este modelo, entretanto, em alguns casos alguns tributos previsores podem não contribuir tanto para o desempenho de um modelo, a partir disso existe o coeficiente de determinação ajustado.

$$R^2 = 1 - \frac{\sum_{i=1}^{i}(x_i - y_i)^2}{\sum_{i=1}^{i}(x_i - M_i)^2} \tag{14}$$

Mesmo o coeficiente de determinação ser uma forma interessante de se avaliar

uma série temporal, ela não pondera a quantidade de previsores o que, por exemplo, pode ter um custo computacional maior desnecessário. Analisando as fórmulas acima, o R^2 é composta pelo somatório do valor real subtraído pelo valor predito y_i sendo esta primeira parte a variação do modelo, no denominador temos o somatório da subtração de x_i pela M média aritmética, a parte do denominador é considerada a variação em relação à média.

2.6 Divisão dos Dados

Em contextos científicos e de modelagem, a correta divisão de conjuntos de dados em treinamento, validação e teste desempenha um papel fundamental na construção e avaliação robusta de modelos. Esta prática representa uma salvaguarda contra armadilhas que podem comprometer a generalização e a confiabilidade dos resultados obtidos.(Hastie et al., 2009)

Nesse contexto, o conjunto de treinamento é utilizado para alimentar o modelo durante seu desenvolvimento. Nele, o algoritmo aprende padrões, ajusta parâmetros e se adapta à complexidade dos dados. No entanto, apenas avaliar o desempenho em dados de treinamento pode levar a uma otimização excessiva, na qual o modelo se adapta demais aos detalhes específicos desse conjunto, perdendo a capacidade de generalização para novos dados. (Hastie et al., 2009)

A introdução do conjunto de validação é crucial para mitigar esse problema. Durante o treinamento, o conjunto de validação é usado para ajustar hiperparâmetros e escolher a melhor configuração do modelo. Esta etapa permite avaliar o desempenho em dados não utilizados no treinamento, ajudando a evitar o sobreajuste e garantindo que o modelo seja capaz de lidar com uma variedade mais ampla de cenários.(Raschka; Mirjalili, 2019)

O conjunto de teste representa a última fase de avaliação. Ele permanece intocado durante o desenvolvimento do modelo e é reservado para avaliar a capacidade de generalização final. A avaliação no conjunto de teste fornece uma estimativa mais realista do desempenho do modelo em condições do mundo real, garantindo que os resultados sejam confiáveis e replicáveis.(Raschka; Mirjalili, 2019)

Assim, a divisão adequada em treinamento, validação e teste não apenas fortalece a validade científica dos modelos, mas também contribui para a transparência e confiança nas conclusões alcançadas. Essa prática consagrada é um pilar na construção de modelos sólidos e na aplicação responsável da modelagem em diversos domínios científicos.(Raschka; Mirjalili, 2019)

Em séries temporais, a ordem dos dados é crucial, pois cada ponto de dados é dependente dos pontos anteriores, formando uma sequência temporal que deve ser mantida. Quando se faz a divisão dos dados para treinamento e teste de modelos

preditivos, utilizar uma amostragem aleatória pode levar a uma situação onde o modelo é treinado com dados futuros e testado com dados passados. Isso cria um cenário irrealista e não aplicável a problemas do mundo real, onde previsões futuras devem ser feitas exclusivamente com base em dados históricos. Por exemplo, ao prever vendas de uma loja, usar vendas de um dia futuro para prever um dia passado distorce a análise, porque no dia em questão, as informações futuras não estariam disponíveis.

Após a análise dos conceitos relacionados à divisão de conjuntos de dados, é pertinente contextualizar essa prática dentro do panorama mais amplo de pesquisas e trabalhos correlatos ao tema da previsão do nível de água. Na próxima seção, serão revisados estudos anteriores e trabalhos que exploraram diversas abordagens metodológicas e técnicas para prever o nível de água em canais similares ao Canal São Gonçalo. Essa revisão proporcionará uma visão abrangente das estratégias adotadas, dos desafios encontrados e das descobertas alcançadas por pesquisadores na área, fornecendo percepções para o desenvolvimento de abordagens eficazes e inovadoras na previsão do nível de água neste contexto específico.

3 TRABALHOS RELACIONADOS

Nesta seção serão abordados os trabalhos relacionados sobre a previsão do nível de água.

Este capítulo apresenta a revisão da literatura realizada para compreender como a previsão do nível de água tem sido explorado nos últimos anos. Para tal, foi adotado o processo de Mapeamento Sistemático (MS)

Para a busca de estudos foram utilizados os principais motores de busca de estudos científicos, sendo eles *IEEE Explore*, *Science Direct*, *ACM* e *Google Scholar*, nos quais foram mapeados estudos completos publicados em revistas ou anais de evento. Visto que alguns motores resultaram em um número considerável de estudos, foram analisados somente os primeiros 10 de cada motor.

Na busca inicial foi encontrado o total de 100 artigos usando a *string* abaixo. A *string* foi adaptada conforme o motor de busca para englobar todos os termos e conectivos lógicos presentes nela.

> ((’water level forecast’ OR ’water level’) AND (’machine learning’ OR ’deep learning’ or ’time series’))

Para os artigos inicialmente, foi realizada a leitura do Título, Resumo e Palavras-chave, observando os critérios de inclusão (CI) e exclusão (CE) presentes na Tabela 1. Para ser incluído, o estudo deveria obedecer todos os CI e não ser categorizado como nenhum dos CE.

Utilizou-se como *string* de busca: séries temporais, associado à predição de nível de água, onde teve como intervalo de tempo os trabalhos posteriores à 2017, conforme se pode observar na tabela 1

Com isso, o montante de estudos foi reduzido para 20, os quais foram lidos na íntegra. Durante a leitura dos artigos que compreenderam a lista final, foram catalogadas, além de informações gerais (Título, Revista/Evento publicado e Ano) as seguintes informações de cada estudo, para entender a contribuição de cada um e organizá-los em padrões.

- ***Modelo***: Modelos utilizados para a previsão do nível de água.

Tabela 1 – Critérios de Seleção

Inclusão	Exclusão
CI-1 Artigo completo;	CE-1 Artigo incompleto ou sem resultados;
CI-2 Artigo com acesso gratuito;	CE-2 Artigo com acesso pago;
CI-3 Relacionado com predição de Nível de Água;	CE-3 Artigo não relacionado com predição de Nível de Água;
CI-4 Implementados em Rios ou Bacias;	CE-4 Artigos não implementados em Rios ou Bacias;
CI-5 Posteriores a 2017	CE-5 Inferiores a 2017
CI-6 Artigo escrito em Inglês;	CE-6 Revisões da literatura;
CI-7 Publicado em Revistas ou Anais de Eventos.	CE-7 Teses, Dissertações ou TCC's;
	CE-8 Artigo não escritos em inglês.

- ***Entrada***: Parâmetros utilizados para a entrada no modelo, ou seja, se utiliza somente dados de nível ou também fatores ambientais.
- ***DataSet***: Tamanho do *DataSet* utilizado.
- ***Híbrido***: Se utiliza somente um modelo preditívo ou a combinação de modelos para a previsão do nível de água.

Na Tabela 2 observam-se um resumo dos modelos utilizados para a previsão de séries hidrológicas. Inicialmente, foram analisados os modelos mais utilizados para a previsão de séries hidrológicas e alguns trabalhos foram propostos. Devido às características do problema em questão, os modelos exclusivos de séries temporais utilizados foram o ARIMA e RNN tendo em vista os bons resultados encontrados na literatura.

Alguns trabalhos foram descartados por não se enquadrarem no escopo desta proposta. Por exemplo, alguns são destinados ao acionamento de uma determinada bomba de água segundo determinados parâmetros, problema este de classificação e não de regressão.

Continuando a análise da Tabela 2, verificam-se alguns trabalhos publicados e o modelo ARIMA aparece na maioria deles, principalmente quando a sua utilização é destinada em dados lineares e com margem de previsão não tão grande, como em (Ghimire, 2017) e (Yu et al., 2017).

Outros modelos analisados como: RF, SVR e RNN e com um considerável desempenho é relatado em (Wang; Wang, 2020),(Zhang et al., 2020)e (Hrnjica; Bonacci, 2019).

Mesmo com relativo sucesso na utilização destes modelos para a previsão do Nível de água, alguns fatores ambientais como velocidade do vento, insolação, entre outros não são considerados na maioria dos trabalhos. Assim sendo, com a adição de dados

Tabela 2 – Revisão de Literatura

Referência	Modelo	Entrada	Conjunto	Híbrido
Birylo,2018	ARIMA	Nível de Água e Dados Climatológicos	1979-2015	Não
Ghimire,2017	ARIMA	Nível de Água	2000-2006	Não
Sun, 2017	ANN	Nível de Água e Dados Climáticos	2007-2012	Não
Yu, 2017	ARIMA	Nível de Água	2012-2015	Não
Zhang, 2020	RNN	Nível de Água	91 meses	Não
Hrnjica, 2019	RNN, FFN	Nível de Água	38 anos	Não
Wang, 2020	RF,KNN,MLR,GP,M5P	Nível de Água e Dados Climáticos	2002-2014	Não
Choi, 2019	DT,RF,SVM	Nível de Água e Dados Climáticos	2009-2015	Não
Castillo, 2018	SVR,ELM,GPR e ANN	Nível de Água	1997-2015	Não
Xu,2019	ARIMA/RNN	Nível de Água e Dados Climáticos	2009-2018	Sim
Xie, 2019	ARIMA/SVR	Nível de Água	2010-2015	Sim
Nguyen, 2020	ARIMA,RF,KNN,RNN e SVR	Nível de Água e Dados Climatológicos	2008-2015	Sim

não lineares e a dificuldade preditiva de modelos tradicionais de aprendizagem de máquina, tornou-se indicado a utilização de modelos híbridos como visto nos trabalhos de (Xu et al., 2019), (Xie; Lou, 2019) e (Nguyen et al., 2020).

Ao analisar a Tabela 2, destaca-se que o conjunto de dados empregado nos trabalhos relacionados é, em sua totalidade, menor do que aquele utilizado neste estudo. Essa discrepância apresenta vantagens e desafios. Por um lado, a dimensão reduzida pode facilitar a manipulação e o processamento dos dados. No entanto, por outro lado, constitui um desafio, pois se lida com informações menos discutidas, exigindo um entendimento aprofundado e um pré-processamento detalhado.

É importante ressaltar que muitos desses trabalhos utilizam apenas o nível de água como entrada, como discutido no primeiro capítulo e quando utilizam dados climatológicos são em sua totalidade no máximo três preditores. Contudo, como enfatizado anteriormente, o nível de água não é o único fator que influencia a previsão do atributo dependente. Essa limitação destaca a necessidade de considerar outros parâmetros relevantes para obter uma previsão mais abrangente e precisa. Portanto, a utilização de um conjunto de dados mais extenso neste trabalho busca abordar essa lacuna e contribuir para uma análise mais completa e representativa.

A utilização de um único modelo não é suficiente para predizer uma série temporal, principalmente no estudo de caso deste trabalho, que possui um *dataset* com informações diárias e de tamanho expressivo. Uma série temporal hidrológica, possui características lineares e não lineares e existem modelos suficientemente bons para informações lineares, como o ARIMA, entretanto, este não apresenta um bom desempenho preditivo em uma janela maior de previsão, necessitando assim de outros

modelos para aumentar o horizonte preditivo e capturar informações úteis em dados não lineares.

Nos modelos híbridos propostos por (Xie; Lou, 2019) e (Nguyen et al., 2020) apesar de apresentarem resultados interessantes utilizando modelos híbridos em conjunto não tão expressivos, estes utilizam-se somente informações de nível de água. Já em (Xu et al., 2019), a arquitetura desenvolvida envolvendo a separação dos dados não lineares e lineares bem como a utilização tanto do nível de água como dados climáticos é interessante, entretanto, o mesmo é treinado em um *dataset* reduzido e não apresenta resultados expressivos.

Na revisão da literatura que conduzimos, observamos que muitos autores empregam modelos de previsão baseados em séries temporais exclusivas como em (Xie; Lou, 2019),(Xu et al., 2019) entre outros. Esses modelos realizam previsões imediatamente após o término da série temporal, resultando em uma dependência crítica da constante atualização do conjunto de dados para manter a precisão das previsões.

Ao contrário dessa abordagem, nosso segundo modelo adota uma perspectiva diferente ao lidar com previsões. Neste caso, enfrentamos um problema de aprendizagem de máquina supervisionado, no qual os usuários inserem manualmente as informações climatológicas e de nível, eliminando a necessidade de depender de atualizações constantes do conjunto de dados. Essa autonomia oferece uma maior flexibilidade no uso do modelo, permitindo uma adaptação mais eficaz a diferentes condições e exigências. Neste cenário o processo de manipulação de um problema de séries temporais com aprendizagem de máquina é único e agrega vários benefícios preditivos.

É interessante notar que, enquanto a maioria dos estudos revisados estabelece uma pequena janela preditiva como um dia frente, encontrado no estudo de (Xie; Lou, 2019) e máxima de 30 dias, conforme exemplificado por (Xu et al., 2019), nosso modelo estende essa capacidade ao fornecer previsões para um horizonte mais amplo de 45 dias à frente. Essa extensão na janela preditiva pode se revelar valiosa em contextos específicos, oferecendo uma visão mais abrangente e antecipada das tendências e variações climáticas, contribuindo assim para uma tomada de decisão mais informada.

Em contraste com a abordagem comum que emprega apenas um conjunto de treinamento e teste para avaliar os dados, neste caso presente em todos os trabalhos analisados, nosso trabalho adota uma estratégia mais abrangente, empregando três conjuntos distintos: treinamento, validação e teste. Esta escolha metodológica foi feita com base na necessidade de garantir uma avaliação mais precisa e robusta do desempenho do modelo. O conjunto de treinamento é utilizado para ajustar os parâmetros do modelo, enquanto o conjunto de validação é empregado para ajustar hiperparâmetros e evitar o sobreajuste (*overfitting*). Por fim, o conjunto de teste é reservado para avaliar

o desempenho final do modelo em dados não vistos durante o treinamento. Ao incorporar esses três conjuntos distintos, nosso trabalho busca mitigar vieses e fornecer resultados mais confiáveis e generalizáveis, contribuindo assim para uma pesquisa mais sólida e confiável no campo da análise de dados.

O estudo de caso em questão, o qual é a previsão do nível de água no canal São Gonçalo envolvendo também dados climáticos, é único, tanto pelas características físicas do meio que está inserido, bem como pelo fato de ainda não existirem trabalhos publicados para a predição deste importante fator que é o nível de água, no qual pode beneficiar toda uma região que depende deste canal.

Nas seções a seguir, serão abordados aspectos relacionados a revisão da literatura sobre os artigos selecionados para o arquivamento.

3.1 Modelo ARIMA

Nesta seção serão relatados assuntos relacionados a modelos preditivos que utilizam somente o modelo ARIMA.

O modelo ARIMA é um dos modelos estatísticos lineares mais conhecidos e eficazes para previsão de séries temporais. Em (Birylo et al., 2018), os autores utilizaram o modelo ARIMA para prever o nível do lençol freático em três locais na região da Polônia, que requer três parâmetros: precipitação, vazão superficial e evapotranspiração, sendo a precipitação o fator que mais influi na previsão. Os resultados da previsão apontam que para doze meses este modelo obteve bons resultados, como parâmetros de configuração utilizaram-se os seguintes valores: p=2,d=0 e q=2. Além da previsão futura, um dos objetivos deste trabalho é responder perguntas como: o fator causador das mudanças no ciclo da água. Lembrado que nesse estudo temos previsões mensal e não diárias como em nossos modelos.

Da mesma forma, (Ghimire, 2017) desenvolveu um modelo baseado no ARIMA para prever as vazões em duas estações hidrológicas em um rio nos EUA. Para a avaliação deste modelo foram utilizadas as métricas *Root Mean Square Error* e *Root Mean Square Deviation*.

O critério de AIC é utilizado neste trabalho para se obter os melhores parâmetros para o ARIMA, logo o menor valor é considerado a melhor escolha para p e q. Um fator a se considerar neste artigo é que com dados diários e séries longas, a sazonalidade praticamente não existiu. O conjunto de dados utilizado foi de seis anos com os seguintes parâmetros de configuração: estação1: (1,1,2) e estação2: (2,1,1).

Na pesquisa de Yu et al.(Yu et al., 2017) foi utilizado o modelo ARIMA para prever o nível de água diário de três estações presentes no rio Yangtze. Observa-se que a precisão do modelo ARIMA diminui à medida que aumenta o horizonte de previsão, ou seja, apresenta bons resultados para as previsões de curto prazo, mas não para

previsões do nível de água a longo prazo.

O modelo é treinado utilizando dados históricos de quatro anos (2012-2015), o modelo é avaliado em um conjunto de teste do ano de 2016. Para determinar o melhor *p* e*q* foi utilizado *Bayesian information criterion* (BIC) onde o menor valor apresentando, logo a escolhida para *p*, *d* e *q* foram (2,1,1).

O desempenho do modelo é avaliado por RMSE, *mean absolute percentage error* MAPE e *percent of bias* PB. O desempenho do modelo ARIMA segundo o autor, é consistente para todas as três estações, e capaz de prever diariamente o nível de água com precisão. No entanto, o período de previsão tem grande influência na precisão do modelo, ou seja, para um dia o RMSE é 0.1202m (metros) para dois dias 0.2452m e três dias 0.3692m.

Os autores afirmam que grande erros aparecem quando ocorrem mudanças bruscas do nível de água diário. A previsão precisa de séries temporais hidrológicas ainda é uma tarefa desafiadora devido a suas complicadas características não lineares e não estacionárias. Considerando a não linearidade e a complexidade da maioria dos problemas de previsão de séries temporais e as desvantagens do ARIMA podem ser necessários a incorporação de outros modelos.

3.2 Outros modelos de Aprendizagem de Máquina

Com o passar dos anos, métodos estatísticos de aprendizado de máquina têm contribuído para o avanço dos sistemas de previsão que fornecem soluções com bom desempenho utilizando séries históricas de dados de nível de água.(Wang; Wang, 2020) comparou o desempenho de vários modelos estatísticos de aprendizado de máquina como: MLR *Regression Linear Multiple*, GP *Process Gaussian*, *Random Forest*, M5P e KNN *k-nearest neighbor*, para a previsão do nível de água do Rio Erie. Foram utilizados como informações de entrada dados o nível de água e variáveis ambientais como: temperatura do ar, radiação solar, velocidade do vento e umidade relativa. O conjunto de dados é do tipo diário e compreende o período de 2002 a 2014, sendo o intervalo de treinamento utilizado foi de 2002 a 2012 e de teste 2013 à 2014.

Segundo o autor (Wang; Wang, 2020), de janeiro a março de 2013, devido à baixa frequência de precipitação, os níveis de água são relativamente estáveis e tornam-se mais baixos. A partir de abril, o nível da água sobe com chuvas frequentes na região. De setembro a dezembro, com a redução da precipitação, o nível da água diminui novamente.

As predições dos modelos de *Machine Learning* também foram comparadas ao *hydrologic prediction system* (AHPS), este modelo pode ser usado para capturar padrões sazonais e interanuais de níveis de água, este modelo já foi utilizado no estudo de caso em questão. As medidas de avaliação utilizadas neste trabalho foram o *RMSE*

e *MAE*. Neste caso, o MLR e M5P obtiveram melhores resultados, ou seja, errou 0,02 e 0,01 metros para um dia a frente.

O Trabalho proposto em (Sun et al., 2017), mostrou a aplicação de uma Rede Neural do tipo *Back Propagation*, no lago *Zhari Namco*, que é um dos maiores lagos do *Tibet*, onde os valores de entrada foram: nível de água, precipitação e a temperatura em um conjunto dados de 5 anos, ou seja, compreendendo o período de 2007 a 2012. Foi verificado uma grande quantidades de informações duplicadas e nulas, onde os dados duplicados foram removidos e as informações nulas foram substituídas pela média dos dados adjacentes de 15 dias.

Seguindo a análise do trabalho de (Sun et al., 2017), para a normalização dos dados foi utilizado a função *Min-Max Scaling*, para aumentar a velocidade de convergência da rede. As funções de ativação utilizadas foram a *tansig* e *purelin*. O resultado deste trabalho, aponta que este tipo de Rede Neural apresenta menor erro preditivo em relação aos demais, ou seja, a rede proposta pelo autor teve um *RMSE* de 0,2983 e a rede *Back Propagation* original teve um *RMSE* de 0,3264 para um horizonte preditivo de uma dia a frente.

Em (Zhang et al., 2020), foi proposto um modelo baseado em RNN, para a previsão do nível de água no rio *Ningbo*, na China, foram selecionados cinco reservatórios e os dados faltantes foram preenchidos com a média do mês anterior. Para o valor previsto de cada reservatório, os dados de cada cinco meses adjacentes são usados como tamanho da janela deslizante para prever o nível de água do próximo mês.

Seguindo a análise do trabalho (Zhang et al., 2020), para se obter um conjunto otimizado de *Learning Rate* e *hidden sizes* , diferentes taxas foram testadas no processo de aprendizado de máquina. Para o *Learning Rate*, três casos são aplicados, os quais são 0,001, 0,0005 e 0,0001.

A quantidade de dados para treinamento foi de 70% e de teste de 30%. Para a previsão do próximo mês de nível de água foi utilizado os cinco meses anteriores e os dados faltantes foram preenchidos pela média de todos os dados. Como resultado, o modelo proposto apresenta melhor desempenho preditivo em relação a ANN e LSTM, em outras palavras, possuindo um erro absoluto de 1,19 e os demais de 1,29 e 1,37 respectivamente, o artigo não menciona a escala utilizada nesses resultados.

Na pesquisa de (Hrnjica; Bonacci, 2019), foi aplicado dois modelos de Redes Neurais para a previsão do nível de água no lago *Vrana* na Croácia, sendo que o primeiro modelo aplicado foi o *Feed and Forward* e o segundo o *RNN*.

Os dados utilizados como entrada foram somente o nível de água e a série temporal inicia-se em 1978 e termina em 2016. Algumas características sobre a configuração dos modelos foram: elemento ativador Adam, três camadas sendo a quantidade de nerônios testados foram: 6,9,12,15 e 18, número de épocas 5000 learning rate 0,1. Segundo (Hrnjica; Bonacci, 2019), quando o nível de água tende a valores críticos de

redução a predição é comprometida, uma das explicações são as mudanças climáticas e a fatores humanos. Para testar os modelos treinados, estes foram comparados com modelos clássicos de séries temporais como: ARIMA e PROFET, sendo a arquitetura proposta pelo autor apresentou o menor RMSE. Neste estudo o autor não comenta a escalda dos dados se é em metros ou centímetros.

Em (Choi et al., 2019), é implementado os modelos de *Decision Tree*, *Random Forest*, *Artificial Neural Network* e *Suport Vector Machine* para a previsão do nível de água em pântanos alagados. Neste caso o intervalo foi de 2009 a 2015 e foram usados como variáveis dependentes o nível de água e dados climatológicos. O modelo *Random Forest* obteve menor taxa de erros, mesmo assim seu desempenho deixou a desejar em pontos de pico, ou seja, e fases chuvosas nos meses de julho, agosto e outubro.

(Choi et al., 2019) relata que no conjunto de dados em questão, o nível de água fica em torno de dois a três metros e durante a estação chuvosa (junho a agosto), houve uma mudança repentina no nível da água devido as fortes chuvas. Como variáveis independentes, a temperatura média diária, temperatura mínima diária, temperatura máxima, precipitação diária, velocidade diária do vento e o nível de água foram utilizados, sendo estas as mesmas variáveis utilizadas neste trabalho desenvolvido.

A temperatura (média, mínima, máxima) mostrou-se moderadamente relacionada com o nível da água. A precipitação e a velocidade do vento (máxima instantânea, média) exibiram uma fraca relação, e o nível de água a montante mostrou uma forte relação. Os dados de um dia atrás teve uma maior relação que os dados de dois ou três dias atrás, no caso da precipitação, a informação mútua foi fraca.

Na RNN, como parâmetros de configuração foi utilizado até dez neurônios e testados até cem épocas para cada neurônio. No *Decision Tree*, as árvores dividem repetidamente os dados, encontrando pontos de ramificação que minimiza o erro quadrático médio. À medida que o processo de divisão dos dados é repetido, a probabilidade de *overfiting* aumenta, então a validação cruzada deve ser usada para determinar o tamanho da árvore que minimiza o erro, o número de *folds* foram de dez.

As previsões no trabalho de (Choi et al., 2019) são realizadas para um horizonte preditivo de um a três dias a frente. O artigo menciona que para uma previsão de três dias a frente o RMSE foi de 0,14 e o coeficiente de correlação 0,94.

No trabalho de (Castillo; Cspedes; Cuchango, 2018), é proposto a utilização de Redes Neurais Auto Regressivas, Regressão de Vetores de Suporte, *Multi-Layer Percipton*, *Extreme Learning Machine* e *Gaussian Process Regression*, para prever o nível de água em uma determinada região da Colômbia. O modelo proposto pode servir para prever enchentes que podem afetar atividades produtivas como agronomia e pecuária; bem como poderia ser usado para evitar danos na indústria energética e infraestrutura de transporte.

O autor (Castillo; Cspedes; Cuchango, 2018) traz alguns conceitos sobre informações hidrológicas como, por exemplo, o monitoramento das medições de chuva e do nível dos rios e barrancos, via pluviômetros e hidrometria que, permitem conhecer a situação hidrológica que podem desencadear em uma inundação. Os medidores de chuva fornecem informações do volume de água no solo (vindo da chuva), enquanto as escalas hidrométricas fornecem informações sobre o aumento do nível de água nos corpos d'água.

As variáveis utilizadas neste modelo incluem: dados históricos de nível de água em diferentes estações em um intervalo de 1997 a 2015 como dados de entrada. As variáveis utilizadas neste modelo incluem: dados históricos de nível de água em diferentes estações. Os resultados mostraram um bom desempenho do sistema para dois, três e cinco dias utilizando três, seis e 12 neurônios respectivamente. Neste estudo, o horizonte preditivo utilizado foi o de uma semana apenas, onde o modelo *Multi-Layer Percepton* obteve o menor RMSE 0,0780 metros.

3.3 Modelos Híbridos

Em muitos dos casos a utilização de um único modelo preditivo não é o suficiente para a obtenção de bons resultados, tendo em vista as características lineares e não lineares de uma série temporal hidrológica. Logo, alguns modelos híbridos foram propostos, como em (Xu et al., 2019), que utilizou um modelo híbrido combinando ARIMA com RNN para a previsão do nível de água.

Neste modelo, os dados lineares, são previstos pelo modelo ARIMA e os componentes não lineares são preditos pela RNN. O modelo usa os dados de nível de água diários e fatores ambientais relacionados, como vetores de entrada, logo, obtém-se como vetor de saída o nível de água nos próximos 30 dias.

Segundo(Xu et al., 2019), devido à influência de vários fatores externos, séries temporais hidrológicas contém componentes tanto lineares como não lineares e um único modelo de previsão é muitas vezes incapaz de prever com precisão o nível da água.

O modelo proposto é dividido em quatro etapas, sendo a primeira a divisão do *dataset* hidrológico em outros dois, um somente com o nível de água e outro com os dados climáticos/ambientais. No segundo passo, o *dataset* linear é submetido ao ARIMA, onde se verifica a estacionalidade, e aplicam-se as diferenças sucessivas caso necessário. Já no terceiro passo, os dados climáticos, são normalizados e juntamente com a parte não linear do modelo ARIMA são submetidas a RNN, então é predito a sequência residual, como último passo, a predição do primeiro e segundo passo são utilizados para a previsão dos próximos trinta dias. Os resultados experimentais indicaram que o modelo combinando pode capturar melhor a tendência geral

e a flutuação de amplitude.

Outra técnica híbrida foi proposta em (Xie; Lou, 2019), onde se observa a conjunção do ARIMA com SVR e a utilização da transformada de wavelet para a decomposição dos aspectos lineares e não lineares da série para a previsão do nível de água. Este modelo diferente dos demais por fazer a decomposição desses aspectos antes do treinamento dos dados.

Na base de decomposição wavelet, o modelo combina as vantagens do modelo ARIMA em lidar com séries temporais lineares, como simplicidade, princípio científico do SVR para lidar com baixo erro de generalização de séries temporais não lineares.

Este modelo híbrido prediz as variáveis lineares e não lineares da série temporal hidrológica separadamente após a decomposição de wavelet, e então reconstrói os resultados previstos pela transformada. Os dados de nível de água são diários em um estacão da Bacia de *Nanjing* na China e como saída obtém-se a predição para o dia seguinte. O conjunto de dados utilizado inicia-se em 2010 e termina em 2015.

Para (Xie; Lou, 2019), é difícil para séries temporais hidrológicas dependerem de um único modelo não linear para prever o nível da água, e é difícil para um único modelo descrever a complexidade de séries temporais hidrológicas e prever com precisão. O modelo proposto é dividido em passos, primeiramente o conjunto de dados é normalizado, como segundo passo é realizado a decomposição em tendência e flutuação, no terceiro passo, o modelo ARIMA é utilizado tendo com entrada a parte linear do *dataset*, como quarto passo, aplica-se o método *GridSearch* para obter os melhores parâmetros de configuração, por fim, é realizado a reconstrução da transformada.

O efeito de regressão do modelo SVR é amplamente influenciado pela escolha da função de *kernel* e dos parâmetros do modelo. O núcleo RBF função usada neste artigo, tem forte capacidade de generalização e menos parâmetros controlados. O modelo SVR usando *kernel* RBF é afetada principalmente pelo parâmetro de penalidade C, e a função de perda o parâmetro p. No modelo ARIMA, o mesmo foi configurando com os seguintes valores de p,q e d (1,2,1). O modelo proposto, é comparado com o SVR e ARIMA, onde o modelo ARIMA apresentou um RMSE de 0,3943, SVR de 0,3063 e o modelo ARIMA-SVR de 0,2666 para uma janela preditiva de um dia a frente.

No trabalho de (Nguyen et al., 2020), é proposto a utilização da combinação do modelo *ARIMA* com vários modelos de *Machine Learning* como: *Random Forest*, KNN, SVR, RNN para a previsão do nível de água em alguns conjuntos de dados, sendo o tamanho médio de cada conjunto é de nove anos. A hipótese para com a utilização de modelos híbridos é que, usando dois tipos de modelos para capturar as partes não lineares em séries temporais de nível de água, padrões ocultos nos dados de séries temporais poderiam ser melhor revelados em comparação com o tipo único de abordagem do modelo.

Segundo o autor (Nguyen et al., 2020) não existe um modelo universal que pode capturar com sucesso os relacionamentos lineares e não lineares. Modelos estatísticos lineares, como o ARIMA, podem não ser bom em modelar as relações não lineares na série temporal, mas é suficiente para modelagem de componentes lineares. Enquanto isso modelos estatísticos de aprendizado de máquina (não paramétricos), como SVM, RF, KNN ou LSTM podem modelar quaisquer componentes não lineares (aproximadores universais).

Os modelos, ARIMA, PARMA e GAMRMA são os candidatos mais relevantes para modelar o componente linear dos dados de séries temporais, nos quais ARIMA foi creditado, por ser o mais adequado dos três para lidar com séries temporais não estacionárias. Para capturar o componente não linear da série temporal, RF, SVM, KNN e LSTM são os modelos utilizados. Foram selecionados por vários motivos. Primeiro, todos eles são métodos estatísticos de aprendizagem (não paramétricos) com fundamentos teóricos relativamente firmes (por exemplo, com garantia teórica de aprendizagem, consistência de aprendizagem e universalidade). Em segundo lugar, são representantes de classes grandes e populares de técnicas estatísticas de aprendizado de máquina. Em terceiro, eles foram aplicados com sucesso na modelagem/aprendizagem de séries temporais hidrológicas.

Continuando a análise de (Nguyen et al., 2020), alguns parâmetros dos modelos foram utilizados, entre eles destacam-se no modelo ARIMA, os melhores p,d e q foram (5,1,1) para a estação 1, (5,1,5) na estação 2 e (5,1,3) para a estação 3. No SVM, o *kernel* utilizado foi o RBF, com o valor de C=1,50 e 50 para as três estações. No *Random Forest*, o número de árvores utilizados foram 100,1000 e 3000, respectivamente.

No *LSTM*, como principais parâmetros utilizados foram *Batch Size* com o valor de 3, o otimizador selecionado foi o *Adam* e o número de épocas utilizadas para o treinamento foi de 200. Como forma de validação do modelo foi utilizado*5-fold cross-validation*. Já no modelo ARIMA, foi aplicado o auto-arima, para encontrar os melhores p,d e q. Foi considerado o componente de sazonalidade para todos os conjuntos de dados ao treinar os modelos.

Os resultados mostram um melhor desempenho para os modelos combinados com: ARIMA-RF e ARIMA-KNN. Apesar deste artigo apresentar bons modelos preditivos, apenas consideram o nível de água para a previsão, o que pode ser uma limitação. No caso do ARIMA-KNN para uma previsão máxima de cinco dias a frente temos um RMSE de 0,3668.

Nos trabalhos selecionados, observa-se uma diversidade nas escalas de expressão dos resultados, sendo comum encontrar dados de nível apresentados em unidades como centímetros ou hectômetros cúbicos. Diante dessa variedade de unidades de medida, uma adaptação necessária para possibilitar comparações significativas é a conversão do *RMSE* para a escala adotada neste trabalho, expressa em metros.

Embora a comparação direta possa não ser perfeita devido às particularidades inerentes a cada conjunto de dados, essa abordagem oferece uma noção interessante da adequação dos modelos desenvolvidos. A uniformização para a mesma escala permite avaliar o desempenho relativo dos modelos, facilitando a interpretação e a compreensão de sua eficácia em diferentes contextos e condições.

Após a revisão dos trabalhos relacionados, onde foi explorado diversas abordagens e estratégias utilizadas na previsão do nível de água em meios semelhantes ao Canal São Gonçalo, é crucial agora avançar para a implementação e metodologia do trabalho propriamente dito. Neste próximo capítulo, será detalhado como foi aplicado as percepções adquiridas durante a revisão da literatura para desenvolver uma abordagem robusta e eficaz para prever o nível de água neste cenário específico. Desde a coleta e preparação dos dados até a seleção e implementação dos modelos preditivos, esta seção fornecerá uma visão detalhada de todas as etapas envolvidas no processo de desenvolvimento e execução do trabalho. Ao conectarmos a revisão dos trabalhos relacionados com a implementação prática deste estudo, será estabelecido uma base para compreender como as pesquisas anteriores influenciaram e informaram nosso próprio trabalho de previsão hidrológica no Canal São Gonçalo.

4 METODOLOGIA DE PREVISÃO HIDROLÓGICA NO CANAL SÃO GONÇALO

Neste capítulo, após ter estabelecido todo o referencial teórico para a previsão do nível de água no Canal São Gonçalo, iremos adentrar ao cerne do trabalho, onde será apresentado a metodologia adotada, o conjunto de dados utilizados, os modelos propostos e os processos de otimização empregados. Esta seção busca fornecer uma visão detalhada e abrangente das etapas práticas envolvidas no desenvolvimento e na implementação dos modelos de previsão hidrológica específicos para o Canal São Gonçalo. Ao delinear nossa abordagem metodológica, será descrito os passos para coleta, preparação e análise dos dados, bem como a fundamentação teórica por trás dos modelos selecionados. Além disso, será discutido os métodos de otimização utilizados para refinar e ajustar os modelos, visando obter previsões mais precisas e confiáveis do nível de água neste contexto particular.

4.1 Conjunto de Dados

Nesta seção serão detalhados os conjuntos de dados utilizados nos experimentos. Esta pesquisa, é baseada em quatro conjunto de dados, onde os parâmetros utilizados são: Dados de Nível, contendo o Nível Jusante Eclusa, Nível Montante Eclusa este dois compreendem a parte da barragem São Gonçalo, existem também os níveis dos centros de medições de Santa Vitória do Palmar e Santa Isabel.

Os conjuntos de dados originais, possuíam escalas diferentes, por exemplo, em determinados conjuntos parte dos dados de níveis monitorados eram expressos em centímetros e a outra parte em metros, como convenção utilizou-se a escala de metros, pelo fato de que a grande maioria dos dados eram nesta convenção. Outra dificuldade encontrada foi que em determinados centros de monitoramento e em partes menores do conjunto de dados, as medições eram realizadas a cada hora, neste caso foi desenvolvido um algoritmo para pegar estes dados expressos em hora, realizar a média e deixá-los na convenção diária, o qual é a utilizada na maior parte dos conjuntos de dados.

Os dados climáticos utilizados nos conjuntos de dados foram: evaporação, insolação total, precipitação total, temperatura mínima, temperatura média e temperatura máxima, umidade relativa do ar, velocidade média do vento e direção do vento. Nesses dados climáticos, somente a direção do vento não foi utilizada da base de dados do Inmet, neste caso foram utilizados os dados fornecidos pelos centros de monitoramento de Santa Vitória do Palmar e Santa Isabel. Cabe ressaltar que toda a elaboração dos conjuntos de dados foram acompanhados por Engenheiros Hídricos da UFPEL.

Segundo detalhamento fornecido pelos Engenheiros Hídricos nosso atributo previsor é o dado de Nível Montante Eclusa, logo essa informação será o atributo dependente e os demais serão atributos independentes nos experimentos a serem realizados.

Foram realizados experimentos com um conjunto de dados inicial considerando todos os dados climáticos mencionados acima, entretanto, somente com um atributo de nível que é a informação de Montante Eclusa. Neste primeiro experimento, conforme (Lima; Marques; Orth, 2023), observa-se o estudo pré-eliminar envolvendo todos os modelos utilizados neste experimento, entretanto, considerando um conjunto de dados menor, ou seja, tendo início em 29/01/1979 e findando em 31/05/2018. De acordo com uma segunda análise do conjunto de dados realizada pelos Engenheiros Hídricos, observou-se que considerar somente um dado de nível não representaria o cenário preditivo em sua totalidade, sendo necessário considerar outras partes da Lagoa.

4.1.1 Considerações sobre os Conjunto de Dados

As informações climatológicas utilizadas, são expressas em determinadas escalas, sendo a evaporação de piche é expressa em mililitros e é a quantidade de água evaporada a partir de uma superfície porosa, mantida permanentemente umedecida por água. Outro dado utilizado é a insolação total em horas, que se refere ao número total de horas em que um determinado local ou região recebeu luz solar durante um período específico. Essa medida é frequentemente usada em meteorologia e climatologia para descrever a quantidade de sol que atinge uma área ao longo de um dia, mês ou ano.

Continuando a análise das escalas temos, a precipitação total, sendo a sua escala expressa em mm, refere-se à quantidade total de água, na forma de chuva, neve, granizo, orvalho ou qualquer outra forma de água que atinja a superfície da Terra durante um período específico de tempo. Essa medida é usada para quantificar a quantidade total de precipitação que ocorreu em uma determinada localidade, seja ao longo de um dia, mês, ano ou qualquer outro intervalo de tempo.

Os registros de temperatura média, máxima e mínima são medidas meteorológicas que descrevem as condições térmicas em um determinado local durante um período

específico, geralmente em um dia, mês ou ano. As temperaturas são expressas em graus Celsius (°C) ou Fahrenheit (°F), dependendo do sistema de unidades utilizado. Neste caso utilizamos graus Celsius (°C).

A umidade relativa do ar é uma medida que expressa a quantidade de vapor d'água presente no ar em relação à quantidade máxima que o ar poderia conter a uma determinada temperatura. A umidade relativa do ar é expressa como uma porcentagem e fornece informações sobre o nível de saturação do ar com vapor d'água. Velocidade do vento é comumente expressa em metros por segundo (m/s). Essa unidade de medida é a mais utilizada no sistema internacional de unidades (SI) para descrever a velocidade do vento e é amplamente adotada em contextos meteorológicos e científicos.

A direção do vento é geralmente expressa em graus, indicando a direção de onde o vento está vindo. A convenção mais comum é medir a direção no sentido horário em relação ao norte, de modo que 0 graus represente o norte, 90 graus represente o leste, 180 graus represente o sul e 270 graus represente o oeste. Os dados de nível de água são expressos em metros e indicam o nível vertical da água em relação a um ponto de referência específico. Essa medida é fundamental para monitorar corpos d'água, como rios, lagos, mares e oceanos, e é usada em diversas aplicações, incluindo previsão de inundações, gestão de recursos hídricos, monitoramento ambiental e estudos hidrológicos.

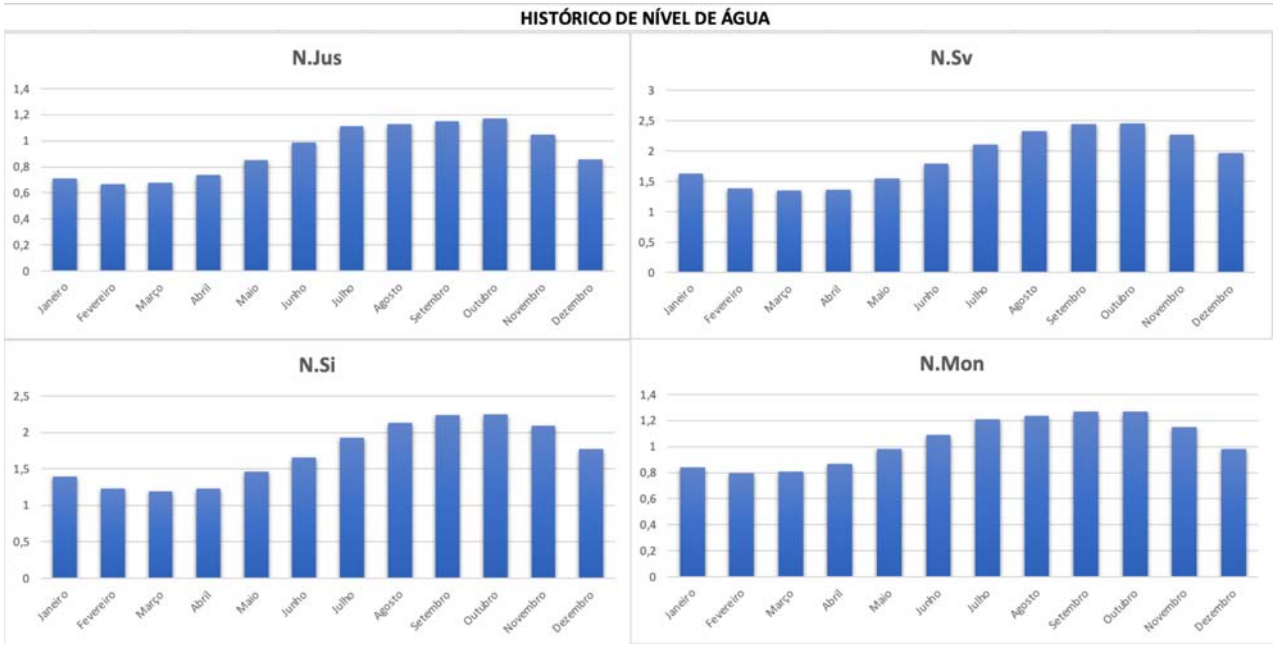

Figura 15 – Dados de Nível de Água - 29/01/1979 à 31/05/2018

Entender o nível de água em um conjunto histórico de dados climatológicos é de suma importância para diversos campos, desde a gestão hídrica até a compreensão das dinâmicas ambientais. O nível de água em rios, lagos ou outras fontes hídricas

são um indicador crítico que reflete a interação complexa entre fatores climáticos, hidrológicos e ambientais. Analisar historicamente esses dados fornece percepções valiosas para tomada de decisões, previsão de eventos extremos, gestão sustentável dos recursos hídricos e compreensão das respostas dos ecossistemas aquáticos às mudanças ambientais.

No contexto específico, conforme a Figura 15, temos informações cruciais sobre os níveis de água de diferentes estações de monitoramento. A inclusão dos dados de Santa Vitória, Santa Isabel e São Gonçalo, Montante e Jusante adiciona camadas significativas de detalhes à análise. Essas estações oferecem perspectivas abrangentes sobre as variações nos níveis de água ao longo do tempo, considerando diferentes locais geográficos e posições ao longo de um curso d'água.

A análise dos dados médios de nível de água para cada mês do ano revela padrões sazonais que oferecem entendimentos sobre o comportamento hidrológico ao longo do tempo. O gráfico proporciona uma visão panorâmica das variações nos níveis de água em diferentes estações de monitoramento, permitindo uma compreensão mais profunda das dinâmicas sazonais que caracterizam a região.

Numa análise preliminar, destaca-se um comportamento sazonal distintivo, no qual ocorre um aumento nos níveis de água em todas as estações a partir do mês de junho. Este aumento persiste por aproximadamente cinco meses, indicando um padrão recorrente de variação nos níveis hídricos ao longo do ano. Essa tendência sazonal pode ser atribuída a uma variedade de fatores climatológicos, como mudanças nas precipitações e outros fatores que discutiremos nas próximas seções.

A identificação desse padrão sazonal é crucial para a gestão e o planejamento de recursos hídricos. Compreender quando ocorrem os aumentos e quedas nos níveis de água possibilita uma melhor preparação para eventos extremos, como inundações sazonais ou períodos de seca prolongada. Além disso, essa análise sazonal pode ser fundamental para setores que dependem diretamente das condições hídricas, como agricultura, geração de energia hidrelétrica e gestão de ecossistemas aquáticos.

Além do comportamento sazonal destacado, é igualmente crucial observar que, nos períodos compreendidos entre dezembro e abril, ocorre uma considerável diminuição nos níveis de água. Essa fase de redução, que abrange os meses de final de ano até o início do outono, representa uma parte significativa do ciclo anual e evidência a complexidade das flutuações hidrológicas ao longo do tempo.

A diminuição nos níveis de água durante esses meses pode estar associada a diversos fatores climatológicos. Possíveis explicações incluem uma diminuição nas chuvas, a transição para períodos de seca ou fenômenos climáticos específicos que impactam negativamente os recursos hídricos. Compreender essa diminuição sazonal é essencial para a gestão sustentável dos recursos hídricos, possibilitando estratégias adaptativas e medidas de conservação durante esses períodos críticos.

Essa análise mais detalhada do comportamento sazonal dos níveis de água, destacando não apenas os aumentos, mas também as diminuições, fornece uma base mais completa para a tomada de decisões informadas. O entendimento dessas variações sazonais é fundamental para antecipar eventos climáticos extremos, otimizar o uso dos recursos hídricos e promover práticas de gestão que assegurem a resiliência dos ecossistemas aquáticos e o fornecimento sustentável de água ao longo do ano.

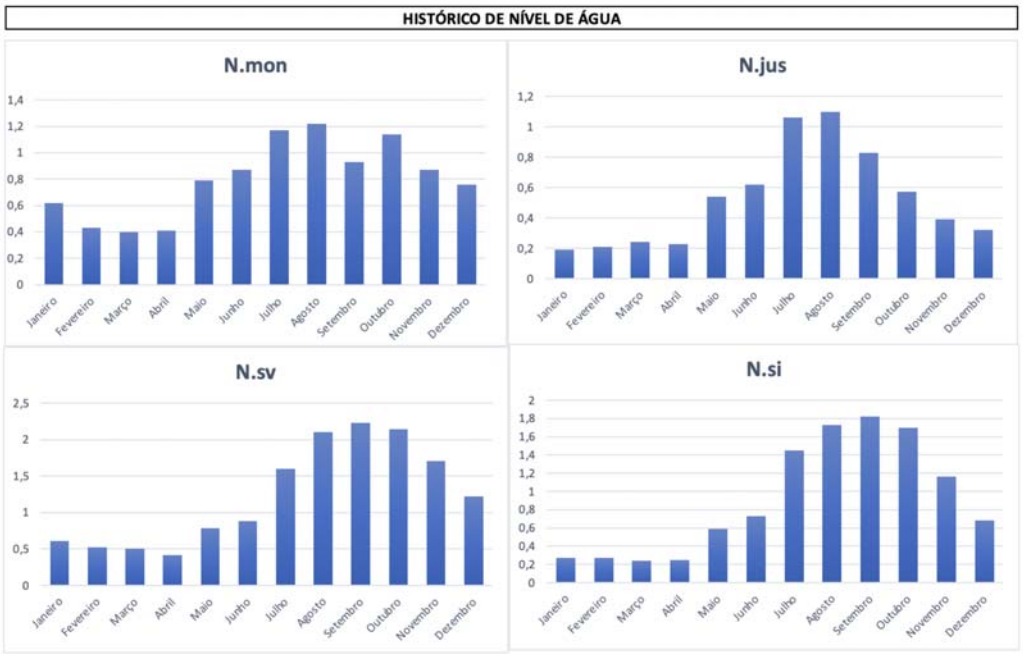

Figura 16 – Dados de Nível de Água - 05/01/2021 à 30/06/2023

A análise dos dados mais recentes, conforme a Figura 16, apesar da menor quantidade de informações, revela padrões significativos nos níveis de água das estações de Santa Isabel e Santa Vitória. Uma observação inicial aponta para níveis de água consideravelmente baixos, especialmente nos meses de janeiro a abril, indicando um fenômeno sazonal que merece atenção especial.

A concentração de níveis de água mais baixos durante esse período pode ser associada a diversas variáveis climáticas ou hidrológicas. Possíveis influências incluem a redução nas precipitações, características específicas da bacia hidrográfica ou eventos climáticos extremos que impactam negativamente o abastecimento hídrico dessas estações.

A identificação desses períodos críticos é essencial para a gestão adequada dos recursos hídricos, uma vez que pode sinalizar vulnerabilidades sazonais e informar a implementação de medidas preventivas. A compreensão detalhada dessas variações, mesmo com uma quantidade reduzida de dados, destaca a importância da análise contínua e atualizada para promover a resiliência e a sustentabilidade em face das flutuações naturais do ambiente hídrico.

Essa análise inicial fornece uma base sólida para investigações mais aprofunda-

das e destaca a necessidade de monitoramento contínuo, especialmente em períodos críticos, para desenvolver estratégias adaptativas e garantir uma gestão eficaz dos recursos hídricos nessas regiões específicas.

Durante a análise dos dados hidrológicos referentes aos níveis de água nas regiões Jusante e Montante, observamos padrões distintos ao longo dos meses. Consideravelmente, constatamos um aumento consistente nos níveis de água durante os meses de maio a setembro, seguido de uma diminuição mais acentuada em Jusante que se inicia em outubro e se estende até abril.

O aumento nos níveis de água nos meses de maio a setembro pode ser atribuído a uma série de fatores sazonais. Durante esse período, muitas regiões experimentam condições climáticas que contribuem para o aumento do fluxo de água, como chuvas sazonais mais intensas ou eventos meteorológicos específicos associados a esses meses. Esse padrão de aumento sazonal é observado tanto em Jusante quanto em Montante, sugerindo uma influência climática global nesse fenômeno.

No entanto, a análise revela uma dinâmica peculiar nos meses subsequentes. A diminuição mais acentuada nos níveis de água em Jusante, iniciando em outubro e persistindo até abril, aponta para um comportamento hidrológico diferenciado entre as duas regiões. Essa diminuição pode ser resultado de uma combinação de fatores, incluindo uma redução nas chuvas, evaporação mais intensa ou mudanças nos padrões de escoamento na área Jusante.

A marcante sazonalidade nos períodos de aumento e diminuição dos níveis de água reforça a influência direta das variações climáticas na dinâmica hidrológica da região. A sazonalidade indica uma resposta sensível do sistema hidrológico aos ciclos naturais, evidenciando a interconexão entre os fatores meteorológicos e os níveis de água.

Essa análise mais abrangente, considerando múltiplas estações, enriquece a compreensão dos padrões hidrológicos na região, destacando a importância de uma abordagem integrada para o monitoramento e gestão de recursos hídricos em diferentes localidades ao longo do curso d'água.

4.1.1.1 Conjunto de Dados I

Conforme os detalhes fornecidos pelos Engenheiros Hídricos, nosso atributo preditivo são os dados do nível de água da parte a montante. Esta informação será o atributo dependente, e as demais serão atributos independentes nos experimentos a serem realizados. A diferença desse estudo em relação ao(Lima; Marques; Orth, 2023) é o fato deste considerar, três dados de nível.

Neste conjunto de dados, utilizou-se todos os dados climatológicos e todos dados de nível, iniciando-se em 29/01/1979 e findando em 31/05/2018 totalizando 14369 registros. Todavia, este conjunto possui algumas lacunas contendo informações sem

registros, conforme a Figura 17, sendo que estes foram preenchidos pela média de tendência central.

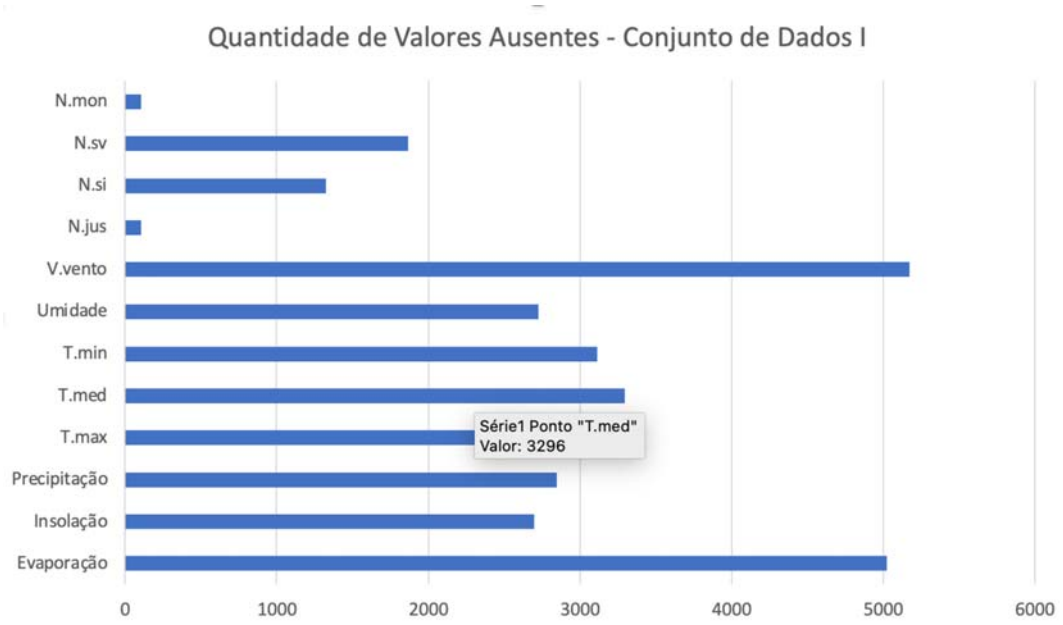

Figura 17 – Somatório Total de Dados Faltantes

No processo de substituição de valores ausentes em um conjunto de dados climatológicos, foi realizado um teste empregando a técnica da média móvel de 90 dias representado uma estação do ano, como estratégia para substituir os dados faltantes. A intenção era avaliar se essa abordagem, que leva em consideração padrões de curto prazo, proporcionaria melhorias significativas em relação à simples utilização da média geral.

Entretanto, os resultados obtidos revelaram que a média móvel de 90 dias não apresentou melhorias substanciais em comparação com a média geral. A análise comparativa, que incluiu estatísticas descritivas e visualizações dos dados substituídos, indicou que ambas as abordagens proporcionaram resultados semelhantes.

É importante ressaltar que a eficácia de uma técnica de substituição pode depender fortemente das características específicas do conjunto de dados e dos padrões climatológicos presentes na região em questão. Em alguns casos, a simplicidade da média geral pode ser suficiente para uma representação adequada do comportamento médio ao longo do tempo.

Na Figura 17, verifica-se que os dados climatológicos de velocidade do vento e evaporação são as informações que possuem a maior quantidade de dados ausentes. Seguindo a análise, os dados de nível, apresentam uma menor quantidade de valores faltantes, lembrando que estas informações são retiradas da base fornecida pela Agência da Lagoa Mirim.

No processo de análise de conjuntos de dados, especialmente em contextos como dados climatológicos, não apenas a substituição de valores faltantes, mas também o

tratamento de *outliers* desempenha um papel crucial na obtenção de resultados mais precisos e confiáveis. *Outliers* são pontos de dados que se desviam significativamente do padrão geral dos dados e podem distorcer a interpretação e a eficácia de modelos estatísticos.

Ao lidar com dados climatológicos, a presença de *outliers* pode ocorrer devido a eventos extremos, erros de medição ou outras anomalias. Ignorar ou não abordar esses *outliers* pode impactar negativamente a qualidade da análise, prejudicando a precisão das conclusões e a confiabilidade dos modelos.

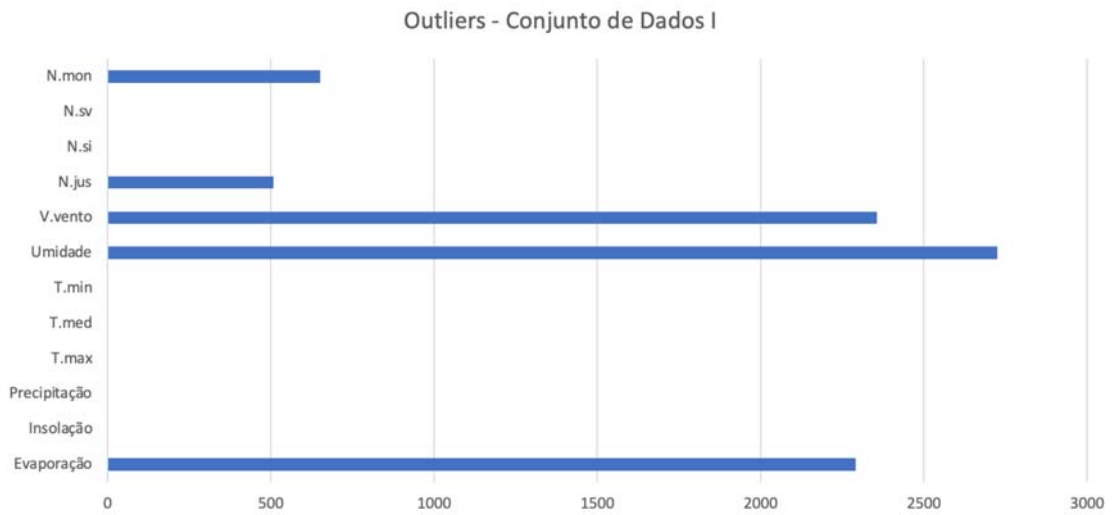

Figura 18 – Outliers do Conjunto de Dados I

No processo de análise dos dados, adotou-se a técnica do Intervalo Interquartil (IQR) para lidar com a presença de *outliers* no conjunto de dados. O IQR é uma medida estatística robusta que quantifica a dispersão dos dados em torno da mediana, oferecendo uma abordagem eficaz para identificação e tratamento de valores extremos.

Primeiramente, calculou-se o IQR, obtendo a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Em seguida, estabeleceram-se limites para identificar *outliers*: um limite inferior, calculado subtraindo 1,5 vezes o IQR de Q1, e um limite superior, obtido somando 1,5 vezes o IQR a Q3. Pontos de dados que ultrapassaram esses limites foram considerados *outliers*.

Para tratar esses *outliers*, optou-se por uma abordagem que envolveu tanto o ajuste quanto a transformação dos valores extremos. Em vez de remover diretamente os *outliers* do conjunto de dados, foram realizadas adaptações, substituindo os valores extremos por valores mais moderados próximos aos limites estabelecidos pelo IQR.

A transformação aplicada visou reduzir a influência dos *outliers*, garantindo uma representação mais fiel da distribuição dos dados. Essa estratégia permite mitigar os efeitos adversos dos valores extremos, preservando, ao mesmo tempo, a integridade e a qualidade do conjunto de dados.

Ao realizar uma análise aprofundada dos dados climatológicos, observou-se que algumas variáveis específicas apresentaram uma quantidade significativa de *outliers*, indicando uma dispersão mais ampla em relação à tendência central. As variáveis que se destacaram nesse aspecto foram a umidade, a velocidade do vento e a evaporação.

No contexto da umidade, a presença de *outliers* sugere que ocorreram eventos climatológicos excepcionais que resultaram em variações significativas nos níveis de umidade. Essas discrepâncias podem ter sido influenciadas por fatores como chuvas intensas ou períodos de seca prolongada, impactando diretamente a umidade do ar registrada.

Da mesma forma, a variável relacionada à velocidade do vento apresentou uma quantidade considerável de *outliers*, indicando momentos em que as condições climáticas foram marcadas por ventos atípicos. Esses eventos podem ser associados a tempestades, frentes atmosféricas ou fenômenos meteorológicos específicos que provocaram variações substanciais na velocidade do vento.

Quanto à evaporação, a detecção de *outliers* sugere episódios em que ocorreram taxas extremas de evaporação, muitas vezes vinculadas a condições climáticas excepcionais, como temperaturas extremamente altas ou baixa umidade relativa do ar.

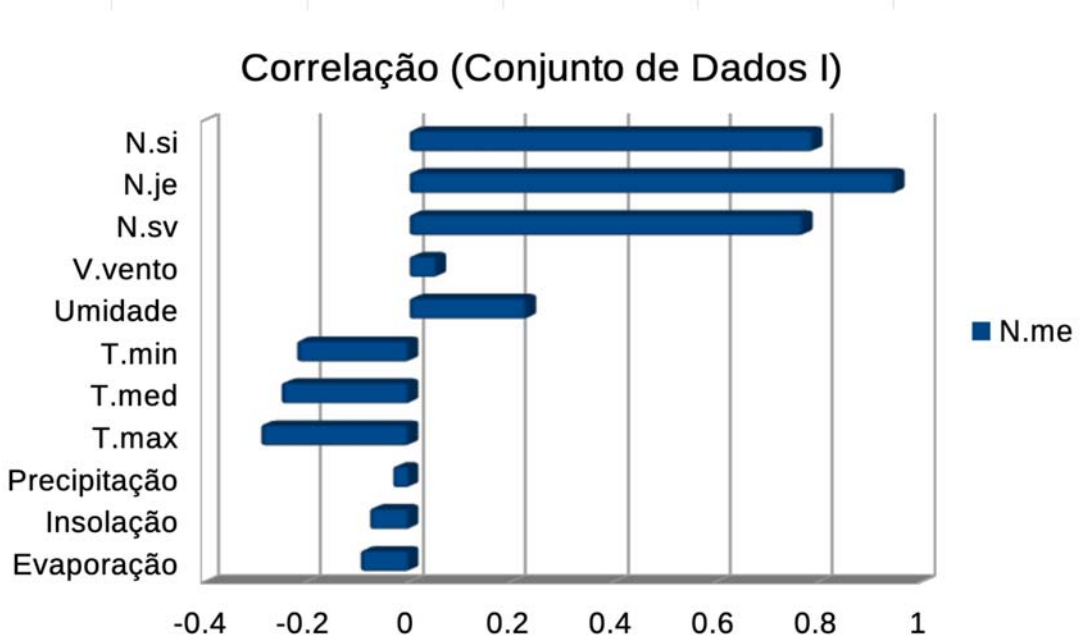

Figura 19 – Correleção do Conjunto de Dados I

Na Figura 19, observa-se a correlação de *Sperman* aplicada no Conjunto de Dados I, neste caso se optou de visualizar as informações diretamente relacionadas ao atributo dependente. No âmbito da análise exploratória deste conjunto de dados, a correlação de *Spearman* foi empregada como uma ferramenta robusta para avaliar as relações entre as diferentes variáveis. Este método é particularmente útil para identificar correlações monotônicas, ou seja, relações que podem ser crescentes ou decrescentes, mas não necessariamente lineares.

A correlação de *Spearman* é uma medida estatística que avalia a força e a direção da relação monotônica entre duas variáveis. Em contraste com a correlação de *Pearson*, que mede a relação linear entre variáveis, a correlação de *Spearman* não assume uma relação linear específica. Em vez disso, ela avalia se há uma relação monotônica geral, o que significa que, à medida que uma variável aumenta, a outra também aumenta ou diminui, mas não necessariamente a uma taxa constante.

A principal vantagem da correlação de *Spearman* é a sua robustez em relação a *outliers* e a capacidade de lidar com dados ordinais e não lineares. Essa medida é especialmente útil quando os dados não seguem uma distribuição normal ou quando a relação entre as variáveis não é linear.

Ao analisar as correlações resultantes, observou-se que entre os atributos relacionados ao nível da água, o dado de nível jusante da eclusa apresentou a mais expressiva correlação. Isso sugere uma forte associação entre as medições de nível jusante e os outros parâmetros relacionados ao fluxo de água, fornecendo percepções sobre a dinâmica do sistema fluvial ou hidrológico sob investigação.

No que diz respeito aos dados climáticos, destacou-se a temperatura como a variável que apresentou a mais significativa correlação. Essa observação sugere que as flutuações na temperatura estão fortemente relacionadas a outros parâmetros climáticos presentes no conjunto de dados. Essa informação pode ser fundamental para compreender padrões sazonais, identificar tendências climáticas e antecipar possíveis impactos ambientais associados a variações de temperatura.

Durante a análise abrangente deste conjunto de dados, foi observado que entre todas as variáveis estudadas, a precipitação e a velocidade do vento foram as que apresentaram a menor contribuição em relação ao atributo de interesse, o qual é o nível de água a montante da eclusa.

A precipitação, apesar de ser um fator climatológico fundamental, pareceu ter uma correlação baixa relativamente com o nível da água da montante. Isso sugere que, em termos gerais, a quantidade de precipitação registrada não demonstrou uma influência significativa nas variações observadas no nível da água. Outros fatores ou variáveis do ambiente fluvial podem estar mais diretamente relacionados às mudanças no nível de água.

De maneira similar, a velocidade do vento também revelou uma relação menos expressiva com o atributo de interesse. Embora a velocidade do vento seja um componente importante no estudo de condições climáticas, sua contribuição específica para variações no nível da água pode ser relativamente limitada neste contexto específico.

Essas observações destacam a complexidade das interações entre as variáveis estudadas e ressaltam a importância de uma análise criteriosa para compreender as variantes do sistema em estudo. A compreensão de como diferentes fatores climatológicos e ambientais contribuem para as variações no nível da água pode ser crucial

para previsões mais precisas e para a gestão eficiente de recursos hídricos.

4.1.1.2 Conjunto de Dados II

Neste conjunto de dados, o qual faz parte da segunda etapa deste estudo, considerou todos os atributos independentes, ou seja, os mesmo considerados no primeiro conjunto de dados, a sua única diferença é relacionada a quantidade de dados que neste, é maior. Esta segunda etapa, foi idealizada em separado em virtude desses novos dados serem fornecidos pelos engenheiros hídricos posteriormente.

Este conjunto de dados inicia em 01/09/1979, mesma data que inicia o primeiro conjunto de dados, no entanto, finda em 31/01/2021. Totalizando 15345 registros diários, um acréscimo de 936 registros em relação ao primeiro conjunto.

Neste segundo conjunto de dados, notamos que embora os atributos independentes sejam os mesmos em comparação com o conjunto anterior, a principal distinção reside na quantidade de dados disponíveis. A ampliação do conjunto de dados oferece uma série de benefícios que contribuem significativamente para uma previsão mais precisa e robusta.

O aumento da quantidade de dados proporciona uma representação mais abrangente e diversificada das relações subjacentes entre os atributos independentes e a variável de interesse. Isso é crucial para a construção de modelos de predição mais robustos, uma vez que um conjunto maior oferece maior variabilidade, permitindo capturar uma gama mais ampla de padrões e comportamentos presentes nos dados. (Müller; Guido, 2016)

Além disso, um conjunto de dados mais extenso geralmente resulta em uma melhor capacidade de generalização do modelo. Modelos treinados em conjuntos de dados maiores tendem a ter uma maior capacidade de se adaptar a diferentes cenários, o que é fundamental para garantir que as previsões sejam aplicáveis a uma variedade de condições e situações no mundo real.

A quantidade adicional de dados também pode contribuir para a redução do impacto de *outliers*, eventos raros ou variações extremas nos dados. Com mais observações, o efeito desses pontos atípicos pode ser minimizado, melhorando a estabilidade e a confiabilidade do modelo de predição.

Em outras palavras, a expansão do conjunto de dados oferece vantagens significativas para a construção de modelos preditivos mais acurados. Aumentar a quantidade de dados não apenas enriquece a análise exploratória, mas também fortalece a capacidade do modelo de aprender e generalizar padrões, resultando em previsões mais confiáveis e representativas do comportamento do sistema em estudo.

Da mesma forma do processo realizado no primeiro conjunto, foi realizado a substituição em uma janela de 90 dias nos dados ausentes, não fornecendo melhorias significativas, logo foi utilizada a média geral de cada atributo independente.

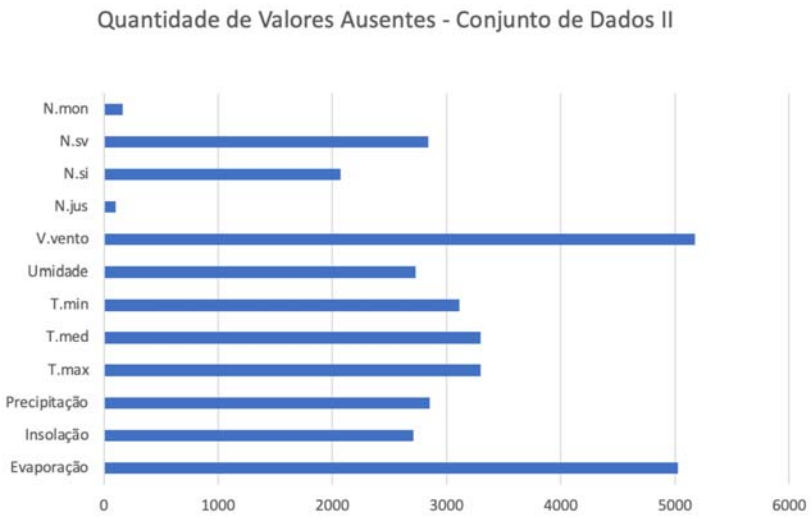

Figura 20 – Somatório Total de Dados Faltantes

Na Figura 20, observa-se de modo geral os mesmo atributos que possuem a maior quantidade de valores ausentes do conjunto 1, aparecem no conjunto 2, isto se justificar por os dados serem os mesmos, e também o acréscimo, ser relativamente pequeno considerando o tamanho total do conjunto de dados.

Neste conjunto de dados, observou-se que a variável dependente, apresenta uma quantidade limitada de dados faltantes. A presença de poucos valores ausentes para a variável dependente é positiva, pois permite uma análise mais robusta e completa, contribuindo para a confiabilidade das conclusões extraídas.

A existência de poucos dados faltantes também simplifica o processo de análise exploratória e estatística descritiva, permitindo uma compreensão mais clara da distribuição, tendências e comportamentos da variável dependente. Isso é fundamental para formular hipóteses, realizar inferências e extrair considerações significativas a partir dos dados.

Além disso, a presença de poucos dados faltantes na variável dependente facilita a execução de análises de correlação e regressão, proporcionando resultados mais confiáveis e representativos das relações entre essa variável e as variáveis independentes.

Na análise da Figura 21, observa-se um aumento da quantidade de *outliers* de 231 valores no atributo N.si e precipitação de 200 valores em relação ao conjunto anterior. Esses são os aumentos mais significativos, mantendo-se os mesmos valores nos restantes dos atributos.

As adaptações foram as mesmas utilizadas no conjunto anterior que ao invés da remoção do *outlier*, foi realizada a substituição por valores próximos aos limites estabelecidos pelo IQR.

A presença limitada de *outliers* em um conjunto de dados oferece benefícios substanciais para a qualidade e confiabilidade das análises realizadas. *Outliers* são valores

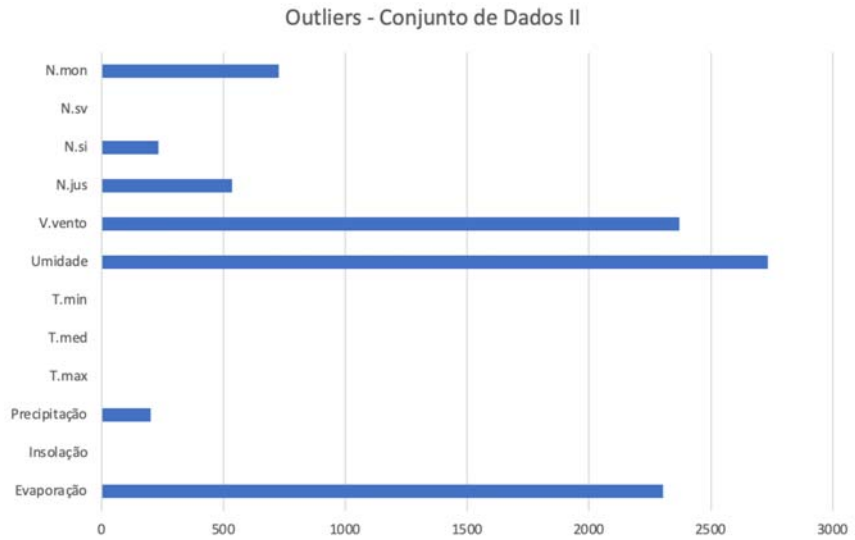

Figura 21 – Outliers do Conjunto de Dados II

atípicos que se desviam significativamente do padrão geral dos dados e, quando minimizados, proporcionam diversos pontos positivos.(Müller; Guido, 2016)

Em primeiro lugar, a escassez de *outliers* facilita a interpretação e compreensão dos dados. Isso porque valores extremos podem distorcer a média e a variância, impactando negativamente as estatísticas descritivas e prejudicando a representação fiel do comportamento geral do conjunto de dados.

Além disso, a presença limitada de *outliers* torna os modelos estatísticos mais robustos e estáveis. Modelos construídos em conjuntos de dados com poucos valores extremos são menos suscetíveis a influências distorcidas, resultando em previsões mais precisas e confiáveis.

A redução de *outliers* favorece a generalização dos modelos para novos conjuntos de dados. Modelos treinados em dados mais consistentes têm maior probabilidade de se adaptar a situações variadas, resultando em uma aplicabilidade mais ampla e eficaz das descobertas obtidas.

Neste conjunto de dados, observamos um cenário promissor em que as variáveis climáticas de umidade e precipitação demonstraram uma correlação mais acentuada com o atributo dependente em comparação com outras variáveis. Um fator-chave para esse resultado positivo foi a inclusão de um volume adicional de dados, consideravelmente mais atuais.

A introdução de dados mais recentes pode ter desempenhado um papel significativo no aprimoramento da correlação, proporcionando uma visão mais atualizada das condições ambientais. Isso é particularmente valioso em contextos climatológicos, onde as condições meteorológicas podem variar ao longo do tempo e a disponibilidade de dados mais recentes reflete de maneira mais precisa o estado atual do ambiente.

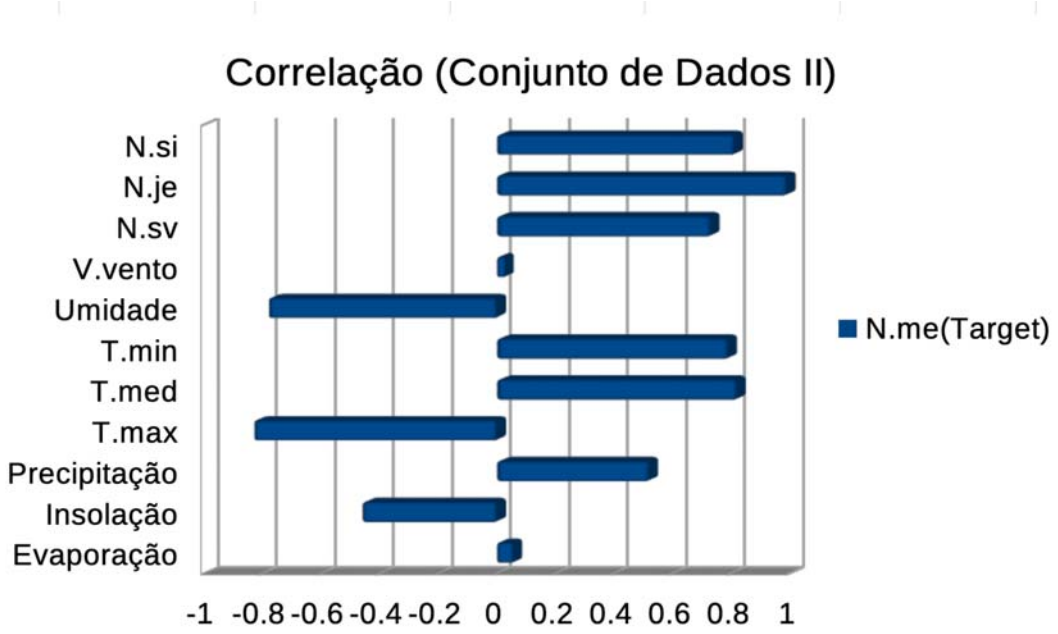

Figura 22 – Correleção do Conjunto de Dados II

A variável de umidade, ao lado da precipitação, emergiu como uma influência mais forte no atributo dependente, sugerindo que as condições atmosféricas atuais exercem um impacto mais substancial nas variações observadas. Essa informação é crucial para a compreensão de padrões climáticos, eventos extremos e suas implicações no atributo dependente em questão.

A relação entre evaporação e nível de água em corpos hídricos pode ser complexa e indireta. Embora a evaporação represente a perda direta de água para a atmosfera, sua influência sobre o nível de água pode ser mediada por uma variedade de fatores ambientais. Fatores como temperatura, umidade do ar, velocidade do vento e exposição solar afetam a taxa de evaporação, que por sua vez pode afetar o nível de água ao longo do tempo. Em dias quentes e ensolarados, com baixa umidade e ventos moderados, a taxa de evaporação tende a ser mais alta, levando a uma potencial redução no nível de água do corpo hídrico. No entanto, a influência da evaporação no nível de água pode ser mitigada por outros fatores, como a precipitação, a presença de vegetação ao redor do corpo d'água e a disponibilidade de fontes de água adicionais. Portanto, embora a evaporação possa não ter uma relação direta e imediata com o nível de água, ela desempenha um papel significativo no balanço hídrico geral do sistema e pode influenciar seu estado ao longo do tempo em conjunto com outros processos hidrológicos.

Em resumo, enquanto a evaporação é um fator importante no ciclo hidrológico, a temperatura, umidade e insolação têm uma relação direta com uma variedade de processos, incluindo a evaporação, a formação de nuvens e a ocorrência de precipitação. Portanto, é provável que esses fatores tenham uma correlação mais forte com o nível de água em comparação com a evaporação, que é apenas um componente do

balanço hídrico geral.

A inclusão de dados mais atuais não apenas aprimora a correlação, mas também fortalece a validade e a relevância das análises realizadas. Essa atualização contínua do conjunto de dados possibilita uma adaptação dinâmica às mudanças nas condições ambientais, proporcionando uma base mais sólida para a modelagem e a previsão.

Em outras palavras, a combinação de dados climáticos mais recentes, especialmente relacionados à umidade e precipitação, contribuiu para uma correlação mais significativa com o atributo dependente neste conjunto de dados, enriquecendo a análise e proporcionando percepções mais pertinentes e aplicáveis às condições atuais.

Neste conjunto de dados, identificou-se que as variáveis de evaporação e velocidade do vento apresentaram a menor correlação com o atributo dependente, semelhante ao que foi observado no conjunto de dados anterior. Essa observação reforça a consistência dos padrões identificados, apontando para uma tendência geral em ambas as análises.

A variável de evaporação, que reflete a taxa de vaporização da água, e a velocidade do vento, indicativa da movimentação atmosférica, revelaram correlações menos pronunciadas com o atributo dependente. Isso sugere que, apesar de serem fatores climáticos relevantes, essas variáveis podem ter uma influência mais limitada nas variações observadas no atributo dependente quando comparadas a outras variáveis climáticas.

Essa consistência nos resultados entre conjuntos de dados distintos pode indicar uma tendência robusta e replicável, enfatizando a importância relativa de certos fatores ambientais em relação ao atributo dependente em estudo. A análise dessas variáveis menos correlacionadas oferece concepções interessantes sobre os elementos que têm uma influência mais sutil ou complexa nas condições do sistema.

4.1.1.3 Conjunto de Dados III

Na fase seguinte deste estudo, foi elaborado o terceiro conjunto de dados, neste caso foram reduzidos os atributos preditivos, uma solicitação advinda dos Engenheiros Hídricos, tendo em vista a facilidade de conseguir estes dados. As informações consideradas neste conjunto foram: todos os dados de nível e as informações climatológicas de precipitação e velocidade do vento.

O terceiro conjunto de dados, inicia em 29/01/1979 esta é a mesma data que os demais iniciam, e finda 31/01/2021 mesma data que terminam os registro do segundo conjunto de dados. No total são 15345 registros, ou seja, a mesma quantidade do segundo conjunto.

A decisão de reduzir as variáveis preditivas no terceiro conjunto de dados, especialmente considerando que estas são de natureza climatológica e a previsão se destina ao nível de água, pode trazer benefícios significativos para a modelagem e interpreta-

ção dos resultados.

Uma redução cuidadosa nas variáveis preditivas simplifica o modelo, tornando-o mais compreensível e interpretável. (Hastie et al., 2009) Em contextos climatológicos, onde a interação entre variáveis pode ser complexa, a simplificação do modelo facilita a identificação de padrões relevantes e a compreensão das relações causais.

A decisão de reduzir as variáveis preditivas no terceiro conjunto de dados, especialmente considerando que estas são de natureza climatológica e a previsão se destina ao nível de água, pode trazer benefícios significativos para a modelagem e interpretação dos resultados.

Além disso, a redução de variáveis contribui para evitar a multicolinearidade, um fenômeno no qual variáveis independentes estão altamente correlacionadas, dificultando a identificação do impacto individual de cada variável sobre a variável dependente. (Hastie et al., 2009). Isso é particularmente importante em previsões climatológicas, onde a sobreposição de efeitos de diferentes variáveis pode ser desafiadora.

Essa abordagem também pode contribuir para mitigar o risco de *overfitting*, um fenômeno no qual o modelo se ajusta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalização para novos dados. A redução de variáveis ajuda a criar modelos mais robustos e adaptáveis a diferentes condições climáticas.

Em resumo, a redução do número de atributos previsores neste conjunto de dados climatológicos enriquecidos com informações de nível de água representa uma abordagem direcionada para otimizar a precisão e a eficiência do modelo preditivo. Ao simplificar e concentrar-se nas variáveis mais relevantes, busca-se construir um modelo mais robusto e capaz de oferecer previsões mais precisas e úteis para o contexto climático em questão.

Além da redução do número de atributos previsores, é relevante destacar que a gestão de dados ausentes e a identificação de *outliers* foram abordadas de maneira consistente com o conjunto de dados anterior. A manutenção dessa abordagem busca garantir a consistência metodológica e a confiabilidade na análise, mesmo diante da redução de variáveis previsoras.

A identificação e tratamento de *outliers* também seguiram a mesma metodologia adotada no conjunto de dados anterior. A atenção a esses valores extremos permanece crucial para evitar distorções no modelo preditivo, garantindo que a previsão do nível de água seja influenciada apenas por padrões significativos e não por eventos atípicos que poderiam comprometer a precisão das previsões.

Dessa forma, a consistência nas práticas de gerenciamento de dados, mesmo com a redução do número de atributos previsores, destaca o compromisso com a integridade e a confiabilidade dos resultados. Manter uma abordagem sólida em relação a dados ausentes e *outliers* contribui para a robustez do modelo preditivo, sustentando

a qualidade das análises realizadas no contexto climatológico.

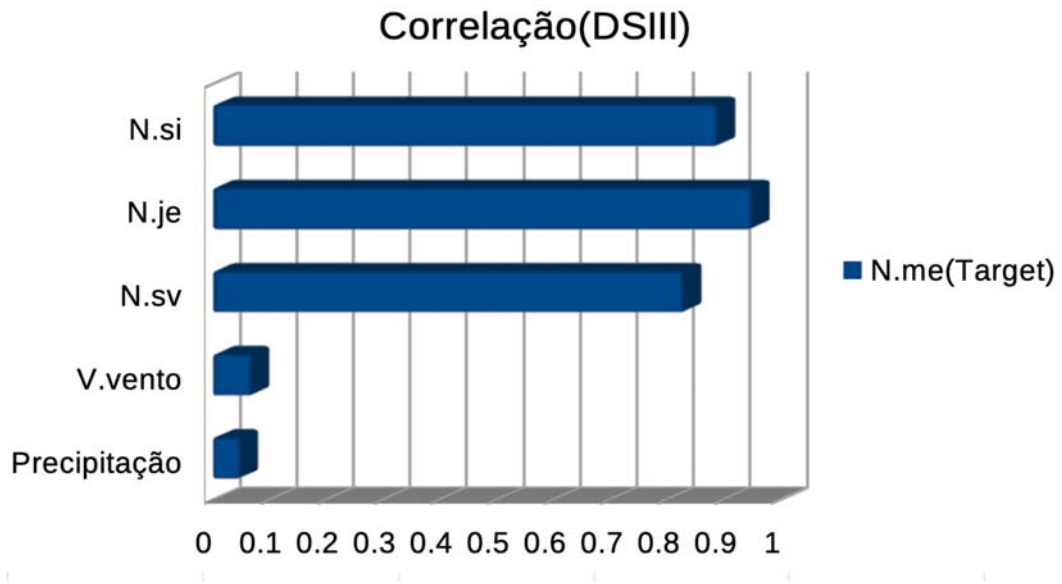

Figura 23 – Correleção do Conjunto de Dados III

Na Figura 23, observa-se a correlação do terceiro conjunto de dados, onde a redução do número de atributos previsores neste conjunto de dados climatológicos, embora tenha trazido benefícios em termos de simplicidade e eficiência, teve impacto na matriz de correlação entre as variáveis remanescentes. Essa mudança merece atenção, pois a correlação entre as variáveis é fundamental para entender as relações subjacentes e orientar a construção de modelos mais precisos.

A redução de atributos pode resultar em uma alteração nas dinâmicas de correlação, uma vez que variáveis importantes anteriormente presentes na análise podem ter sido removidas. A correlação entre as variáveis restantes pode se manifestar de maneira diferente, afetando a compreensão do impacto relativo de cada fator climatológico no nível de água.

Por outro lado, a redução do número de atributos pode simplificar a identificação de correlações mais diretas e facilmente interpretáveis, focando em fatores mais impactantes e eliminando a influência de variáveis menos relevantes.

Assim, ao analisar as mudanças na matriz de correlação após a redução de atributos, é fundamental compreender como essas alterações podem influenciar a capacidade do modelo em capturar as percepções das relações climatológicas relevantes para a predição do nível de água. A adaptação estratégica do modelo às novas circunstâncias é crucial para garantir a validade e a eficácia das previsões.

4.1.1.4 Conjunto de Dados IV

Neste novo conjunto de dados, observamos uma mudança significativa em relação ao conjunto anterior. Apesar de ter um tamanho consideravelmente menor, este

conjunto possui uma característica valiosa: a inclusão dos previsores de direção do vento das estações de monitoramento de Santa Isabel e São Gonçalo. Essa adição é particularmente relevante, pois essas informações podem desempenhar um papel crucial na compreensão das condições climatológicas e, por conseguinte, na predição do nível de água.

Embora a quantidade de dados seja menor em comparação com o conjunto anterior, a vantagem de possuir informações mais atuais pode ser crucial para análises mais precisas e alinhadas com as condições climáticas recentes. Dados mais recentes proporcionam uma visão mais dinâmica e adaptável das mudanças climáticas, permitindo que o modelo preditivo capture melhor as variações no nível de água associadas a condições climáticas específicas.

A inclusão dos previsores de direção do vento amplia a gama de informações climatológicas consideradas no modelo, oferecendo uma perspectiva mais abrangente sobre como os padrões de vento podem impactar o sistema em estudo. A direção do vento é um fator crítico, pois pode influenciar a distribuição espacial de elementos climatológicos, como chuvas e temperatura, afetando diretamente o nível de água (Maslin, 2014).

Embora o tamanho menor do conjunto de dados possa trazer desafios em termos de representatividade estatística, a inclusão de previsores adicionais pode compensar essa limitação, fornecendo percepções valiosas para a construção de um modelo mais robusto.

Portanto, este novo conjunto de dados, apesar de sua dimensão reduzida, apresenta um potencial significativo devido à inclusão de informações atualizadas e à consideração de variáveis adicionais relacionadas à direção do vento. Essas características podem enriquecer a análise e a modelagem, proporcionando uma compreensão mais refinada das relações climatológicas e suas implicações no nível de água.

Este conjunto de dados abrange um período substancial, iniciando em 05 de janeiro de 2021 e estendendo-se até 30 de junho de 2023, totalizando 908 registros. Apesar de apresentar um tamanho menor em comparação com conjuntos de dados anteriores, a riqueza temporal dessas informações é inegável.

A limitação em termos de quantidade é compensada pela qualidade temporal e pela inclusão dos previsores de direção do vento das estações de monitoramento de Santa Isabel e São Gonçalo. Essas variáveis adicionais podem oferecer informações interessantes sobre a influência da circulação atmosférica local nas condições do sistema. Dessa forma, mesmo com um conjunto de dados mais compacto, a combinação de informações atuais e variáveis climatológicas específicas pode contribuir significativamente para uma análise mais focada e uma modelagem preditiva mais precisa.

Analisando a Figura 24, no que se refere à integridade dos dados, destaca-se que a variável de precipitação se destaca pela ausência completa de registros não pre-

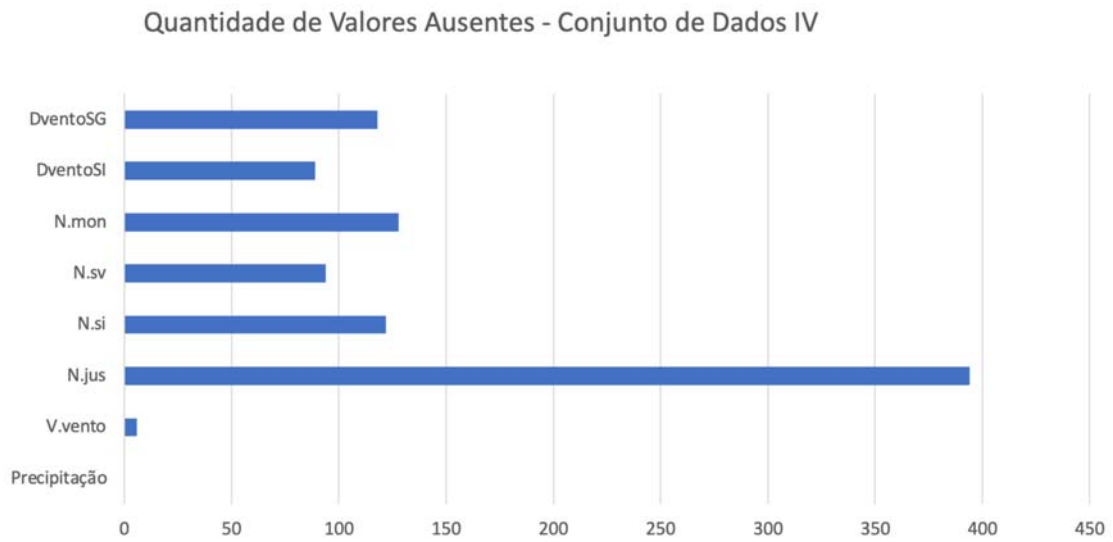

Figura 24 – Somatório Total de Dados Faltantes do Conjunto IV

enchidos, proporcionando uma base robusta e confiável para análises subsequentes. A disponibilidade completa desses dados é crucial, especialmente ao considerar seu papel significativo nas condições climáticas e na dinâmica hidrológica associada.

No entanto, ao analisar os dados de direção do vento, observa-se que a estação de monitoramento de São Gonçalo apresenta uma quantidade substancial de dados ausentes em comparação com a estação de Santa Isabel. Outra informação prejudicial, foi a excesso de dados faltantes de nível Jusante.

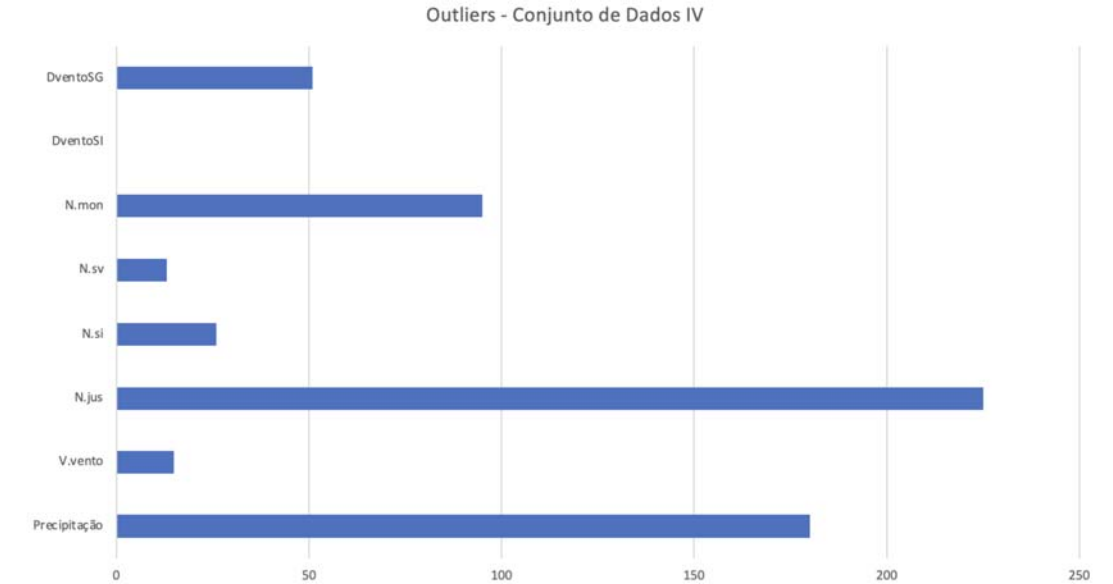

Figura 25 – Outliers do Conjunto de Dados IV

Na Figura 25, temos a análise dos *outliers* neste conjunto de dados que revela padrões intrigantes, especialmente na variável de precipitação. A presença de inúmeros *outliers* nessa variável sugere eventos climáticos extremos, como o El Niño e La

Niña, que podem provocar variações substanciais nos padrões de precipitação. Esses fenômenos climáticos, conhecidos por sua influência global, podem desencadear condições meteorológicas excepcionais, resultando em chuvas intensas ou períodos de seca acentuada.

Essas flutuações extremas na precipitação, refletidas pelos *outliers*, têm implicações diretas nos níveis de água tanto a jusante quanto a montante. A intensificação dos *outliers* nos níveis de água sugere uma resposta sensível do sistema hidrológico a eventos climáticos extraordinários. Quando a precipitação atinge extremos, os rios e corpos d'água podem experimentar aumentos significativos nos níveis, causando tanto impactos imediatos quanto efeitos prolongados nas áreas a jusante e montante.

A identificação e compreensão desses *outliers* são fundamentais para avaliar a vulnerabilidade do sistema hidrológico a eventos climáticos extremos e para antecipar possíveis consequências, como inundações ou secas prolongadas. Portanto, ao considerar os *outliers*, é crucial contextualizar essas observações extraordinárias no cenário climático mais amplo, reconhecendo o potencial impacto dos fenômenos climáticos globais na dinâmica local da água.

Ao contrastar com a variável de precipitação, observa-se um número reduzido de *outliers* no atributo de direção do vento em Santa Isabel neste conjunto de dados. Essa relativa escassez de valores atípicos sugere uma estabilidade mais consistente na direção do vento ao longo do período considerado. Ao contrário de variáveis mais suscetíveis a eventos climáticos extremos, como a precipitação, a direção do vento pode exibir uma menor variação em situações normais.

A diminuição dos *outliers* na direção do vento pode indicar uma certa regularidade nas condições atmosféricas associadas a essa variável. Isso pode ser atribuído à prevalência de padrões climáticos mais estáveis ou à menor sensibilidade da direção do vento a eventos climáticos excepcionais, pelo menos no período considerado. Essa estabilidade relativa pode contribuir para uma melhor previsibilidade nas condições climáticas relacionadas à direção do vento, facilitando análises e modelagens mais precisas.

Durante a análise dos dados de nível de água na parte Jusante, observamos um considerável número de *outliers*. Esses *outliers* podem ter um impacto significativo nas análises estatísticas e, mais especificamente, na correlação com a variável dependente, a qual é o nível Montante.

A presença de *outliers* pode distorcer as medidas de tendência central e afetar a interpretação das relações entre as variáveis. No contexto dos dados de nível de água, é crucial entender como esses *outliers* na parte Jusante podem influenciar a correlação com o nível Montante.

A correlação entre as duas variáveis pode ser distorcida devido à presença desses pontos atípicos. Se os *outliers* na parte Jusante estão associados a condições ex-

tremas ou eventos incomuns, sua influência na correlação pode ser desproporcional. Esses valores extremos podem afetar negativamente a estabilidade e a consistência das análises, dificultando a identificação de padrões reais entre as variáveis.

Além disso, é importante considerar se os *outliers* na parte Jusante são resultados de erros de medição, falhas nos sensores ou eventos anômalos genuínos. A abordagem para lidar com esses outliers dependerá da causa subjacente.

Ao considerar a interação entre a direção do vento e os níveis de água, a redução de *outliers* nessa variável sugere uma resposta hidrológica mais controlada em relação às variações na direção do vento. Esse entendimento é crucial para uma avaliação das condições climáticas e do impacto subsequente no sistema hidrológico, permitindo uma abordagem mais refinada na interpretação dos dados.

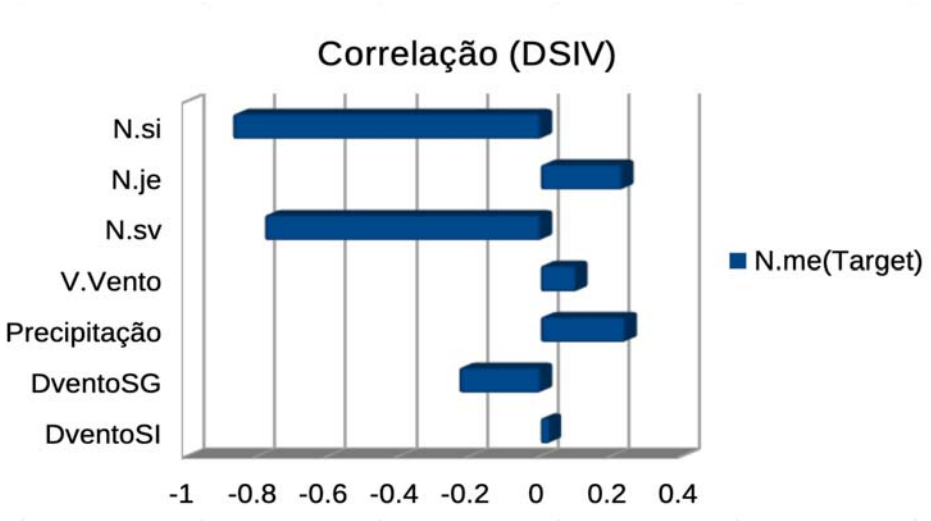

Figura 26 – Correlação do Conjunto de Dados IV

Conforme se observa na Figura 26 a inclusão dos dados de direção do vento provenientes das estações de monitoramento de São Gonçalo e Santa Isabel neste conjunto de dados trouxe variantes interessantes às relações entre as variáveis climatológicas e o nível de água. Ressaltamos que, uma mudança nas características que exercem maior influência sobre a variável dependente.

Diferentemente dos conjuntos de dados anteriores, onde a correlação era predominantemente influenciada pela variável de Nível Jusante Eclusa, agora identificamos que aos dados de nível de Santa Isabel desempenham um papel mais proeminente na variação do nível de água da variável dependente. Essa alteração nas características determinantes destaca a importância da consideração específica das direções do vento provenientes de São Gonçalo e Santa Isabel na modelagem hidrológica.

Anteriormente, nos conjuntos de dados, a correlação predominante das informações de nível era direcionada mais fortemente pela estação de jusante, Eclusa São Gonçalo. No entanto, com a inclusão dos dados de direção do vento e também pelo fato dos dados serem mais atuais, observamos uma mudança considerável, onde o

nível da estação de monitoramento de Santa Isabel emerge como um fator de maior influência no nível de água da variável dependente.

A análise mais aprofundada das variáveis climatológicas revelou padrões interessantes em relação ao nível de água no montante São Gonçalo. Dentre essas variáveis, a precipitação e a direção do vento proveniente de São Gonçalo emergem como fatores de influência mais proeminentes. Esta observação sugere que, muito provavelmente, a proximidade geográfica dessas duas estações exerce um papel significativo nas condições hidrológicas da região.

A direção do vento, quando oriunda de São Gonçalo, aparentemente, desempenha um papel mais relevante no comportamento do nível de água no montante. Essa correlação mais forte pode ser explicada pela proximidade espacial das estações de monitoramento, indicando que as condições atmosféricas específicas nessa região têm um impacto mais direto nas variações do nível de água.

Essa observação ressalta a importância de considerar não apenas a direção do vento, mas também a precipitação, ao avaliar as influências climatológicas nas condições hidrológicas. O conhecimento detalhado dessas interações específicas pode informar estratégias mais precisas de gerenciamento de recursos hídricos e aprimorar a capacidade de previsão em áreas onde fatores geográficos e climatológicos desempenham um papel crucial.

Adicionalmente, ao analisar as variáveis climatológicas, destaca-se que a direção do vento proveniente da estação de monitoramento de Santa Isabel e a velocidade do vento nessa região mostraram ter uma influência relativamente menor no nível de água da variável dependente, a qual é o montante São Gonçalo.

A falta de uma correlação tão forte pode ser atribuída a diversos fatores, como a distância geográfica, características topográficas específicas da região ou padrões climatológicos locais. A direção do vento e a velocidade do vento em Santa Isabel podem não exercer uma influência tão direta no comportamento hidrológico na área do montante São Gonçalo, mas pode ter uma relação indireta ajudando outros atributos na predição.

Essa compreensão diferenciada ressalta a importância de avaliar cada variável em relação à sua relevância específica para as condições locais. A análise detalhada desses padrões permite uma interpretação mais precisa dos fatores que verdadeiramente impactam o nível de água, proporcionando uma base sólida para estratégias de gerenciamento de recursos hídricos e modelos de previsão mais refinados.

Observa-se na Figura 27 que a variação do nível de água em determinadas áreas está intrinsecamente ligada a diversos fatores, sendo a direção do vento um dos elementos-chave que influenciam esse ciclo hidrológico. No período compreendido entre maio a agosto, observamos um ciclo ascendente no nível de água, contrastando com a fase de declínio que ocorre nos meses subsequentes, de setembro a abril.

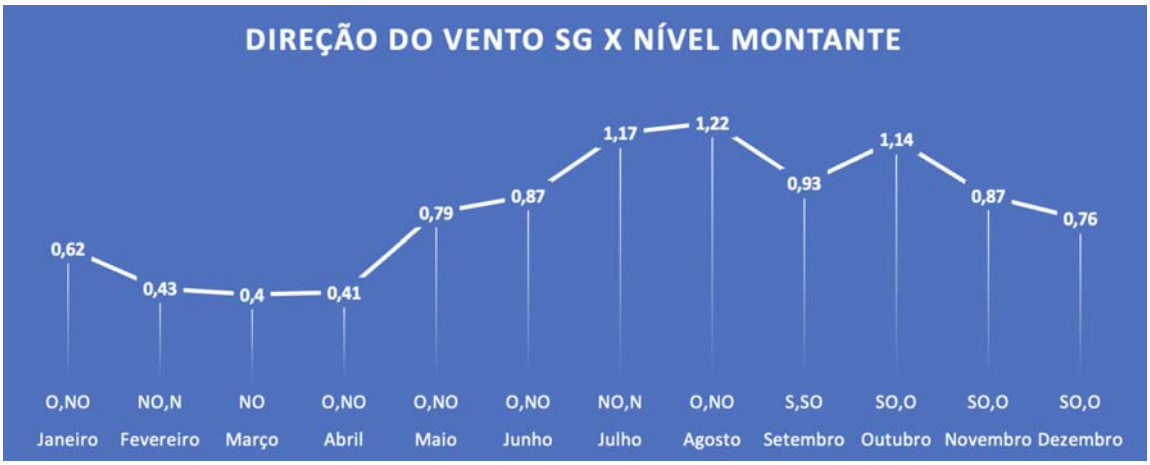

Figura 27 – Direção do Vento SG x Nível de Água

Este fenômeno revela uma relação direta entre a direção predominante do vento e as oscilações no nível de água. Durante o ciclo de aumento, a direção predominante do vento assume uma orientação oeste ao noroeste. Essa configuração propícia favorece a acumulação de água, conduzindo ao aumento do nível nos corpos hídricos em questão.

À medida que adentramos os meses de setembro a abril, a direção predominante do vento altera-se para sul a sudoeste, desencadeando um período de diminuição no nível de água. Esta mudança na orientação do vento promove a dissipação e evaporação das águas acumuladas, contribuindo para a fase descendente do ciclo hidrológico.

A compreensão dessa interconexão entre a direção do vento e as variações no nível de água é crucial para o manejo sustentável desses recursos hídricos. Estratégias de gestão e previsão que levem em consideração esses padrões climáticos podem ser implementadas para otimizar o uso e mitigar impactos adversos, garantindo a sustentabilidade desses ecossistemas ao longo do tempo.

Figura 28 – Direção do Vento SI x Nível de Água

Ao examinar a Figura 28 que retrata a variação do nível de água em uma região

específica, observamos um padrão semelhante ao previamente discutido. Neste cenário, a dinâmica do ciclo hidrológico mantém-se consideravelmente constante, onde a diminuição do nível de água está intrinsecamente relacionada à direção predominante do vento para sul-Sudoeste, enquanto o aumento correlaciona-se com a orientação oeste a noroeste.

Ao longo dos meses de maio a agosto, período caracterizado pelo aumento do nível de água, a direção predominante do vento assume uma orientação oeste a noroeste. Esse fenômeno facilita a acumulação de água na região, criando um cenário propício para o acúmulo hídrico.

Por outro lado, nos meses subsequentes, de setembro a abril, a mudança na direção predominante do vento para sul a sudoeste desencadeia a fase de diminuição no nível de água. Este movimento do vento contribui para a dispersão e evaporação da água acumulada, caracterizando a parte descendente do ciclo hidrológico.

Segundo (Hartmann; Harkot, 1990), a direção do vento desempenha um papel crucial nas variações do nível da água no Canal São Gonçalo, situado em Pelotas. Este fenômeno é influenciado diretamente pela proveniência do vento, especificamente quando este se manifesta entre Nordeste e Noroeste, ou entre Sudeste e Sudoeste.

Quando o vento sopra de Nordeste a Noroeste, sua força e trajetória exercem uma pressão favorável que resulta em um considerável aumento do nível de água no canal. Este aumento pode ser atribuído ao empurrão do vento sobre a superfície da água, criando uma acumulação significativa que se manifesta como um aumento mensurável no nível hídrico. Essa relação entre a direção do vento e o aumento do nível de água é consistentemente observada e observada por estudos anteriores.

Contrastando com essa dinâmica, quando o vento provém das direções Sudeste a Sudoeste, ocorre um efeito oposto. O vento nessa faixa de direções exerce uma pressão que contrai a superfície da água, resultando em uma diminuição mensurável do nível do canal. A influência do vento proveniente dessas direções específicas é evidente ao analisar os padrões de variação do nível da água ao longo do tempo.

4.1.1.5 Importância dos Atributos Independentes

Na seção anterior, exploramos a correlação de Spearman para compreender as relações diretas entre variáveis em nosso conjunto de dados. Essa análise nos proporcionou percepções sobre como diferentes variáveis se relacionam com a variável alvo. No entanto, compreender não apenas as relações diretas, mas também a importância relativa das *features* pode ser fundamental para a análise.

Nesta seção, será explorado a funcionalidade de importância das *features* do algoritmo *Random Forest*. Enquanto a correlação de *Spearman* nos ajudou a entender as relações lineares ou monotônicas, a análise de importância das *features* nos permitirá identificar quais previsores são mais relevantes para prever nossa variável alvo,

mesmo que as relações não sejam simples ou lineares.

Ao investigar a importância dos previsores, estaremos em busca de percepções sobre quais variáveis são mais informativas para nosso modelo. Isso nos ajudará a focar nossos esforços nas *features* mais relevantes, melhorando potencialmente a precisão e a interpretabilidade de nossos modelos de previsão. Vamos agora explorar como a análise de importância das *features* do *Random Forest* pode enriquecer nossa compreensão dos dados e orientar nossas futuras decisões analíticas.

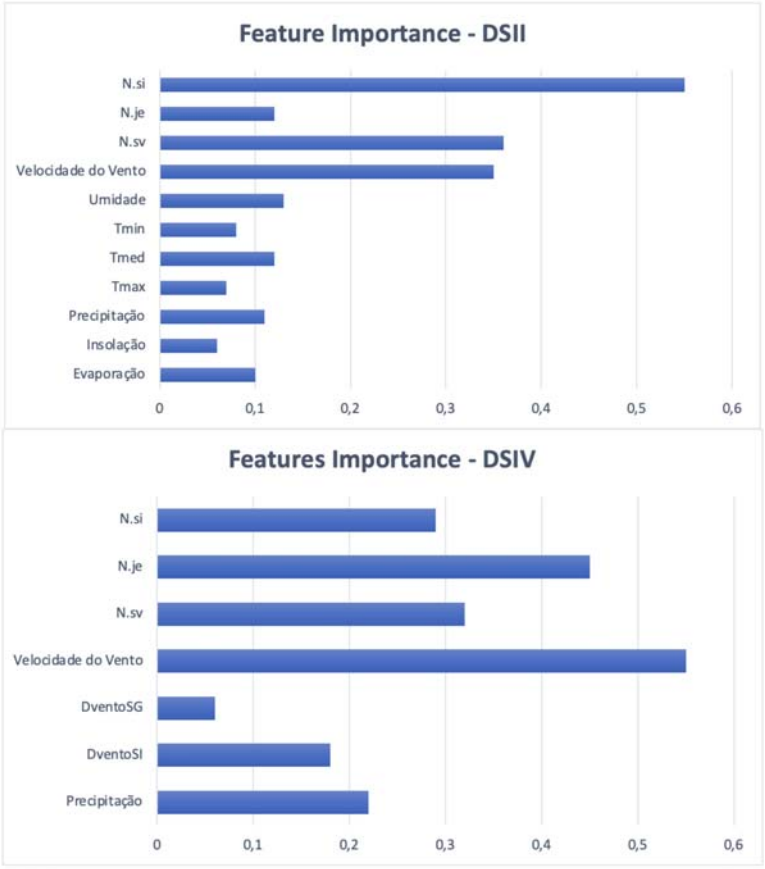

Figura 29 – Importância das Features na Previsão da Variável Alvo

Analisando o primeiro gráfico da Figura 29 , onde todos os atributos independentes estão incluídos, fica evidente que os dados do nível de Santa Isabel emergem como o atributo mais relevante para a previsão da variável alvo. Isso sugere que o nível de água em Santa Isabel é um preditor significativo da variação do nível de água total. Essa observação é particularmente interessante, pois destaca a importância específica de um local de medição em relação aos outros para a previsão do nível de água geral.

Além disso, é de se considerar a mudança em relação à análise de correlação de *Spearman*. Enquanto na análise de *Spearman*, as variáveis climáticas como temperatura, umidade e insolação foram identificadas como mais fortemente correlacionadas com o nível de água, agora, com essa análise, a velocidade do vento emerge como

o atributo climático mais forte. Isso sugere que, apesar de sua menor correlação de *Spearman*, a velocidade do vento desempenha um papel significativo na previsão do nível de água quando consideramos a interação complexa entre todas as variáveis independentes. Essa descoberta destaca a capacidade do *Random Forest* de identificar relações não lineares e complexas entre as variáveis, fornecendo percepções que podem não ser capturados por medidas de correlação tradicionais.

Além disso, é interessante notar que, apesar da evaporação não ter sido destacada como um atributo significativo na análise de correlação de *Spearman*, sua importância na previsão do nível de água aumentou consideravelmente na análise de importância das *features*. Isso sugere que, embora a evaporação possa não estar fortemente correlacionada linearmente com o nível de água, sua influência indireta e complexa no sistema hídrico pode ser capturada de maneira mais precisa pelo *Random Forest*, destacando sua capacidade de detectar relações não lineares e interações complexas entre as variáveis independentes. Essa observação ressalta a importância de considerar uma variedade de métodos analíticos para uma compreensão abrangente e precisa dos dados.

Na análise do segundo gráfico, que envolve o conjunto de dados IV, observamos que a velocidade do vento novamente emerge como o atributo mais importante na previsão da variável alvo. Isso sugere que a velocidade do vento continua sendo um fator significativo, mesmo quando consideramos um conjunto de dados diferente. Essa consistência na importância da velocidade do vento destaca sua influência substancial na dinâmica do sistema hídrico.

Além disso, o dado climático de precipitação surge como outro atributo de destaque em termos de importância. Isso indica que a quantidade de precipitação desempenha um papel fundamental na variação do nível de água, o que não é surpreendente, considerando seu impacto direto na entrada de água nos corpos hídricos.

Outro aspecto interessante é a Direção do Vento de Santa Isabel, que também é identificada como uma variável importante. Isso sugere que a direção do vento pode influenciar indiretamente a dinâmica do sistema hídrico, possivelmente afetando padrões de evaporação, distribuição de chuvas e outras variáveis climáticas.

Por fim, é considerável que o dado de nível Jusante surge como o atributo de nível mais influente na variável alvo. Isso pode ser justificado pelo fato de que o nível Jusante representa a condição do fluxo de água após a região de interesse, refletindo diretamente a influência das condições hidrológicas anteriores no comportamento atual do sistema hídrico. Essa observação destaca a importância de considerar não apenas as condições locais, mas também o contexto hidrológico mais amplo ao prever variações no nível de água.

Após a análise dos conjuntos de dados disponíveis para o estudo da previsão do nível de água no Canal São Gonçalo, é fundamental avançar para a próxima etapa,

onde serão apresentados os modelos propostos. Nesta seção, serão discutidas as estruturas e características dos modelos desenvolvidos, levando em consideração as particularidades dos conjuntos de dados utilizados. Ao conectar a análise dos dados com a proposta dos modelos, se estabelece uma base para o desenvolvimento de abordagens preditivas que sejam adaptadas para este contexto específico.

4.2 Modelo I

Nesta seção será fornecido detalhes do primeiro modelo preditivo criado, para este modelo foi utilizado o Conjunto de Dados I como base. O motivo dessa escolha é que na época foi disponibilizado somente esses dados. Basicamente, esse modelo híbrido é composto pelos modelos ARIMA e *Random Forest*, ou seja, os dados lineares, como, por exemplo, os dados de nível, são utilizados com entrada pelo ARIMA e os climatológicos não lineares pelo *Random Forest*.

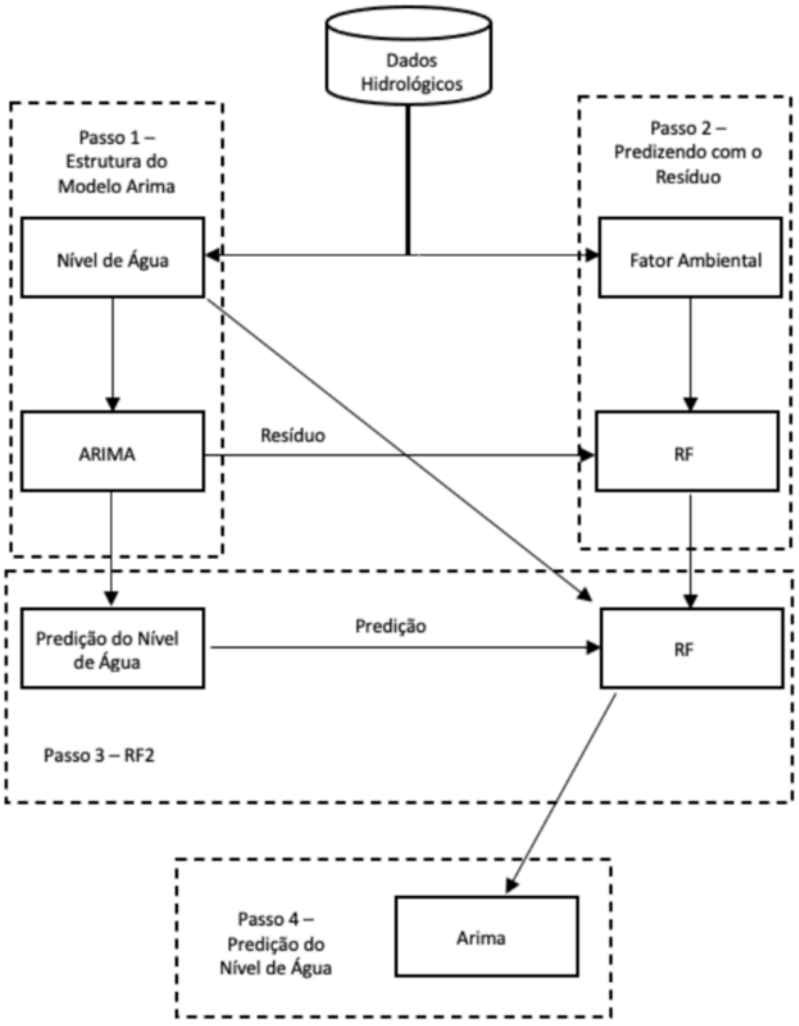

Figura 30 – Arquitetura Proposta I

Na Figura 30, pode-se observar a arquitetura proposta, onde existe inicialmente uma base de dados hidrológicos. Tendo este novo conjunto de dados, o mesmo foi dividido em dois outros *datasets* o primeiro somente com o nível de água e o segundo com os dados climáticos que são: evaporação, insolação, precipitação, temperatura máxima, média e mínima, umidade do ar e velocidade do vento.

Continuando a análise da Figura 30, no passo 1, observa-se que os dados de nível, que possuem características lineares, são submetidos ao modelo ARIMA. No segundo passo, os resíduos do modelo ARIMA, ou seja, a parte não linear, servem como entrada para o modelo *Random Forest* com os dados ambientais. No passo 3 temos a segunda camada com o RF, onde tem como entrada a predição do nível anterior com a predição do ARIMA, considerando os dados originais. Por fim, no passo 4, a predição do modelo anterior é utilizado como entrada para a última camada, que utiliza o modelo ARIMA. Nesta última camada, os dados são divididos em 70% para o treinamento e 30% para o teste. Esta divisão se justifica pelo fato de apresentar um melhor desempenho preditivo nos testes executados.

No modelo proposto, os melhores valores encontrados para p,d e q foram *(2,1,2)*, ou seja, é utilizado um componente autorregressivo p de valor 2, as médias móveis q como o valor de 2 o qual tenta explicar o efeito do ruído na série. A diferenciação d, de ordem, 1 torna a série estacionária em uma mesma média.

4.2.1 Algorítimo e Análise

Nesta seção será abordado o algoritmo proposto intitulado de ARIMA-RF.

Conforme o Algoritmo 1, observa-se inicialmente a variável act, que recebe a autocorrelação total, servindo de base para a escolha do número de lags da série hidrológica, ou seja, quantos períodos para trás será utilizado para a previsão. Em d é aplicado a função das diferenças para deixar a base estacionária na média 0. É aplicado também a função AIC para obter o melhor modelo ARIMA e os melhores "p", "d"e "q".

Continuando a análise do Algoritmo 1 as variáveis: pdc e res armazenam a predição do modelo ARIMA e os resíduos. No modelo RF1, a variável $baserf1$ armazena os resíduos do modelo anterior com os dados de satélite. $x1$ contém os dados normalizados com as informações preditoras e $y1$ possui o dado dependente, também normalizado.

O modelo RF2 possui como entradas: pdc, $rnn1$ e $base_h$, ou seja, tem como parte que compõe a variável $baserf2$, a predição do ARIMA, as predições do modelo anterior e a base de nível original. Esta etapa realiza o processo inverso da normalização, para que os dados de nível sejam expressos em $metros$ e, assim, se obtém as predições. O resultado das predições são armazenados em $rf2$.

Na última camada aplica-se o modelo ARIMA. A entrada é o nível de água do

Algoritmo 1: ARIMA-RF

Entrada: base-h, base hidrológica base-e, base climatológica
Saída: predição do nível de água

```
   // Modelo ARIMA 1
 1 act = ACF(base-h);
 2 d = diff(base-h);
 3 para p=0 até pmax faça
 4     para d=0 até dmax faça
 5         para q=0 até qmax faça
 6             m = ARIMA(base-h,(p,d,q)) //Contrução do Modelo Arima;
 7             AIC = aic(m)
 8         fim
 9     fim
10 fim
11 m = ARIMA(base-h,(p,d,q) //Determinação do Modelo;
12 pdc = m.prdict();
13 res = m.residual();
   // Modelo RF1
14 baserf1 = res + base-e + base-h;
15 x1 = normalize(baserf1[:,0:9]);
16 y1 = normalize(baserf1[:,9]);
17 rf1 = RF(x1,y1);
   // Modelo RF2
18 baserf2 = pdc + rf1 + base-h;
19 x2 = normalize(baserf1[:,0:9]);
20 y2 = normalize(baserf1[:,9]);
21 rf2 = RF(x2,y2);
   // Modelo ARIMA 2
22 basearima2 = rf2;
23 train-size = (basearima2 * 2/3);
24 train = basearima2[:, train-size];
25 test = basearima2[train:];
26 out = ARIMA[test];
   // Processo Reverso da normalização e exibição a predição do nível
      de água
```

modelo anterior $rf2$ é armazenado em $basearima2$. O conjunto é dividido em treino e teste, onde os 70% primeiros registros são utilizados para treinamento do modelo e 30% por cento para teste. A variável out armazena a predição final do modelo.

Neste estudo, buscamos aprimorar as capacidades preditivas em relação ao nível de água, implementando dois modelos distintos: o ARIMA-RNN proposto por (Xu et al., 2019) e o ARIMA-RF, desenvolvido por nossa equipe. Ambos os modelos foram aplicados à nossa base de dados, permitindo uma comparação direta de seu desempenho.

O modelo ARIMA-RNN proposto por (Xu et al., 2019) combina a robustez do ARIMA, um modelo linear tradicional, com a capacidade de aprendizado não linear de uma Rede Neural Recorrente (RNN). Essa abordagem híbrida visa capturar padrões temporais complexos e dinâmicas não lineares presentes nos dados climatológicos, oferecendo uma estrutura mais flexível para a modelagem.

Por outro lado, introduzimos o modelo ARIMA-RF, uma combinação entre ARIMA e *Random Forest*, explorando a força preditiva das florestas aleatórias em conjunção com a abordagem de séries temporais do ARIMA. A combinação desses dois métodos visa explorar a diversidade e robustez inerentes às florestas aleatórias, enquanto ainda se beneficia da modelagem temporal precisa do ARIMA.

Ao comparar os resultados obtidos pelos dois modelos, observamos que o ARIMA-RF desenvolvido por nossa equipe apresentou desempenho superior em termos de precisão preditiva em relação ao nível de água. Essa superioridade pode ser atribuída à capacidade do modelo de capturar relações complexas e não lineares nos dados climatológicos, além de lidar eficazmente com padrões temporais.

Como se observa na Tabela 3, o ARIMA-RF apresentou menor erro preditivo em relação ao ARIMA-RNN proposto por (Xu et al., 2019). Foi realizada a implementação deste modelo e os resultados sinalizam uma melhor acurácia no modelo proposto neste trabalho.

Tabela 3 – Comparação do ARIMA-RF com o ARIMA-RNN

ARIMA - RNN	Mse: 0.3523 Mae: 0.2666 Dp: 0.0317
ARIMA - RF	Mse: 0.1292 Mae: 0.0940 Dp: 0.2624

O modelo ARIMA-RF apresenta um erro menor, atingindo uma variação de *0.054*m comparado ao ARIMA-RNN. No que se refere ao erro médio absoluto, a diferença foi de *0,038*m. Esses dados, são gerados na base de testes, onde o modelo treinado não conhece as informações. O experimento deste trabalho comprova a vantagem do modelo combinado na previsão do nível de água, mostra sua racionalidade científ-

fica e apresenta melhor desempenho preditivo. Portanto, no sistema hidrológico, este modelo pode alcançar um bom efeito na previsão do nível da água.

Essa conclusão ressalta a importância de explorar e adaptar modelos às características específicas do conjunto de dados em questão. Embora o ARIMA-RNN de (Xu et al., 2019) seja uma proposta valiosa, nosso ARIMA-RF se destacou na modelagem das relações entre os atributos climatológicos e o nível de água, demonstrando a relevância da escolha do modelo para obter resultados mais precisos em contextos hidrológicos.

4.3 Modelo II

Neste capítulo serão abordados os dados e informações da segunda parte deste estudo, nesta fase foram utilizados mais três conjunto de dados. Na fase anterior tínhamos somente o primeiro conjunto de dados sendo que neste conjunto não tínhamos os dados de nível de Santa Vitória e Santa Isabel, logo a arquitetura proposta caracterizou-se basicamente no estudo e implementação do modelo ARIMA-RF, que forneceu bons resultados e serviu como base para o entendimento desta etapa.

Na fase inicial deste estudo, propusemos um modelo híbrido interessante denominado ARIMA-RF, que combinava elementos do modelo ARIMA com o algoritmo de *Random Forest*. Esta abordagem visava aprimorar a capacidade preditiva em séries temporais, considerando tanto a dependência temporal quanto padrões mais complexos que poderiam ser capturados pela flexibilidade das Florestas Aleatórias.

Agora, na segunda fase do estudo, estamos expandindo nossa análise ao testar diversos modelos híbridos com o ARIMA. A disponibilidade de mais dados nesta etapa representa uma oportunidade significativa para melhorar ainda mais o desempenho preditivo do nosso modelo. A inclusão de um conjunto mais extenso de dados permite uma melhor compreensão das tendências, padrões sazonais e variações nos níveis de água, contribuindo para previsões mais precisas e robustas.

Uma das limitações identificadas no primeiro modelo híbrido é o fato de fornecer previsões apenas para após o término do conjunto de dados. Isso significa que, se o conjunto de dados encerrar em 2018, o modelo oferecerá previsões apenas para os 30 dias subsequentes a esse período específico, por exemplo. Essa limitação temporal impede a obtenção de previsões em tempo real ou para períodos mais recentes.

Nesta nova fase, visamos superar essa limitação, explorando modelos híbridos ARIMA aprimorados que permitam previsões mais atuais. A extensão do conjunto de dados disponível oferece a oportunidade de calibrar e ajustar modelos de maneira mais eficaz, considerando padrões temporais mais recentes e dinâmicas emergentes nos dados.

Essa abordagem aprimorada não apenas contribuirá para a precisão das previ-

sões, mas também aumentará a utilidade prática do modelo, permitindo que seja aplicado a cenários mais próximos da data atual. Ao testar e avaliar diferentes modelos híbridos, almejamos proporcionar uma ferramenta mais robusta e eficiente para prever os níveis de água, um componente crucial na gestão de recursos hídricos e na resposta a eventos hidrológicos extremos.

Na primeira fase deste estudo, apresentamos um modelo híbrido inovador denominado ARIMA-RF, que combinava elementos do modelo ARIMA e *Random Forest*. Agora, na segunda fase, estamos avançando ao testar diversos modelos híbridos com o ARIMA. Esta etapa é marcada por uma vantagem substancial: a disponibilidade de um conjunto de dados mais extenso, proporcionando uma melhoria significativa no desempenho preditivo.

Apesar das limitações identificadas na primeira fase deste estudo, é crucial reconhecer que essa etapa inicial proporcionou uma base sólida para a compreensão do caso em questão. As percepções obtidas, as tendências identificadas e as análises realizadas durante a implementação do modelo ARIMA-RF foram fundamentais para orientar o desenvolvimento e aprimoramento do estudo.

Na segunda fase, aproveitamos essas aprendizagens para implementar mudanças significativas no modelo. O foco principal foi capacitar o modelo para previsões com entradas manuais, permitindo maior flexibilidade e adaptabilidade às necessidades específicas do usuário. Agora, o modelo está configurado para receber informações manuais como *input*, tornando-se uma ferramenta mais versátil e ajustável a cenários específicos.

Essas melhorias visam proporcionar previsões mais dinâmicas e contextualmente relevantes, superando as limitações temporais da fase anterior. Ao possibilitar a inserção manual de dados no modelo, agora somos capazes de gerar previsões mais atuais, atendendo às demandas práticas de uma gestão eficiente dos recursos hídricos.

Dessa forma, a segunda fase não apenas corrige limitações, mas também representa uma evolução substancial, transformando o modelo em uma ferramenta mais ágil e adaptável. Este estudo de caso destaca a importância do aprendizado contínuo e da flexibilidade na abordagem de problemas complexos, proporcionando soluções mais robustas e eficazes ao lidar com a dinâmica específica dos dados hidrológicos.

Em outras palavras, na primeira fase do estudo, ao introduzir o modelo ARIMA-RF, a abordagem estava mais centrada em técnicas específicas de séries temporais. O ARIMA é um modelo amplamente utilizado para análise de séries temporais, enquanto o *Random Forest* é um algoritmo de aprendizado de máquina que integra elementos de árvores de decisão.

Na segunda fase, ao expandir o escopo para testar vários modelos híbridos com o ARIMA, e ao permitir a entrada manual de dados, o problema se torna mais aberto

à flexibilidade e customização, abrangendo técnicas mais amplas de aprendizado de máquina. A inclusão de entradas manuais e a busca por modelos mais flexíveis indicam uma abordagem mais orientada para soluções práticas, com a capacidade de adaptar-se a cenários específicos.

Portanto, enquanto a primeira fase estava mais direcionada a técnicas específicas de séries temporais, a segunda fase incorpora elementos mais amplos de aprendizado de máquina, tornando-se uma abordagem mais abrangente e adaptável ao contexto do problema em questão.

4.3.1 Arquitetura

Tendo em vista as considerações abordadas na seção anterior, mostraremos aqui detalhes da arquitetura proposta na segunda fase. Sendo um modelo híbrido, neste caso não deixamos pré-definido uma segunda camada com um algorítimo já fixo, em vez disso deixamos em aberto, para ser utilizada qualquer outro modelo.

Um dos desafios fundamentais que enfrentamos durante a segunda fase deste estudo foi adaptar um modelo de aprendizado de máquina para lidar efetivamente com problemas de séries temporais. O contexto de séries temporais impõe desafios únicos, uma vez que envolve a análise e previsão de dados que variam ao longo do tempo. Para superar essa complexidade, implementamos estratégias específicas, focadas na geração de *lags* na variável alvo.

A geração de *lags* na variável alvo consistiu em criar versões defasadas da própria variável alvo, representando seus valores em períodos temporais anteriores. Essa abordagem é crucial para capturar padrões temporais e correlações que podem ser cruciais na previsão de séries temporais. Ao introduzir *lags*, o modelo é capacitado para considerar não apenas o estado atual da variável alvo, mas também sua evolução ao longo do tempo.

Essa modificação no modelo foi essencial para aprimorar a capacidade de previsão em problemas complexos de séries temporais, como os relacionados aos níveis de água. Ao incorporar *lags* na variável alvo, o modelo tornou-se mais sensível a padrões temporais, ciclos sazonais e outros fatores que caracterizam os dados hidrológicos.

Esse desafio específico ilustra a importância da customização de abordagens de aprendizado de máquina para atender às peculiaridades de diferentes conjuntos de dados. Ao superar esse desafio, não apenas fortalecemos a capacidade preditiva do modelo, mas também destacamos a necessidade de estratégias especializadas ao lidar com problemas específicos de séries temporais na gestão de recursos hídricos.

Outra alteração significativa implementada nesta fase do estudo foi a transformação do modelo para permitir a geração de previsões futuras em uma única execução. Anteriormente, para prever múltiplos dias à frente, era necessário executar o modelo repetidamente, uma vez para cada período de previsão desejado. Essa modificação

representa uma otimização crucial que aprimora a eficiência do processo de previsão.

Agora, com a capacidade de gerar previsões múltiplas em uma única execução, o modelo tornou-se mais ágil e eficaz, eliminando a necessidade de execuções repetidas para cada ponto no horizonte de previsão. Essa melhoria não apenas acelera o processo de geração de previsões, mas também simplifica a implementação prática do modelo, tornando-o mais acessível e fácil de utilizar em cenários operacionais.

Essa modificação não apenas aprimora a eficiência do modelo, mas também destaca a importância de considerar a usabilidade e a praticidade ao desenvolver ferramentas de previsão. Ao otimizar o processo de execução, maximizamos a utilidade prática do modelo, permitindo uma integração mais fluida e eficiente nas operações de gestão de recursos hídricos. Essa atualização representa um passo significativo em direção a um modelo mais ágil e adaptável às demandas dos Engenheiros Hídricos.

Uma alteração substancial na arquitetura do modelo, em comparação com a abordagem anterior, foi a introdução de uma divisão mais refinada dos conjuntos de dados. Antes, utilizávamos apenas conjuntos de treinamento e teste, mas agora adotamos uma abordagem mais sofisticada, dividindo os dados em conjuntos de treinamento, validação e teste. Essa mudança visa aprimorar a capacidade do modelo de generalizar padrões, otimizando sua performace em dados não vistos.

No novo paradigma, o conjunto de treinamento continua a ser essencial para o treinamento inicial do modelo. No entanto, o conjunto de validação entra em cena como uma ferramenta crucial durante o treinamento, permitindo ajustes contínuos nos hiperparâmetros do modelo. A introdução desse conjunto de validação auxilia na seleção de modelos mais robustos e evita possíveis problemas de *overfitting*, onde o modelo se adapta excessivamente aos dados de treinamento, perdendo a capacidade de generalização.

O conjunto de teste permanece reservado para avaliar o desempenho final do modelo, proporcionando uma avaliação realista de como ele se sairá em dados completamente novos e não utilizados durante o treinamento. Essa abordagem tripartida fornece uma estrutura mais completa para o desenvolvimento e avaliação do modelo, aprimorando sua adaptabilidade e confiabilidade em cenários do mundo real.

Essa nova configuração representa um avanço significativo, garantindo que o modelo seja mais robusto, capaz de lidar com diversas variantes nos dados e, ao mesmo tempo, capaz de fornecer previsões mais precisas e generalizáveis em condições operacionais.

Uma alteração significativa na nova arquitetura do modelo foi a expansão da abrangência dos dados de nível. Ao invés de se restringir apenas ao estudo de caso específico, optamos por incorporar informações de nível de água de todo o percurso da corrente, abrangendo o sentido completo do fluxo, até seu deságue na Lagoa dos Patos. Essa ampliação no escopo dos dados tem implicações estratégicas e operaci-

onais importantes.

Ao incorporar informações de nível de água de todo o percurso da corrente, obtivemos uma visão mais abrangente e integrada das condições hidrológicas ao longo de todo o sistema. Isso possibilita uma modelagem mais precisa, considerando as interações complexas entre diferentes trechos do curso d'água e fornecendo uma base mais sólida para as previsões.

Dessa forma, a utilização de dados de nível de água de todo o cenário, desde o ponto de estudo até o deságue na Lagoa dos Patos, representa um aprimoramento estratégico, permitindo uma modelagem mais abrangente e uma compreensão mais abrangente das condições hidrológicas em jogo. Isso fortalece a capacidade do modelo de fornecer previsões mais robustas e contextualmente relevantes para a gestão eficiente dos recursos hídricos.

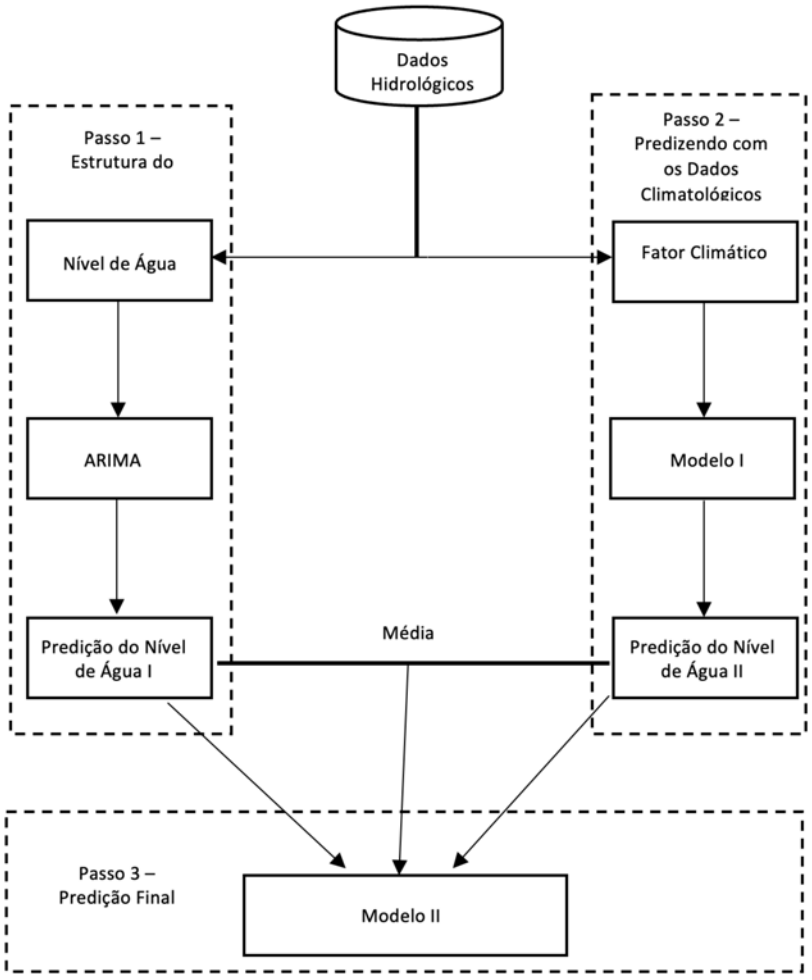

Figura 31 – Arquitetura Proposta II

Conforme a Figura 31, nesta segunda arquitetura implementada, mantivemos uma estrutura semelhante à primeira, na qual os dados de nível de água continuam sendo fundamentais para alimentar o Passo 1. No entanto, uma modificação crucial foi introduzida para aprimorar a capacidade preditiva do modelo: em vez de utilizar o ARIMA

tradicional, optamos pelo ARIMAX.

O ARIMAX, ou ARIMA com variáveis exógenas, é uma extensão do ARIMA que incorpora informações de variáveis externas, além da série temporal principal, para melhorar a precisão das previsões. Nesse contexto, a inclusão de mais de um dado de nível como variável exógena permite ao modelo considerar múltiplos fatores simultaneamente durante o processo de previsão.

Ao empregar o ARIMAX, capacitamos o modelo a capturar relações mais complexas entre os diferentes previsores de nível, proporcionando uma representação mais rica e abrangente dos padrões hidrológicos. Isso é particularmente interessante quando lidamos com sistemas complexos de recursos hídricos, nos quais a interação de múltiplos fatores pode influenciar significativamente os níveis de água.

Essa abordagem avançada fortalece a capacidade do modelo em lidar com as variações dos dados hidrológicos, fornecendo previsões mais refinadas e adaptáveis. Ao integrar o ARIMAX na estrutura, reconhecemos a importância de considerar múltiplos previsores de nível para obter uma visão mais completa e precisa do comportamento das séries temporais hidrológicas. Essa modificação representa um passo crucial em direção a modelos mais sofisticados e eficazes na gestão preditiva dos recursos hídricos.

Na evolução para essa nova arquitetura, introduzimos um segundo passo crucial que envolve as variáveis climatológicas. Em contraste com a arquitetura anterior, optamos por adotar uma abordagem mais flexível nesta fase, permitindo a utilização de qualquer modelo de aprendizado de máquina para lidar com essas variáveis. A única constante mantida é o modelo ARIMAX do primeiro passo, garantindo a coesão e o aproveitamento da informação temporal.

Essa flexibilidade na escolha do modelo de aprendizado de máquina para as variáveis climatológicas abre um leque de possibilidades, permitindo a seleção do algoritmo mais adequado para a natureza específica desses dados. Diferentemente da abordagem anterior, onde a escolha do modelo estava mais rigidamente definida, agora pode-se explorar e experimentar com uma gama diversificada de técnicas de aprendizagem de máquina.

Essa abertura na escolha do modelo é particularmente benéfica quando lidamos com variáveis climatológicas, que podem apresentar padrões complexos e não lineares. Diferentes algoritmos de aprendizado de máquina podem ser mais eficazes em capturar essas complexidades e relações não triviais nos dados climatológicos, contribuindo para previsões mais precisas e robustas.

Ao incorporar essa flexibilidade na arquitetura, visamos otimizar a capacidade do modelo de se adaptar a diferentes características dos dados climatológicos, aprimorando assim a sua eficácia geral na modelagem integrada do sistema hidrológico. Essa estratégia reforça a natureza adaptável e dinâmica da abordagem, promovendo uma

modelagem mais personalizada e eficiente para as condições específicas do estudo.

No terceiro passo desta arquitetura aprimorada, damos continuidade ao processo de integração dos resultados obtidos nos passos anteriores. Aqui, é criado um novo conjunto de dados que serve como base para as etapas subsequentes do modelo de previsão. Esse conjunto de dados é composto por três elementos-chave: as previsões provenientes do modelo ARIMAX do primeiro passo, as previsões geradas pelo modelo de aprendizado de máquina no segundo passo e a média desses valores.

Essa abordagem traz uma dimensão adicional à modelagem, permitindo que múltiplos conjuntos de previsões contribuam para a construção do novo conjunto de dados. A inclusão da média dessas previsões tem implicações interessantes, proporcionando uma visão mais equilibrada e combinada das tendências identificadas por ambos os modelos.

Ao adotar essa estratégia de combinação de previsões, buscamos tirar proveito das forças complementares dos dois modelos. Enquanto o ARIMAX pode ser especialmente habilidoso em capturar padrões temporais e sazonais, os modelos de aprendizado de máquina podem colaborar na identificação de complexidades e relações não lineares nos dados climatológicos. A diversidade de abordagens contribui para a robustez do modelo final, abrindo caminho para previsões mais confiáveis e adaptáveis à dinâmica complexa do sistema hidrológico em questão.

4.3.2 Algorítimo e Análise

Nesta seção serão abordados detalhes da implementação realizada no modelo híbrido proposto. Essa nova arquitetura possui basicamente três partes, sendo a primeira referindo-se aos aspectos relacionados ao modelo ARIMAX, na segunda etapa temos o primeiro modelo de aprendizagem de máquina. Como última etapa temos o segundo modelo, neste caso é adicionado a média como atributo previsor.

Na análise do algoritmo da segunda arquitetura, conforme a Figura2, o processo se inicia com uma avaliação da série temporal em questão. Uma etapa crucial é a verificação da autocorrelação total, que busca identificar padrões de dependência temporal entre observações consecutivas. Esta análise permite visualizar a presença de correlações significativas em diferentes defasagens, orientando a seleção de modelos temporais adequados.

Em seguida, o procedimento incorpora a diferenciação da série temporal, uma técnica essencial para estacionarizar a série e torná-la mais adequada para modelagem. A diferenciação visa remover tendências e padrões sazonais, promovendo a estabilidade dos dados. Essa etapa é fundamental para assegurar que o modelo possa capturar efetivamente as variações e comportamentos temporais da série(Box et al., 2015).

Outro aspecto considerado é a função de Critério de Informação *Akaike (AIC)*.

Algoritmo 2: Algoritmo II

Entrada: baseh, base hidrológica basee, base climatológica
Saída: predição do nível de água

```
// Modelo ARIMAX
act = ACF(base-h);
d = diff(x1,y1);// operação diferencial
para p=0 até pmax faça
    para d=0 até dmax faça
        para q=0 até qmax faça
            m = ARIMAX(base-h,(p,d,q));
            AIC = aic(m)
        fim
    fim
fim
para n=0 até nmax faça
    x1 = baseh[:,0:3];// previsores de nível
    y1 = baseh[:,9];// variável alvo
    yf = y1.shift[-n];
    m1 = ARIMAX(x1,yf(p,d,q));// Determinação do Modelo Arimax
    arm=m1.predict(test);// Predição do Modelo Arimax
fim
// Modelo I
para n=0 até nmax faça
    x2 = baseh[:,4:9];
    y2 = baseh[:,9];
    yf = y2.shift[-n];
    m2 = M2(x2,yf);
    mdl2 = m2.predict(test);
fim
// Modelo II
av = avg(arm,mdl2);// Obtém a média
basef = arm + mdl2 + avg+ y2;
x3 = basef[:,0:2];// previsores
y3 = basef[:,2];// variável alvo
m3 = M3(x3,y3);
mdlf = m3.predict(test);
// Processo Reverso da normalização e exibição da predição do nível
   de água
```

Essa função desempenha um papel vital na escolha do modelo mais apropriado, equilibrando a complexidade do modelo com a qualidade das previsões. O AIC penaliza modelos mais complexos, favorecendo aqueles que oferecem boas previsões com um número menor de parâmetros. Essa abordagem ajuda a evitar o *overfitting*, contribuindo para a generalização e robustez do modelo (Box et al., 2015).

Portanto, esse conjunto de análises iniciais fornece os fundamentos para a seleção e ajuste adequado do modelo ARIMAX no primeiro passo da arquitetura 2. A combinação dessas técnicas visa criar um modelo que não apenas se ajusta aos dados disponíveis, mas também é capaz de fazer previsões precisas e adaptáveis às complexidades inerentes à série temporal hidrológica em consideração.

A implementação de um *loop* para fornecer resultados preditivos consecutivos em uma única execução representa uma evolução significativa em relação à abordagem anterior do modelo. Essa modificação estrutural visa aprimorar a eficiência operacional, permitindo a geração de previsões consecutivas sem a necessidade de múltiplas execuções separadas.

Além disso, destaca-se a inserção do *lag* na variável alvo durante esse processo. A inclusão do *lag*, que representa a defasagem temporal na variável de resposta, é uma estratégia importante em problemas de séries temporais. Essa abordagem permite ao modelo capturar influências temporais passadas na variável alvo, aprimorando assim a capacidade de antecipar padrões e tendências ao longo do tempo.

Essa combinação de um *loop* eficiente e a consideração do *lag* na variável alvo confere ao modelo uma maior capacidade de adaptação às complexidades temporais inerentes aos dados hidrológicos. A capacidade de fornecer previsões consecutivas em uma única execução e a incorporação do *lag* contribuem para a robustez do modelo, tornando-o mais apto a lidar com a natureza dinâmica e fluida das séries temporais hidrológicas. Essa abordagem otimizada é fundamental para uma modelagem mais precisa e eficiente dos padrões hidrológicos.

A próxima fase do algoritmo marca a transição para a utilização do modelo de aprendizagem de máquina, agora adaptado especificamente para lidar com séries temporais e considerando o mesmo conceito de *lag* utilizado na primeira parte do algoritmo. Um ponto a ser considerado nesta etapa é a utilização exclusiva dos dados climatológicos, ao contrário da etapa anterior que incorporou os dados de nível de água para o modelo ARIMAX.

Essa abordagem focada nos dados climatológicos destaca a flexibilidade do algoritmo, que se ajusta de maneira específica às características e influências das variáveis em cada fase do processo de modelagem. Ao direcionar a atenção apenas para os dados climatológicos nesta etapa, o modelo de aprendizagem de máquina busca identificar padrões, tendências e relações específicas nesses dados que contribuam para previsões mais precisas da variável alvo, considerando o *lag* temporal.

Portanto, essa segmentação do processo, focando exclusivamente nos dados climatológicos e utilizando um modelo de aprendizagem de máquina adaptado para séries temporais com o *lag*, destaca a capacidade do algoritmo em se adaptar à natureza multifacetada dos dados hidrológicos, garantindo uma modelagem mais especializada e precisa em diferentes contextos.

Na última etapa deste algoritmo aprimorado, ocorre a inserção da média preditiva dos dois modelos anteriores, juntamente com os resultados preditivos de cada modelo, como variáveis de entrada. Uma distinção crucial em relação ao modelo anterior é que, agora, todos os modelos (etapas) são executados no conjunto de teste. Isso representa uma mudança significativa na estratégia de validação, pois cada fase contribui diretamente para as previsões finais no conjunto de teste.

Ao incorporar a média preditiva dos modelos anteriores, o algoritmo busca agregar as perspectivas distintas oferecidas pelo ARIMAX e pelo modelo de aprendizado de máquina, promovendo uma abordagem integrada e equilibrada. Essa combinação busca capturar a complementaridade das informações fornecidas por ambos os modelos, proporcionando uma visão mais abrangente e robusta para a predição final.

A decisão de executar todos os modelos no conjunto de teste em cada etapa realça a importância de validar cada componente do algoritmo individualmente. Isso contribui para uma avaliação mais precisa do desempenho de cada modelo em condições realistas e em cenários específicos de teste. Essa abordagem de validação contínua reforça a confiabilidade e a adaptabilidade do algoritmo diante de diferentes conjuntos de dados e condições hidrológicas.

Nos modelos híbridos, a última etapa envolve comumente a combinação das previsões dos modelos anteriores, utilizando geralmente a média dessas previsões. Essa abordagem visa aproveitar as vantagens individuais de cada modelo para produzir uma previsão final mais confiável e precisa.

Implementamos também uma terceira camada onde a média das previsões é tratada como um atributo independente. Isso nos permite verificar a eficácia do modelo combinado em um conjunto de dados separado, o que ajuda a evitar problemas de superajuste *overfitting* e garantir que o modelo seja capaz de generalizar bem para novos dados.

No entanto, quando consideramos a aplicação prática do modelo em um ambiente de aprendizado de máquina em tempo real, surge uma questão crucial: a média das previsões se torna um atributo dependente. Isso ocorre porque, ao utilizar o modelo para fazer previsões em um cenário de produção, não temos acesso aos resultados reais ou *y*.

Portanto, é importante reconhecer essa mudança na natureza do atributo, média ao implementar o modelo em um ambiente de produção. Sem a disponibilidade de dados futuros, precisamos adotar abordagens alternativas para lidar com essa limitação e

garantir que o modelo continue a produzir previsões úteis e confiáveis.

Em resumo, a inclusão da média preditiva e a mudança na estratégia de validação para executar todos os modelos no conjunto de teste representam a evolução do algoritmo para uma abordagem mais aberta e dinâmica. Essa adaptação visa otimizar a precisão das previsões, incorporando as vantagens individuais de cada modelo e promovendo uma modelagem hidrológica mais eficaz e versátil.

4.3.3 Otimização e Resultados

Nesta seção, exploraremos estratégias e técnicas fundamentais para otimizar os resultados dos modelos. A otimização desempenha um papel crucial no desenvolvimento de modelos de aprendizagem de máquina, visando aprimorar seu desempenho e capacidade de generalização. Abordaremos temas como ajuste de hiperparâmetros, escolha adequada de algoritmos, tratamento de dados desbalanceados e aplicação de técnicas para lidar com desafios específicos. Nesta seção, exploraremos estratégias e técnicas fundamentais para otimizar os resultados dos modelos. A otimização desempenha um papel crucial no desenvolvimento de modelos de aprendizagem de máquina, visando aprimorar seu desempenho e capacidade de generalização.

4.3.3.1 Otimização

A sintonia de hiperparâmetros desempenha um papel fundamental na construção de modelos de aprendizado de máquina, sendo uma etapa crítica para otimizar o desempenho e a generalização dos modelos. Hiperparâmetros são configurações ajustáveis que não são aprendidas pelo modelo durante o treinamento, e sua escolha adequada pode significar a diferença entre um modelo eficaz e um sub ótimo (Hastie et al., 2009).

A importância de ajustar hiperparâmetros reside na capacidade de encontrar a combinação ideal que melhor se adapta aos dados específicos do problema em questão. Isso é particularmente relevante em um contexto onde diferentes conjuntos de dados e complexidades de problemas demandam abordagens distintas. Tunar hiperparâmetros visa equilibrar o modelo entre a capacidade de aprender com eficácia os padrões presentes nos dados e a capacidade de generalizar para novas instâncias.

Em nossa metodologia, empregamos o *Random Search*, uma técnica que explora de forma aleatória o espaço de hiperparâmetros, permitindo uma busca mais abrangente por configurações ótimas. Além disso, complementamos esse processo com o teste manual de hiperparâmetros, uma abordagem que incorpora o conhecimento especializado para refinar e ajustar os parâmetros com base na compreensão do domínio.

Juntos, o *Random Search* e o teste manual de hiperparâmetros proporcionam uma abordagem integral para sintonizar modelos, maximizando seu desempenho em uma

variedade de cenários. Essa combinação estratégica busca otimizar não apenas a precisão do modelo, mas também sua capacidade de adaptação a diferentes contextos, garantindo soluções mais robustas e eficazes.

Tabela 4 – Seleção dos Hiperparâmetros dos Modelos

Seleção dos Hiperparâmetros		
Dataset	**Modelo**	**Hiperparâmetros**
DSI	RF	n_tree=100,max_depth=32
	SVR	kernel=rbf,C=10,gamma=0.01
	ANN	opt=rmsprop,act=relu,neu=32,batch=10
	RNN	opt=rmsprop,act=relu,neu=32,batch=16
	Arima-RF	Arima(1,1,1),RF n_tree=200,max_depth=32
	Arima-SVR	Arima(1,1,1),SVR kernel=rbf,C=10,gamma=0.01
	Arima-ANN	Arima(1,1,1),ANN opt=adam,act=relu,neu=64,batch=32
	Arima-RNN	Arima(1,1,1),ANN opt=adam,act=relu,neu=16,batch=32
DSII	RF	n_tree=200,max_depth=8
	SVR	kernel=rbf,C=10,gamma=0.01
	ANN	opt=adam,act=relu,neu=64,batch=32
	RNN	opt=adam,act=relu,neu=128,batch=16
	Arima-RF	Arima(1,1,1),RF n_tree=200,max_depth=32
	Arima-SVR	Arima(1,1,1),SVR kernel=rbf,C=10,gamma=0.01
	Arima-ANN	Arima(1,1,1),ANN opt=adam,act=relu,neu=64,batch=64
	Arima-RNN	Arima(1,1,1),ANN opt=adam,act=relu,neu=16,batch=32
DSIII	RF	n_tree=200,max_depth=32
	SVR	kernel=rbf,C=10,gamma=0.01
	ANN	opt=rmsprop,act=relu,neu=32,batch=10
	RNN	opt=rmsprop,act=relu,neu=32,batch=16
	Arima-RF	Arima(1,1,1),RF n_tree=200,max_depth=6
	Arima-SVR	Arima(1,1,1),SVR kernel=rbf,C=100,gamma=0.1
	Arima-ANN	Arima(1,1,1),ANN opt=adam,act=relu,neu=64,batch=32
	Arima-RNN	Arima(1,1,1),ANN opt=adam,act=relu,neu=64,batch=16
DSIV	RF	n_tree=100,max_depth=8
	SVR	kernel=rbf,C=10,gamma=0.01
	ANN	opt=rmsprop,act=relu,neu=32,batch=10
	RNN	opt=rmsprop,act=relu,neu=32,batch=16
	Arima-RF	Arima(2,1,2),RF n_tree=200,max_depth=32
	Arima-SVR	Arima(2,1,2),SVR kernel=relu,C=1,gamma=0.001
	Arima-ANN	Arima(2,1,2),ANN opt=adam,act=relu,neu=32,batch=32
	Arima-RNN	Arima(2,1,2),ANN opt=adam,act=relu,neu=128,batch=16

Ao analisar a Tabela 4, podemos identificar os melhores *hiperparâmetros* para otimizar o desempenho dos modelos em nosso conjunto de dados. Em particular, destacam-se algumas descobertas interessantes, especialmente no que diz respeito ao *Random Forest*, Modelos de Redes Neurais Artificiais (ANN) e Redes Neurais Recorrentes (RNN).

No contexto do *Random Forest*, observamos que o tamanho do conjunto de dados desempenha um papel crucial. Consideravelmente, à medida que aumentamos o tamanho do conjunto de dados, o número de árvores no modelo *Random Forest* tende a crescer de maneira exponencial. Essa observação sugere que, para otimizar o desempenho do *Random Forest*, é vital considerar a relação entre o tamanho do conjunto de dados e o número de árvores, ajustando-os de maneira equilibrada.

No que diz respeito aos Modelos de Redes Neurais Artificiais e Redes Neurais Recorrentes, uma descoberta interessante foi feita durante os testes de diferentes

ativadores. Em todas as situações, o ativador Relu demonstrou consistentemente um desempenho superior. Esse resultado é particularmente significativo quando estamos lidando com problemas de regressão.

A utilização da função de ativação ReLU *Rectified Linear Unit* em problemas de regressão não lineares, como os relacionados a dados climáticos, oferece vários benefícios significativos. A natureza não linear e complexa dos padrões climáticos demanda uma capacidade robusta de modelagem por parte da rede neural. A função ReLU, ao introduzir não linearidades nos neurônios, permite que a rede aprenda e represente relações mais complexas entre as variáveis climáticas.

Ao explorar a eficácia do modelo ARIMA em diferentes conjuntos de dados, identificamos padrões relacionados aos hiperparâmetros ideais, revelando a necessidade de uma abordagem adaptativa conforme a extensão do conjunto de dados.

Em conjuntos de dados mais extensos, observamos que os hiperparâmetros ideais para o modelo ARIMA foram (1,1,1). Esta combinação específica refere-se ao número de termos autorregressivos (p), à ordem de diferenciação (d) e ao número de termos de média móvel (q), respectivamente. A escolha desses hiperparâmetros está profundamente enraizada na capacidade do ARIMA em capturar padrões temporais e sazonalidades. O termo autoregressivo (p) destaca a dependência das observações passadas, enquanto o termo de média móvel (q) lida com os resíduos das observações anteriores.

Em conjuntos de dados menores, os hiperparâmetros ideais se mostraram diferentes, especificamente (2,1,2). Essa variação pode ser atribuída à natureza mais limitada do conjunto de dados, onde os padrões temporais podem ser mais complexos e exigir um ajuste mais sensível para uma modelagem precisa.

A inclusão de um termo adicional de média móvel (q) no modelo ARIMA para conjuntos de dados menores pode ser justificada pela necessidade de capturar percepções mais sutis nas flutuações dos dados. Isso destaca a importância de adaptar os hiperparâmetros às características específicas do conjunto de dados em questão, demonstrando que não há uma abordagem única e fixa para todos os cenários.

Além disso, observou-se um aumento na ordem do termo de autoregressão (p) de 1 para 2 para este conjunto de dados menor. Esse aumento sugere que o modelo ARIMA precisa considerar mais observações anteriores para capturar adequadamente os padrões temporais presentes nos dados. Esta mudança indica uma maior complexidade nos padrões de autocorrelação dos dados menores, ressaltando a importância de uma análise cuidadosa e adaptação dos hiperparâmetros para garantir a precisão das previsões

Em suma, a escolha dos hiperparâmetros ideais para o modelo ARIMA é um fator que combina compreensão do comportamento temporal dos dados, experiência prática e ajuste sensível às particularidades do conjunto de dados. Essa flexibilidade

permite que o ARIMA se destaque em uma variedade de contextos, proporcionando previsões precisas e percepções interessantes, independentemente do tamanho do conjunto de dados em análise.

Ao explorar diferentes otimizadores em modelos de Inteligência Artificial dedicados a problemas de regressão, destaca-se o otimizador Adam como a escolha que consistentemente ofereceu o melhor desempenho preditivo.

O algoritmo Adam, que combina técnicas de *momentum* e *RMSprop*, tem se destacado em problemas de otimização pela sua capacidade de se adaptar dinamicamente às taxas de aprendizado. No contexto de modelos de regressão, onde a precisão na previsão é crucial, a flexibilidade do Adam em ajustar automaticamente as taxas de aprendizado tem sido uma vantagem significativa.

A habilidade do Adam em lidar eficientemente com conjuntos de dados complexos, ajustando as taxas de aprendizado de forma adaptativa para diferentes parâmetros do modelo, contribui diretamente para um desempenho preditivo superior. Essa característica é particularmente relevante em problemas de regressão, onde a relação entre variáveis pode ser não linear e altamente dinâmica, neste caso nas variáveis climatológicas.

Além disso, a implementação eficaz do Adam na gestão de gradientes estocásticos torna-o robusto em lidar com grandes conjuntos de dados, onde a eficiência computacional é crucial. Sua capacidade de lidar com gradientes ruidosos e escassez de dados, com um mecanismo adaptativo de atualização de pesos, confere ao Adam uma vantagem notável.

Dessa forma, o otimizador Adam emerge como uma escolha ideal para impulsionar o desempenho preditivo em modelos de Inteligência Artificial voltados para problemas de regressão, destacando-se pela sua versatilidade, eficiência computacional e capacidade de adaptação dinâmica aos desafios inerentes à complexidade dos dados.

Essas observações destacam a importância de uma abordagem criteriosa ao ajustar hiperparâmetros, levando em consideração as características específicas do conjunto de dados e a natureza do problema em questão. Compreender como diferentes hiperparâmetros afetam o desempenho do modelo é crucial para desenvolver modelos mais eficazes e precisos, resultando em melhores resultados em tarefas de previsão e análise de dados.

Outro assunto que destacamos aqui é sobre a alocação apropriada de dados nos conjuntos de treinamento, validação e teste que desempenha um papel crucial no desenvolvimento de modelos de *machine learning* robustos. Ao observar as divisões utilizadas em diferentes conjuntos de dados, notamos uma adaptação significativa nas proporções, refletindo uma abordagem consciente para lidar com as características específicas de cada conjunto.

Nos conjuntos de dados maiores, optamos por uma divisão de 75% para treina-

mento, 15% para validação e 10% para teste. Essa distribuição equilibrada é fundamentada na premissa de que, em conjuntos de dados extensos, ter uma porção substancial, mas não tão alta para o treinamento permite ao modelo aprender padrões complexos e intrínsecos. A validação desempenha o papel de ajustar hiperparâmetros e evitar *overfitting*, enquanto o teste avalia o desempenho geral em dados não vistos.

No entanto, ao lidar com um conjunto de dados consideravelmente reduzido, ajustamos a divisão para 85% para treinamento, 5% para validação e 10% para teste. Essa mudança reflete a necessidade de maximizar a quantidade de dados disponíveis para treinamento, considerando a limitação da amostra. Em conjuntos pequenos, é crucial proporcionar ao modelo uma exposição mais extensa aos dados de treinamento para extrair padrões significativos.

A alocação maior para treinamento no conjunto de dados reduzido visa mitigar o risco de *overfitting*, uma vez que a capacidade do modelo de aprender com poucos exemplos pode ser limitada. A parcela menor para validação ainda permite ajustar hiperparâmetros com base em um conjunto representativo, enquanto o conjunto de teste permanece uma ferramenta vital para avaliar a generalização do modelo.

Em resumo, as variações nas proporções de treinamento, validação e teste são estratégias adaptativas para otimizar o desempenho do modelo, levando em consideração a quantidade e complexidade dos dados disponíveis em cada conjunto. Essa abordagem flexível reflete uma prática sólida, garantindo uma modelagem eficaz em diferentes contextos e cenários.

Na Tabela 5, pode-se observar que ao examinar os tempos de treino e ajuste registrados em uma tabela abrangendo os quatro conjuntos de dados distintos, notamos uma disparidade interessante, particularmente nos modelos que envolvem Redes Neurais, quando comparados aos outros modelos.

Os tempos de ajuste mencionados na tabela referem-se especificamente ao período gasto pelo algoritmo *Random Search* para explorar diferentes combinações de hiperparâmetros. É crucial destacar que esses tempos não incorporam o ajuste manual, uma prática que, sem dúvida, demanda um período ainda mais extenso.

A análise inicial da tabela revela que os modelos de Redes Neurais, impulsionados por técnicas como o *Random Search*, apresentam tempos de ajuste consideravelmente maiores em comparação com métodos convencionais. Essa disparidade pode ser justificada por várias razões. Primeiramente, a complexidade inerente aos modelos das Redes Neurais, especialmente aqueles baseados em redes neurais profundas ou algoritmos complexos, contribui para uma estrutura mais complexa, demandando assim mais tempo para o ajuste de hiperparâmetros.

Além disso, a aplicação frequente de modelos de RNs em conjuntos de dados extensos, caracterizados por uma quantidade substancial de amostras e características, aumentando a complexidade do espaço de busca de hiperparâmetros. O aumento no

Tabela 5 – Tempo de Treinamento e de Ajustes

	Tempo de Treinamento		
DataSet	**Modelo**	**T.Treino**	**T.Ajuste**
DSI	RF	22s	2h:30min
	SVR	25s	3hs:12min
	ANN	1h:31min	12hs:21min
	RNN	1h:50min	14:hs:03min
	Arima-RF	15min	3hs:02min
	Arima-SVR	16min	3hs:50min
	Arima-ANN	1h:42min	13hs:34min
	Arima-RNN	1h:56min	14hs:51min
DSII	RF	30s	2h:57min
	SVR	28s	3hs:49min
	ANN	1h:46min	13hs:35min
	RNN	2hs:15min	14:hs:55min
	Arima-RF	21min	3hs:40min
	Arima-SVR	23min	3hs:55min
	Arima-ANN	1h:54min	14hs:08min
	Arima-RNN	2hs:20min	15hs:02min
DSIII	RF	24s	2h:45min
	SVR	26s	3hs:32min
	ANN	1h:39min	13hs:15min
	RNN	2hs:05min	14:hs:22min
	Arima-RF	18min	3hs:31min
	Arima-SVR	20min	3hs:42min
	Arima-ANN	1h:48min	14hs:28min
	Arima-RNN	2hs:15min	15hs:51min
DSIV	RF	15s	25min
	SVR	18s	29min
	ANN	52min	2hs:22min
	RNN	1hs:02min	2:hs:41min
	Arima-RF	5min	51min
	Arima-SVR	8min	1hs:02min
	Arima-ANN	1h:08min	2hs:35min
	Arima-RNN	2hs:17min	2hs:53min

tamanho do conjunto de dados implica em mais iterações necessárias para explorar de maneira abrangente esse espaço, prolongando o tempo de ajuste.

Finalmente, a estratégia de busca aleatória adotada pelo *Random Search*, embora eficaz na descoberta de boas combinações de hiperparâmetros, contribui para um tempo de ajuste mais extenso. A exploração aleatória de uma ampla gama de configurações possíveis pode exigir mais iterações para convergir para as configurações ideais, prolongando assim o processo de ajuste. Essas considerações destacam a complexidade e os desafios inerentes à otimização de modelos de IA, justificando os tempos de ajuste mais significativos observados.

A observação detalhada da Tabela 5 ressalta um padrão interessante relacionado à redução no número de atributos preditores e ao tamanho do conjunto de dados. Nota-se uma correspondência proporcional entre a diminuição desses fatores e uma redução significativa nos tempos de ajuste e treinamento dos modelos.

Quando o número de atributos preditores é reduzido, como evidenciado em conjuntos de dados com menos características, os modelos de forma geral demonstram um tempo de ajuste menor. Essa tendência pode ser atribuída à diminuição da complexidade do espaço de busca de hiperparâmetros, uma vez que há menos dimensões a serem exploradas. A redução na quantidade de atributos preditores simplifica o processo de ajuste, permitindo que o *Random Search* alcance configurações ótimas de forma mais eficiente.

Além disso, a análise revela que a redução no tamanho do conjunto de dados, como observado no *DataSet IV*, também está associada a um tempo de ajuste e treinamento inferior. Com menos amostras disponíveis para treinamento, o modelo é exposto a um conjunto de dados menos extenso, o que implica em menos iterações durante o ajuste de hiperparâmetros. Esse fenômeno destaca a influência direta do tamanho do conjunto de dados na eficiência do processo de ajuste, mostrando que conjuntos menores podem resultar em tempos de treinamento mais ágeis.

Em suma, a análise detalhada dos tempos de ajuste e treinamento em relação ao número de atributos preditores e ao tamanho do conjunto de dados destaca a natureza dinâmica do processo de ajuste. A redução em ambos os fatores contribui para uma otimização mais eficaz, evidenciando a importância de considerar a complexidade do conjunto de dados ao escolher estratégias de otimização de hiperparâmetros. Essas observações são cruciais para a compreensão e implementação eficaz de modelos de Inteligência Artificial em diferentes contextos e condições de dados.

4.3.4 Resultados

Nesta seção, apresentaremos os resultados obtidos a partir da aplicação de modelos em nossos conjuntos de dados. Ao avaliar a desempenho dos modelos, utilizamos diversas métricas essenciais para medir a precisão e robustez de nossas previsões.

As métricas destacadas incluem o Erro Quadrático Médio (RMSE), o Erro Médio Absoluto (MAE), o Coeficiente de Determinação e o Percentual Médio de Erro Absoluto (MAPE).

Como mencionado anteriormente, o RMSE é uma métrica que avalia a dispersão dos erros, proporcionando uma medida robusta da precisão geral do modelo. Já o MAE representa a média dos valores absolutos dos erros, fornecendo uma visão direta da magnitude dos desvios entre as previsões e os valores reais. O Coeficiente de Determinação, indica a proporção da variabilidade em uma variável dependente explicada pela variabilidade nas variáveis independentes em um modelo de regressão, enquanto o MAPE quantifica a precisão percentual média das previsões.

Ao utilizar essa gama de métricas, buscamos fornecer uma análise abrangente e detalhada do desempenho dos modelos em diferentes aspectos. Esses indicadores são cruciais para a validação e seleção adequada dos modelos, permitindo-nos compreender não apenas a acurácia geral, mas também a consistência e confiabilidade das previsões em diferentes cenários. A seguir, apresentaremos os resultados detalhados, proporcionando uma visão aprofundada do impacto e eficácia de cada modelo em nosso contexto específico.

A análise da Tabela 6 revela padrões distintos de desempenho preditivo entre modelos, com resultados interessantes no *DataSet I*. Destaca-se o modelo Arima-RNN, que demonstrou uma performace considerável em comparação com os modelos tradicionais de Aprendizagem de Máquina.

No contexto de previsão de um dia à frente, o Arima-RNN apresentou um RMSE interessante de 0,072 metros. Essa métrica indica uma precisão elevada na estimativa dos valores reais, sugerindo que o modelo foi capaz de capturar de maneira eficaz as variantes e padrões presentes nos dados para previsões imediatas.

O desempenho ainda mais destacado ocorreu ao estender a previsão para 45 dias à frente, onde o Arima-RNN registrou um RMSE de 0,2418 metros. Mesmo em um horizonte de previsão mais longo, o modelo manteve uma precisão significativa, indicando uma capacidade robusta de generalização e previsão em períodos mais extensos.

Para previsões de um dia à frente com um RMSE de 0,072 metros, observamos uma precisão considerável nas estimativas do modelo. O RMSE, que representa a raiz do erro médio quadrático, fornece uma medida da discrepância entre as previsões do modelo e os valores reais. Neste caso, um RMSE baixo sugere que o modelo Arima-RNN foi capaz de fazer previsões muito próximas aos valores reais para o horizonte de um dia. Isso indica uma capacidade de capturar padrões imediatos nos dados e uma boa adaptação às variações diárias.

Ao estender a previsão para 45 dias à frente, observamos um aumento no RMSE para 0,2418 metros. Esse aumento é esperado, uma vez que prever eventos mais

Tabela 6 – Performace dos modelos na Base de Teste

Modelos	Lag	DS1				DS2				DS3				DS4			
	1	RMSE	MAE	R2	MAPE	RMSE	MAE	R2	MAPE	RMSE	MAE	R2	MAPE	RMSE	MAE	R2	MAPE
RF		0.0962	0.0654	0.9513	10.08%	0.0689	0.0332	0.9466	5.39%	0.1019	0.0670	0.9410	10.15%	0.1063	0.0871	0.8863	12.75%
SVR		0.1424	0.1134	0.9142	11.50%	0.0943	0.0698	0.9212	10.17%	0.1274	0.0999	0.9216	13.62%	0.1367	0.1079	0.8128	20.10%
ANN		0.0988	0.0673	0.9496	10.97%	0.0769	0.0482	0.9421	7.74%	0.1004	0.0680	0.9438	6.98%	0.1580	0.1176	0.7780	16.67%
RNN		0.0943	0.0637	0.9532	9.42%	0.0666	0.0358	0.9505	5.28%	0.0982	0.0661	0.9460	10.09%	0.1254	0.0963	0.8559	14.09%
ARIMA-RF		0.0910	0.0631	0.9397	10.05%	0.0678	0.0328	0.9468	5.37%	0.0876	0.0627	0.9631	8.97%	0.1077	0.0879	0.8854	13.99%
ARIMA-SVR		0.1037	0.0754	0.9736	11.02%	0.0698	0.0339	0.9724	5.40%	0.0977	0.0680	0.9402	10.25%	0.1365	0.1066	0.8122	14.07%
ARIMA-ANN		0.0837	0.0570	0.9474	9.10%	0.0685	0.0363	0.9737	5.38%	**0.0249**	0.0169	0.9450	6.20%	**0.0721**	0.0646	0.9853	7.98%
ARIMA-RNN		**0.0723**	0.0377	0.9603	8.53%	**0.0664**	0.0356	0.9506	5.26%	0.0980	0.0665	0.9430	9.98%	0.1021	0.0924	0.8774	12.66%
	5																
RF		0.1571	0.1096	0.8750	13.56%	0.1056	0.0528	0.8752	6.89%	0.1458	0.1023	0.8804	18.88%	0.2290	0.1740	0.5223	26.90%
SVR		0.1928	0.1511	0.8402	19.98%	0.1347	0.1018	0.8398	12.33%	0.1845	0.1463	0.8414	24.67%	0.2433	0.2004	0.3781	29.56%
ANN		0.1657	0.1114	0.8622	15.54%	0.1100	0.0556	0.8646	10.21%	0.1538	0.1056	0.8674	21.90%	0.2528	0.1973	0.3738	30.79%
RNN		0.1598	0.1116	0.8730	14.05%	0.1066	0.0565	0.8730	7.06%	0.1475	0.1049	0.8794	19.83%	0.2410	0.1942	0.3846	30.38%
ARIMA-RF		0.1569	0.1094	0.8752	13.54%	**0.1054**	0.0526	0.8755	6.88%	**0.1438**	0.1052	0.8896	13.20%	0.2157	0.1594	0.3344	25.10%
ARIMA-SVR		0.1263	0.0961	0.9536	10.22%	0.1334	0.0666	0.8924	12.28%	0.1465	0.1079	0.8690	17.74%	0.2108	0.1372	0.3354	25.05%
ARIMA-ANN		0.1562	0.1077	0.8758	13.52%	0.1098	0.0559	0.8649	10.18%	0.1470	0.1041	0.8744	17.98%	0.1810	0.1653	0.4095	24.97%
ARIMA-RNN		**0.1168**	0.0614	0.8899	9,12%	0.1065	0.0554	0.8736	7.04%	0.1440	0.1039	0.8798	13.30%	**0.1447**	0.1434	0.9586	15.98%
	10																
RF		0.1844	0.1318	0.8297	18.03%	0.1320	0.0665	0.8061	8,74%	0.1810	0.1318	0.8176	22.62%	0.2062	0.1494	0.6153	25.24%
SVR		0.2178	0.1725	0.8017	17.83%	0.1585	0.1190	0.7763	14.64%	0.2150	0.1709	0.7783	31.55%	0.2201	0.1640	0.5809	26.58%
ANN		0.1935	0.1343	0.8124	18.40%	0.1368	0.0694	0.7921	9.77%	0.1944	0.1347	0.7732	27.80%	0.2221	0.1701	0.5358	28.65%
RNN		0.1957	0.1331	0.8140	18.90%	0.1401	0.0785	0.7850	9.24%	0.1929	0.1319	0.7970	25.51%	0.2277	0.1690	0.5428	30.16%
ARIMA-RF		0.1802	0.1352	0.7436	18.01%	**0.1315**	0.0661	0.8055	8.72%	**0.1808**	0.1316	0.8234	22.61%	0.1946	0.1486	0.6155	25.10%
ARIMA-SVR		0.1886	0.1229	0.9049	18.05%	0.1366	0.0812	0.8534	9.74%	0.2023	0.1430	0.7712	28.92%	0.2198	0.1638	0.5812	26.60%
ARIMA-ANN		0.1805	0.1366	0.7485	18.03%	0.1365	0.0692	0.7919	9.76%	0.2062	0.1497	0.7614	28.97%	0.1853	0.1691	0.6165	24.99%
ARIMA-RNN		**0.1373**	0.0752	0.8486	10.90%	0.1375	0.0764	0.8486	9.78%	0.1925	0.1310	0.7982	25.48%	**0.1740**	0.1725	0.5410	25.77%
	15																
RF		0.2048	0.1519	0.7869	19.90%	**0.1467**	0.0749	0.7663	11.77%	0.2065	0.1515	0.7512	30.05%	0.1761	0.1388	0.6228	32.43%
SVR		0.2330	0.1830	0.7519	23.54%	0.1683	0.1207	0.7348	17.00%	0.2319	0.1824	0.7160	32.23%	0.1701	0.1394	0.6431	30.31%
ANN		0.2236	0.1559	0.7613	23.01%	0.1564	0.0826	0.7344	13.09%	0.2164	0.1546	0.7286	31.98%	0.1858	0.1504	0.6017	32.62%
RNN		0.2156	0.1522	0.7707	20.65%	0.1601	0.1201	0.7338	16.55%	0.2125	0.1504	0.7415	30.10%	0.1668	0.1369	0.6226	28.48%
ARIMA-RF		0.1998	0.1548	0.7180	19.50%	0.1571	0.0799	0.7416	13.10%	**0.1944**	0.1416	0.7701	28.55%	0.1704	0.1338	0.6329	30.35%
ARIMA-SVR		0.2327	0.1824	0.8791	23.50%	0.1681	0.1205	0.7346	16.98%	0.2068	0.1486	0.7377	29.87%	0.1698	0.1392	0.6438	30.29%
ARIMA-ANN		0.2044	0.1557	0.7177	20.05%	0.1562	0.0824	0.7349	13.07%	0.2162	0.1522	0.7296	31.81%	0.1856	0.1498	0.6021	32.60%
ARIMA-RNN		**0.1712**	0.0911	0.7702	17.44%	0.1594	0.0802	0.7426	15.16%	0.2122	0.1502	0.7418	30.08%	**0.1623**	0.1454	0.5590	31.22%
	30																
RF		0.2449	0.1838	0.6878	25.70%	**0.1659**	0.0880	0.6833	14.32%	0.2578	0.1917	0.6293	30.21%	0.2140	0.1295	0.5887	30.01%
SVR		0.2734	0.2129	0.6274	33.42%	0.1869	0.1195	0.6125	19.05%	0.2802	0.2169	0.5726	36.56%	0.2304	0.1732	0.4734	31.42%
ANN		0.2617	0.1969	0.6426	32.48%	0.1839	0.0975	0.6120	19.02%	0.2826	0.2038	0.5772	37.35%	0.2514	0.1721	0.4234	33.78%
RNN		0.2584	0.1908	0.6555	30.19%	0.1842	0.1008	0.6182	19.03%	0.2796	0.1994	0.5948	36.02%	0.2378	0.1611	0.4865	30.07%
ARIMA-RF		0.2342	0.1853	0.5729	24.50%	0.1655	0.0878	0.6832	14.30%	**0.2377**	0.1864	0.6713	30.12%	**0.2138**	0.1288	0.5891	29.98%
ARIMA-SVR		0.2730	0.2127	0.7739	33.40%	0.1866	0.1193	0.6122	19.03%	0.2604	0.1952	0.5720	30.71%	0.2155	0.1594	0.4154	30.00%
ARIMA-ANN		0.2300	0.1832	0.5932	24.21%	0.1705	0.0961	0.6980	18.90%	0.2557	0.1962	0.5611	30.10%	0.2353	0.1586	0.1978	30.28%
ARIMA-RNN		**0.2088**	0.1138	0.6411	19.55%	0.1840	0.1005	0.6188	10.00%	0.2517	0.1940	0.5816	29.98%	0.2376	0.1609	0.4869	30.05%
	45																
RF		0.2799	0.2069	0.5961	35.01%	**0.1988**	0.1035	0.5333	17.01%	0.2878	0.1917	0.6243	36.91%	0.1927	0.1575	0.4085	33.24%
SVR		0.3144	0.2401	0.4900	36.55%	0.2165	0.1248	0.4861	21.58%	0.3183	0.2416	0.4299	38.12%	**0.1793**	0.1459	0.4810	22.35%
ANN		0.3056	0.2249	0.5187	35.82%	0.2142	0.1170	0.4930	21.21%	0.3127	0.2297	0.4450	36.82%	0.1813	0.1412	0.4618	23.10%
RNN		0.3048	0.2178	0.5333	33.99%	0.2169	0.1149	0.4900	17.56%	0.3167	0.2260	0.4484	33.73%	0.1815	0.1357	0.4831	23.21%
ARIMA-RF		0.2764	0.2052	0.4487	34.88%	0.1991	0.1044	0.5321	17.10%	**0.2501**	0.1967	0.5425	32.22%	0.1925	0.1571	0.4082	33.22%
ARIMA-SVR		0.3030	0.2102	0.6902	33.87%	0.2163	0.1245	0.4867	21.56%	0.2907	0.2203	0.4101	36.98%	0.1824	0.1540	0.4790	27.90%
ARIMA-ANN		0.2883	0.2057	0.4848	35.50%	0.2144	0.1175	0.4925	21.26%	0.2933	0.2170	0.4204	36.99%	0.1810	0.1415	0.4620	23.08%
ARIMA-RNN		**0.2418**	0.1333	0.4953	25.60%	0.2167	0.1145	0.4910	17.54%	0.2814	0.2175	0.4309	36.88%	0.1817	0.1358	0.4756	23.25%

distantes no futuro geralmente é mais desafiador devido à incerteza acumulativa e à maior complexidade temporal. Apesar do aumento no erro médio quadrático, o RMSE de 0,2418 ainda é relativamente baixo, sugerindo que o modelo mantém uma precisão razoável mesmo em horizontes de previsão mais longos.

Em contraste, os modelos tradicionais de Aprendizagem de Máquina, especificamente o *Random Forest*, emergiram como destacados no mesmo conjunto de dados. A preferência pelo *Random Forest* pode ser justificada por sua capacidade nata de lidar com relações complexas nos dados, a capacidade de trabalhar bem com características não lineares e a habilidade de mitigar o *overfitting*.

Esses resultados sugerem que a combinação de técnicas ARIMA e RNN no modelo ARIMA-RNN apresentou uma eficácia particular na captura de padrões temporais complexos nos dados do *DataSet I*, elevando a precisão nas previsões. A abordagem híbrida pode ter permitido uma melhor adaptação a diferentes padrões, tanto de curto quanto de longo prazo, destacando-se em comparação com abordagens mais tradicionais.

A análise do segundo conjunto de dados revela uma dinâmica interessante em relação ao desempenho preditivo, consideravelmente aprimorado pela presença de uma maior quantidade de dados. A abundância de informações parece contribuir significativamente para a capacidade dos modelos de realizar previsões mais precisas de modo geral.

Neste cenário, destaca-se novamente o *Random Forest* como um modelo tradicional de Aprendizagem de Máquina, evidenciando sua robustez em lidar com conjuntos de dados mais extensos. Sua habilidade intrínseca de capturar relações complexas nos dados, bem como a resistência a *overfitting*, fazem dele uma escolha sólida, especialmente quando a quantidade de dados é substancial.

Além disso, o modelo híbrido ARIMA-RF surge como uma abordagem considerável. A combinação de técnicas ARIMA e *Random Forest* pode oferecer vantagens ao aproveitar as características complementares de ambas. O destaque desse modelo em particular pode indicar que a fusão de métodos clássicos de séries temporais com técnicas mais avançadas de aprendizagem de máquina proporciona uma abordagem eficaz para lidar com a complexidade dos dados do segundo conjunto.

Um ponto de interesse específico é a observação de que, para previsões de 45 dias à frente, o *Random Forest* apresentou um RMSE de 0,1988. Esse valor sugere uma precisão interessante nas previsões, considerando o horizonte de tempo mais amplo. No entanto, é crucial interpretar esse resultado à luz das características específicas do problema, da sensibilidade às variações nos dados e das expectativas em relação às previsões.

Uma observação relevante é que o coeficiente de determinação que parece diminuir à medida que se amplia o horizonte de previsão. Isso indica que, conforme nos

afastamos do ponto de origem da previsão, a variação nas previsões aumenta, indicando uma maior incerteza ou dificuldade em manter a precisão ao longo do tempo.

Em resumo, a análise do segundo conjunto de dados destaca a importância da quantidade de dados na melhoria do desempenho preditivo. Modelos tradicionais, como o *Random Forest*, continuam a ser eficazes, enquanto abordagens híbridas, como ARIMA-RF, emergem como estratégias promissoras. A consideração cuidadosa do horizonte de previsão, da quantidade de dados disponíveis e do comportamento do coeficiente de determinação é essencial para uma escolha adequada do modelo adaptado às necessidades específicas do problema em questão.

A análise do terceiro conjunto de dados apresenta um panorama interessante, especialmente ao considerar a redução no número de atributos previsores. Nesse cenário, observa-se novamente o destaque do modelo *Random Forest* como uma escolha robusta entre os modelos tradicionais de Aprendizagem de Máquina, demonstrando sua adaptabilidade mesmo diante de conjuntos de dados mais compactos. Além disso, o modelo híbrido ARIMA-RF permanece como uma alternativa promissora, reforçando a eficácia da combinação de métodos clássicos de séries temporais com técnicas de aprendizagem de máquina.

É relevante notar que, no terceiro conjunto de dados, todos os modelos apresentaram uma piora geral no RMSE em comparação com os conjuntos anteriores. Esse aumento na métrica de erro médio quadrático sugere que a redução no número de atributos previsores pode ter impactado a capacidade dos modelos de capturar a complexidade subjacente nos dados, resultando em previsões menos precisas.

Contrastando com essa tendência, destaca-se uma pequena melhora no Coeficiente de Determinação (R^2) para os modelos no terceiro conjunto de dados. O Coeficiente de Determinação é uma medida que indica a proporção da variabilidade nos dados de resposta explicada pelo modelo. A pequena melhora no R^2 pode ser atribuída à simplificação do conjunto de dados devido à redução de atributos. Quando há menos variáveis para considerar, o modelo pode se adaptar mais facilmente aos padrões restantes, resultando em uma explicação mais precisa da variabilidade observada.

Essa aparente contradição entre o aumento do RMSE e a pequena melhora no R^2 destaca a importância de considerar várias métricas de avaliação de desempenho ao analisar modelos. Enquanto o RMSE indica a precisão das previsões em termos absolutos, o R^2 fornece percepções sobre a capacidade do modelo de explicar a variabilidade nos dados, mesmo que a precisão geral das previsões tenha diminuído.

Em resumo, o terceiro conjunto de dados destaca a influência crucial do número de atributos previsores no desempenho dos modelos. A escolha entre complexidade e simplicidade nos conjuntos de dados deve ser cuidadosamente ponderada, considerando o equilíbrio entre precisão e interpretabilidade. A análise conjunta de métricas

como RMSE e R^2 fornece uma visão mais abrangente do desempenho dos modelos, permitindo decisões mais informadas na seleção do modelo mais apropriado para as características específicas dos dados.

No quarto conjunto de dados, observamos uma mudança significativa, caracterizada por uma redução considerável no tamanho do conjunto de dados e a inclusão de novos atributos preditivos. Essa transformação introduz novos desafios e oportunidades para os modelos de previsão, refletindo-se nos resultados obtidos.

Destaca-se, neste contexto, o bom desempenho do modelo híbrido ARIMA-RNN, sugerindo que a combinação de técnicas ARIMA e Redes Neurais Recorrentes pode ser particularmente eficaz em lidar com conjuntos de dados menores e mais complexos. Além disso, o *Random Forest* continua a se destacar como um modelo tradicional de aprendizagem de máquina, mantendo sua capacidade de adaptação mesmo em cenários desafiadores.

É interessante notar que, com a redução no tamanho do conjunto de dados, houve uma piora considerável no Coeficiente de Determinação de modo geral. O R^2, sugere que a capacidade dos modelos de explicar a variação nos dados diminuiu nesse contexto específico. Essa piora no R^2 pode ser atribuída à complexidade introduzida pelos novos atributos preditivos, exigindo dos modelos uma adaptação mais desafiadora para explicar a variabilidade presente.

Paralelamente, observamos um aumento geral no RMSE. Esse aumento indica uma piora na precisão das previsões em termos absolutos, destacando os desafios adicionais introduzidos pelos novos atributos. Essa métrica, combinada com a piora no R^2, destaca a necessidade de uma avaliação abrangente do desempenho do modelo.

Um aspecto a se considerar é a relativamente boa performace do RMSE para previsões com uma janela maior, como 45 dias à frente. Essa melhoria pode ser justificada pela inserção dos atributos de direção do vento no conjunto de dados. A inclusão desses atributos pode ter fornecido informações adicionais e relevantes para a modelagem, permitindo uma melhor captura de padrões sazonais e comportamentos de longo prazo. Essa capacidade de incorporar informações específicas pode ter contribuído para um desempenho mais robusto em previsões mais extensas.

Em resumo, o quarto conjunto de dados destaca a importância da qualidade e relevância dos atributos preditivos na modelagem preditiva. Enquanto o RMSE indica um aumento na imprecisão geral das previsões, a piora no R^2 chama a atenção para os desafios introduzidos pela complexidade dos novos atributos. A análise cuidadosa dessas métricas, juntamente com a compreensão do impacto dos novos atributos, é crucial para interpretar e selecionar o modelo mais apropriado para as características específicas do conjunto de dados em questão.

A análise dos quatro conjuntos de dados revela a complexidade inerente ao processo de modelagem preditiva, especialmente quando se trata da inserção ou exclu-

são de atributos climatológicos. Este desafio decorre da interconexão entre diferentes variáveis e da influência que atributos, aparentemente menos correlacionados com a variável dependente, podem exercer sobre outros atributos independentes.

Ao longo das análises, observamos que a inclusão de novos atributos climatológicos pode ter impactos substanciais no desempenho preditivo dos modelos. Em alguns casos, a adição de informações específicas, como a direção do vento, pode resultar em melhorias interessantes mesmo em um conjunto reduzido de informações, especialmente em previsões de longo prazo. No entanto, a inclusão indiscriminada de atributos pode também introduzir complexidade excessiva, levando a um aumento no erro nas previsões.

A correlação aparentemente fraca entre alguns atributos climatológicos e o atributo dependente pode ser enganadora, uma vez que esses atributos podem ter forte correlação com outros atributos independentes. Dessa forma, a exclusão precipitada desses elementos pode resultar na perda de informações cruciais para a modelagem preditiva.

Portanto, a escolha de atributos climatológicos para inclusão ou exclusão em modelos preditivos deve ser realizada com base em uma compreensão profunda do domínio do problema e em análises cuidadosas da relação entre variáveis. A consideração da relevância, qualidade e possível interdependência dos atributos é crucial para garantir que o modelo seja capaz de capturar de maneira precisa os padrões presentes nos dados.

Além dos desafios relacionados à complexidade dos atributos climatológicos, é crucial considerar outros fatores que permeiam a modelagem preditiva. A redução do tamanho do conjunto de dados, como observado em conjuntos menores, não apenas impacta o desempenho preditivo, mas também oferece vantagens tangíveis. A diminuição do tempo de treinamento em conjuntos menores é um fator a ser levado em conta, possibilitando uma abordagem mais eficiente na construção e ajuste dos modelos. Além disso, em conjuntos de dados mais compactos, a busca por atributos independentes torna-se mais facilitada, permitindo uma análise mais direcionada e eficaz na identificação daqueles que realmente contribuem para a capacidade preditiva do modelo. Essa consideração balanceada entre eficiência computacional e qualidade dos atributos é crucial para a construção de modelos que sejam não apenas precisos, mas também práticos e eficientes em diferentes contextos.

Em última análise, a inserção ou exclusão de atributos climatológicos representa um dilema que requer uma abordagem equilibrada e contextualizada. A análise detalhada das métricas de desempenho, como RMSE e R^2, aliada a uma compreensão aprofundada do impacto dos atributos, é essencial para orientar a escolha de atributos mais eficazes e contribuir para modelos preditivos mais precisos e generalizáveis.

Após uma análise abrangente dos resultados obtidos na base de testes, onde uti-

lizamos métricas como RMSE e MAE para avaliar o desempenho dos modelos de previsão, estamos prontos para prosseguir para a próxima etapa. Na próxima seção, será adotado uma abordagem mais específica, retirando amostras aleatórias da base de testes e analisando os resultados preditivos por meio de gráficos. Essa abordagem nos permitirá uma visualização mais detalhada do desempenho dos modelos em diferentes cenários e nos ajudará a identificar padrões ou tendências que podem não ter sido totalmente capturados na análise da base de testes completa. Ao combinar análises quantitativas e visuais, estaremos mais bem equipados para compreender a eficácia dos modelos em diferentes contextos e, assim, aprimorar ainda mais suas capacidades de previsão

4.3.4.1 Explorando o Desempenho Preditivo dos Modelos: Uma Análise Visual

Nesta seção, apresentaremos os gráficos com as previsões de todos os modelos deste estudo. Para avaliar o desempenho preditivo dos modelos, utilizamos seis amostras aleatórias do conjunto de testes. Cada conjunto de dados será representado em dois gráficos distintos: um para uma janela de um dia e outro para quarenta e cinco dias à frente.

A análise das previsões de curto prazo, com uma janela de um dia, nos permitirá examinar a capacidade dos modelos em capturar variações imediatas e pequenas flutuações nos dados. Por outro lado, a avaliação das previsões de longo prazo, para quarenta e cinco dias à frente, nos fornecerá percepções sobre a capacidade dos modelos em prever tendências e padrões de forma mais ampla no futuro.

Cada gráfico apresentará as previsões dos diferentes modelos em comparação com os valores reais do conjunto de testes. Isso nos permitirá visualizar diretamente a precisão das previsões em relação aos dados observados. Serão destacadas as tendências, desvios e qualquer outro aspecto relevante que surgir da comparação entre as previsões e os valores reais.

Ao examinar esses gráficos, poderemos ter uma compreensão abrangente do desempenho de cada modelo e identificar suas respectivas capacidades de previsão em diferentes horizontes de tempo. Essa análise nos auxiliará na seleção do modelo mais adequado para as necessidades específicas do nosso estudo, considerando tanto a precisão das previsões quanto a relevância para os objetivos de análise de dados propostos.

Os gráficos representam seis amostras selecionadas da base de teste, visando abranger o início, meio e fim do conjunto de dados. No eixo *x*, encontramos a representação das amostras, organizadas conforme a ordem em que foram coletadas. Já no eixo *y*, está indicado o nível de água em metros, oferecendo uma perspectiva visual da variação ao longo do tempo ou do espaço, dependendo do contexto da coleta dos dados.

Essas amostras foram estrategicamente escolhidas para capturar a amplitude do comportamento dos níveis de água ao longo do período de estudo ou do espaço analisado. Dessa forma, podemos observar como os níveis de água se comportam em diferentes momentos ou locais, o que pode fornecer percepções importantes para a compreensão dos padrões hidrológicos ou ambientais em estudo.

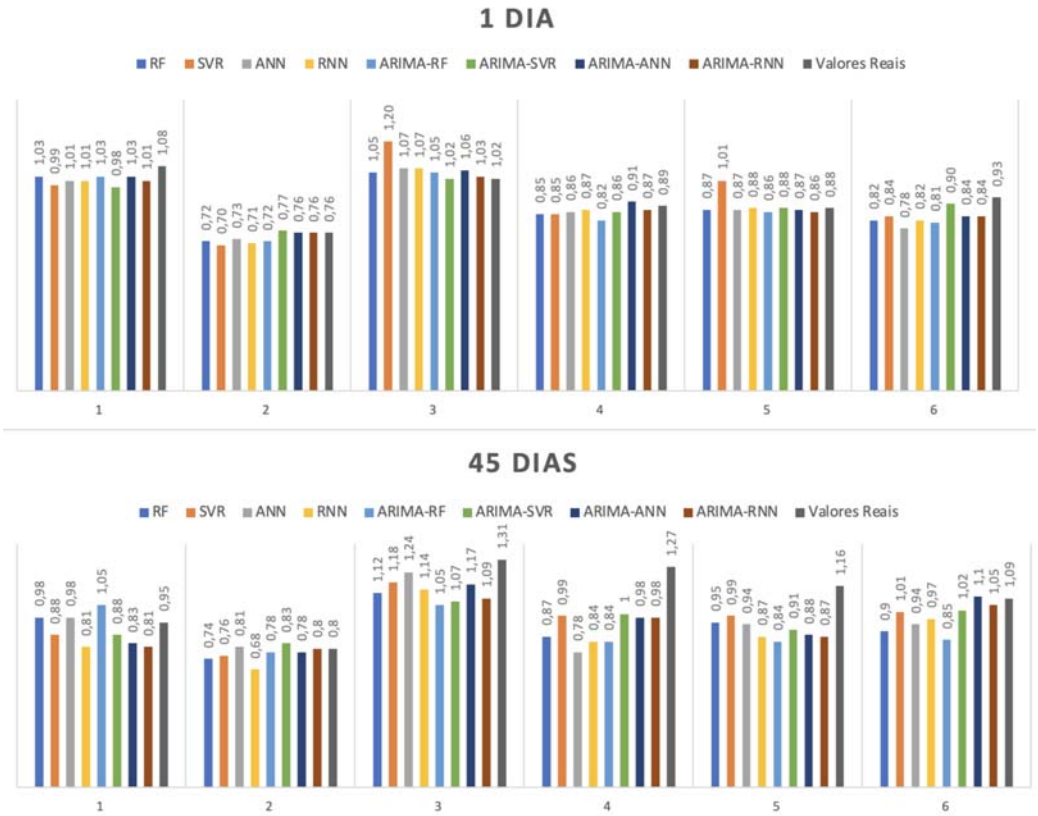

Figura 32 – Análise do Conjunto de dados I - Dados de Nível

Conforme a Figura 32 ao analisar as previsões para um dia à frente, observamos que os modelos Arima-RNN e Arima-ANN se destacaram ao apresentar uma maior proximidade com a variável dependente, enquanto o modelo SVR mostrou o pior desempenho nesse aspecto. Essa observação sugere que os modelos baseados em redes neurais, especialmente quando integrados com o ARIMA, foram mais eficazes em capturar as percepções das variações imediatas nos dados, resultando em previsões mais precisas para horizontes de curto prazo.

No entanto, ao estendermos o horizonte preditivo para os próximos 45 dias, observamos um distanciamento geral dos modelos em relação aos dados reais. Esse fenômeno é comum e se justifica pelo aumento do horizonte preditivo, que torna as previsões mais suscetíveis a erros acumulativos ao longo do tempo.

Nesse contexto, é interessante notar que os modelos SVR e Arima-SVR mostraram uma maior aproximação com a variável dependente para previsões de longo prazo. Isso sugere que, mesmo que os modelos baseados em redes neurais tenham

se destacado para previsões de curto prazo, modelos como o SVR e suas variantes podem ser mais robustos ao lidar com horizontes preditivos mais longos, possivelmente devido à sua capacidade de generalização e tratamento de padrões complexos nos dados.

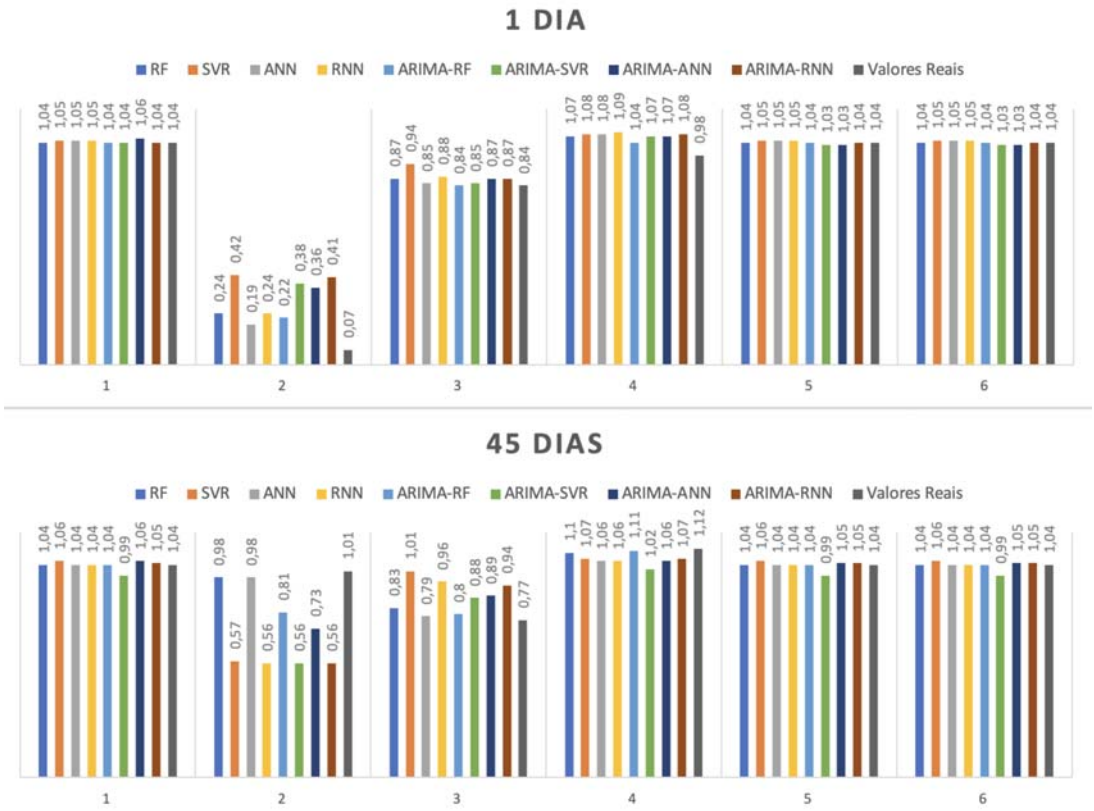

Figura 33 – Análise do Conjunto de dados II - Dados de Nível

Analisando a Figura 33 com base nas seis amostras aleatórias, examinamos as previsões para horizontes preditivos de um e quarenta e cindo dias. Para uma previsão de um dia, observamos que o modelo Arima-RF se destacou, seguido de perto pelo Arima-SVR e pelo Arima-RNN. Essa observação sugere que a combinação do ARIMA com a técnica de Random Forest (RF) foi eficaz em capturar as variações imediatas nos dados, resultando em previsões mais precisas para horizontes de curto prazo. No entanto, notamos que quando a amostra real apresentava um valor baixo, todos os modelos tiveram dificuldade em fornecer previsões precisas.

Para previsões de quarenta e cinco dias, embora ainda haja uma diferença entre os valores previstos e os valores reais, essa distância não é tão considerável em comparação com o conjunto de dados anterior. Mais uma vez, o modelo Arima-RF se destaca como o principal modelo, indicando sua robustez e capacidade de generalização para horizontes de previsão mais longos.

A análise da Figura 34, referente ao Conjunto de Dados III, revela percepções interessantes sobre o desempenho dos modelos de previsão para horizontes de 1 e 45 dias.

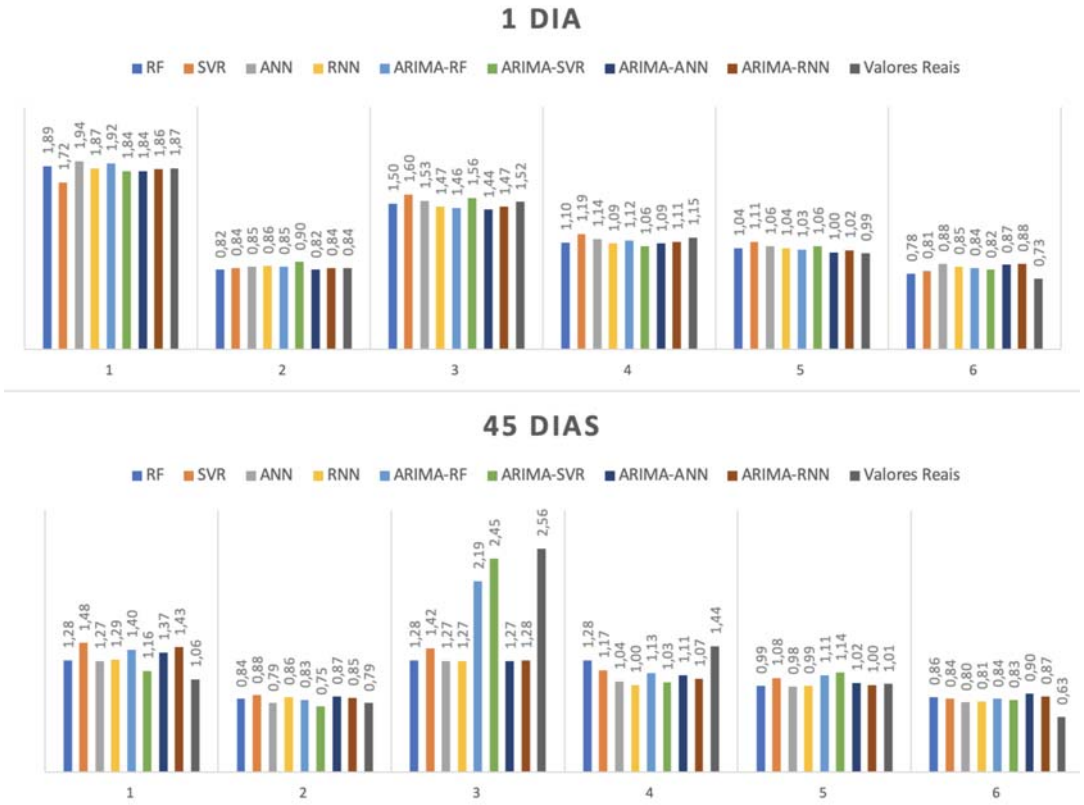

Figura 34 – Análise do Conjunto de dados III - Dados de Nível

Para a previsão de 1 dia, nota-se uma homogeneidade preditiva entre os modelos, indicando que todos eles obtiveram desempenhos semelhantes. No entanto, destaca-se o modelo ANN pela sua boa concordância com o valor real, seguido pelo modelo Arima-RNN. Essa observação sugere que a ANN foi capaz de capturar efetivamente as variações imediatas nos dados, resultando em previsões precisas para um horizonte de curto prazo.

Ao estender o horizonte de previsão para 45 dias, observa-se novamente o destaque do modelo Arima-SVR, que demonstrou uma boa consistência preditiva, semelhante ao que foi observado no Conjunto de Dados I para o mesmo horizonte de previsão.

No entanto, é importante notar que na amostra três, houve uma discrepância considerável entre os valores previstos e os dados reais para a maioria dos modelos. Nesse cenário, os modelos Arima-SVR e Arima-RF foram os que mais se aproximaram dos dados reais, indicando sua capacidade de lidar com padrões complexos e inesperados nos dados.

Ao examinar a Figura 35, que representa o último conjunto de dados analisado, pode-se destacar importantes observações sobre o desempenho dos modelos de previsão para horizontes de 1 e 45 dias.

Para a previsão de 1 dia, os modelos Arima-RF e Arima-RNN se destacam como aqueles que obtiveram uma boa concordância com os valores reais. Essa observação

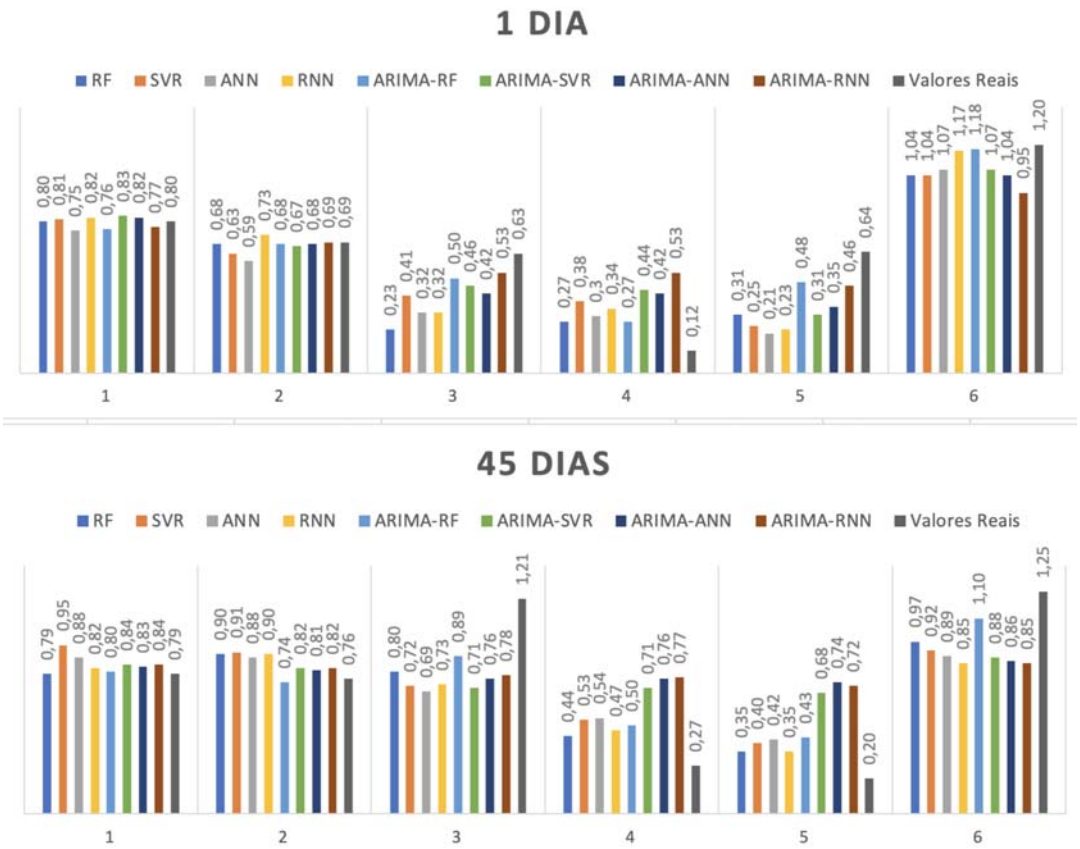

Figura 35 – Análise do Conjunto de dados IV - Dados de Nível

sugere que a combinação do ARIMA com a técnica de *Random Forest* e a rede neural recorrente foram eficazes em capturar as variações imediatas nos dados, resultando em previsões precisas para horizontes de curto prazo.

No entanto, ao analisarmos a terceira amostra, observamos que, quando o valor real é reduzido, os modelos tendem a não captar adequadamente a variabilidade presente nos dados. Isso resulta em previsões que podem ficar um pouco abaixo dos valores reais, destacando a sensibilidade dos modelos às características específicas dos dados de entrada.

Quanto à previsão para 45 dias, notamos que os modelos Arima-RF e RF continuam a apresentar um bom desempenho preditivo, especialmente nas amostras três, quatro e cinco. No entanto, é importante observar que quando os dados reais se afastam do padrão esperado, seja por serem maiores ou menores do que o habitual, os modelos podem não obter uma boa aproximação dos valores reais. Isso ressalta a importância de considerar a variabilidade dos dados e as limitações dos modelos ao realizar previsões de longo prazo.

Essas observações enfatizam a necessidade de uma análise cuidadosa dos resultados de previsão, levando em consideração não apenas a precisão geral dos modelos, mas também sua capacidade de lidar com diferentes cenários e características dos dados. Essa compreensão aprofundada pode orientar na seleção do modelo mais

apropriado e na interpretação correta das previsões para uma aplicação específica.

Após a análise detalhada dos resultados obtidos pelos modelos de previsão para diferentes conjuntos de dados e horizontes de previsão, é fundamental realizar testes adicionais nos dados atuais. Esses testes nos permitirão avaliar a robustez e a generalização dos modelos em cenários reais, onde novos dados podem ser inseridos de forma manual. Na próxima seção, será explorado esses testes, examinando como os modelos se comportam quando confrontados com dados atualizados e inseridos manualmente. Esse passo é essencial para garantir que os modelos de previsão sejam capazes de manter seu desempenho e precisão ao lidar com informações em tempo real, contribuindo assim para uma aplicação prática e eficaz desses modelos em diversos contextos.

4.3.4.2 Teste com dados Atuais

Na próxima fase deste estudo, os modelos serão submetidos a um teste utilizando dados mais recentes, abrangendo o período de 01/08/2023 a 15/09/2023, este intervalo foi utilizado pelo fato de ser possível obter todos os dados climatológicos e os futuros dados de nível para este teste. É importante ressaltar que, para esta etapa, a entrada de dados foi realizada manualmente, garantindo a inclusão de informações atualizadas. Os dados climatológicos foram obtidos a partir da página da Embrapa, enquanto os dados de nível foram retirados da Agência da Lagoa Mirim, proporcionando uma fonte confiável e atualizada para ambas as variáveis e em concordância com os dados do conjunto de treinamento.

Os dados de entrada utilizados nos modelos incluem os dados climatológicos e de nível para a data de 01/08/2023. A evaporação registrada foi de 1,1mm, enquanto a insolação atingiu 10,1 horas. A precipitação foi nula, com 0mm registrados durante o período analisado. As temperaturas máxima, média e mínima foram medidas em 25℃, 16,6℃ e 10,4℃, respectivamente. A umidade relativa do ar foi elevada, alcançando 88,5%. A velocidade do vento foi moderada, registrando 2,1m/s, com a direção do vento em Santa Isabel e São Gonçalo apontando predominantemente para 356,57° e 356,77°, respectivamente. Além disso, os níveis de água em diferentes pontos ao longo do canal apresentaram variações, com o nível em Santa Isabel alcançando 1,81m, em Santa Vitória 1,37m, em Montante 1,2m, e no Jusante 1,1m. Esses dados fornecem uma visão abrangente das condições ambientais e hidrológicas durante o período considerado, essenciais para a análise e modelagem dos processos relacionados à água na região e sua posterior predição.

A escolha de inserir dados manualmente tem sua justificativa na necessidade de avaliar a capacidade de adaptação dos modelos a informações mais recentes e verificar a robustez de suas previsões em situações de entrada de dados prática. Este teste em dados atualizados busca não apenas avaliar a capacidade de generalização

dos modelos, mas também fornecer percepções sobre sua eficácia em lidar com informações em tempo real. A introdução de novos dados pode apresentar desafios únicos, como a variabilidade temporal e a resposta a eventos climáticos imprevistos. Essa abordagem manual na entrada de dados reflete a prática comum em cenários práticos de previsão, onde a disponibilidade de dados em tempo real pode ser crucial para decisões rápidas e informadas.

Ao considerar a fonte confiável dos dados climatológicos e de nível, espera-se que este teste proporcione uma visão abrangente do desempenho dos modelos em condições realistas. A combinação de dados atualizados e fontes confiáveis contribui para a validade e a relevância dos resultados, permitindo uma avaliação mais precisa da capacidade preditiva dos modelos em contextos dinâmicos e em constante mudança.

Tabela 7 – Teste em Dados Atuais

Análise dos Modelos com Informações Atuais									
Data	y	DataSet I		DataSet II		DataSet III		DataSet IV	
		RF	Arima-RNN	RF	Arima-RF	RF	Arima-RF	RF	Arima-RNN
01/08/2023	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2
02/08/2023	1,19	1,2	1,19	1,19	1,18	1,21	1,18	1,29	1,39
03/08/2023	1,15	1,2	1,15	1,19	1,16	1,22	1,18	1,22	1,13
04/08/2023	1,21	1,21	1,21	1,19	1,16	1,21	1,2	1,29	1,35
05/08/2023	1,2	1,21	1,18	1,16	1,13	1,21	1,25	1,13	1,27
06/08/2023	1,35	1,15	1,33	1,13	1,35	1,18	1,16	1,07	1,08
07/08/2023	1,29	1,22	1,25	1,12	1,15	1,18	1,17	1,08	1,16
08/08/2023	1,26	1,2	1,2	1,08	1,15	1,16	1,23	1,04	1,01
09/08/2023	1,4	1,23	1,23	1,11	1,35	1,22	1,22	1,12	1,18
10/08/2023	1,14	1,23	1,23	1,13	1,19	1,21	1,19	1,21	1,15
11/08/2023	1,17	1,2	1,2	1,12	1,16	1,21	1,16	1,17	1,16
12/08/2023	1,19	1,21	1,21	1,14	1,16	1,2	1,23	1,19	1,11
13/08/2023	1,18	1,2	1,2	1,14	1,32	1,22	1,16	1,05	1,09
14/08/2023	1,05	1,21	1,07	1,17	1,16	1,25	1,22	1,18	1,13
15/08/2023	1,02	1,21	1,21	1,18	1,23	1,22	1,16	1,09	1,07
16/08/2023	1,19	1,19	1,19	1,21	1,18	1,25	1,22	1,02	1,04
17/08/2023	1,1	1,26	1,2	1,2	1,3	1,26	1,22	1,18	1,11
18/08/2023	1,08	1,23	1,2	1,21	1,24	1,31	1,23	1,28	1,01
19/08/2023	1,2	1,24	1,18	1,23	1,24	1,3	1,26	1,01	1,05
20/08/2023	1,2	1,23	1,22	1,27	1,39	1,34	1,22	1,06	1,07
21/08/2023	1,2	1,23	1,19	1,24	1,23	1,27	1,22	1,06	1,07
22/08/2023	1,1	1,24	1,24	1,26	1,24	1,29	1,22	1,02	1,06
23/08/2023	1,1	1,24	1,24	1,23	1,22	1,31	1,19	1,02	1,06
24/08/2023	1,1	1,23	1,12	1,22	1,18	1,28	1,21	1,2	1
25/08/2023	1,1	1,23	1,23	1,23	1,21	1,28	1,23	1,22	1,14
26/08/2023	1,1	1,2	1,2	1,24	1,23	1,27	1,21	1,1	1,1
27/08/2023	1,14	1,21	1,21	1,22	1,12	1,25	1,19	0,97	0,97
28/08/2023	1,05	1,25	1,25	1,22	1,17	1,27	1,22	1,03	1,03
29/08/2023	1,05	1,25	1,25	1,22	1,17	1,27	1,22	1,03	1,03
30/08/2023	1,1	1,23	1,23	1,23	1,23	1,25	1,17	1,09	1,09
31/08/2023	1,05	1,22	1,22	1,25	1,19	1,28	1,23	1,03	1,03
01/09/2023	1,4	1,25	1,25	1,26	1,25	1,29	1,16	0,95	0,95
02/09/2023	0,96	1,26	1,26	1,29	1,21	1,3	1,22	1,14	1,14
03/09/2023	1,05	1,27	1,27	1,3	1,28	1,32	1,19	1	1
04/09/2023	1,01	1,25	1,25	1,25	1,29	1,31	1,2	0,96	0,96
05/09/2023	1,16	1,27	1,27	1,28	1,28	1,3	1,16	1,03	1,03
06/09/2023	1,2	1,26	1,26	1,28	1,29	1,3	1,19	1,01	1,01
07/09/2023	1,34	1,26	1,26	1,25	1,28	1,3	1,16	1,05	1,11
08/09/2023	1,3	1,26	1,26	1,26	1,28	1,27	1,17	0,92	0,98
09/09/2023	1,5	1,2	1,27	1,25	1,28	1,22	1,19	1,01	1,01
10/09/2023	1,95	1,25	1,25	1,26	1,29	1,25	1,17	0,9	0,95
11/09/2023	2,05	1,24	1,22	1,26	1,28	1,24	1,19	1	1,02
12/09/2023	1,93	1,24	1,28	1,26	1,28	1,24	1,19	1,01	1,02
13/09/2023	1,76	1,23	1,23	1,26	1,27	1,22	1,17	1,01	1,04
14/09/2023	1,87	1,2	1,53	1,23	1,28	1,19	1,19	1,01	1,03
15/09/2023	1,56	1,26	1,36	1,24	1,28	1,2	1,19	1,02	1

Após a análise prévia realizada no conjunto de teste anterior, é proposto agora um teste adicional, conforme a Tabela 7 considerando os modelos que demonstraram os melhores resultados preditivos nessa base. Mesmo que a base de teste não apresente uma defasagem significativa, esse teste adicional aprofundará a avaliação dos modelos mais eficazes em cenários mais recentes. Normalmente, os dados são avaliados na base de testes, conforme verificado nos estudos dos capítulos anteriores. No entanto, reconhecendo a importância de adaptar a análise aos dados mais atualizados, optamos por realizar um teste manual com dados um pouco mais recentes. Isso nos permitirá verificar como os modelos se comportam em situações mais próximas da realidade atual e avaliar sua capacidade de generalização para novos dados.

A escolha de focar nos modelos de melhor desempenho da base de teste anterior é guiada pelo objetivo de explorar a consistência e a adaptabilidade desses modelos em diferentes contextos temporais. Essa abordagem também permite uma comparação direta entre os modelos, proporcionando percepções sobre sua robustez e capacidade de manter um desempenho superior ao longo do tempo.

Ao considerar os modelos que se destacaram anteriormente, busca-se uma compreensão mais aprofundada de como esses modelos lidam com dados mais recentes, incorporando variantes temporais e possíveis mudanças nos padrões observados. Essa estratégia contribui para uma avaliação mais abrangente da aplicabilidade prática dos modelos, consolidando a confiança em sua eficácia em situações dinâmicas e em evolução constante.

Tabela 8 – Resultado em Dados Atuais

Conjunto	**DataSet I**		**DataSet II**		**DataSet III**		**DataSet IV**	
Métrica	RF	Arima-RNN	RF	Arima-RF	RF	Arima-RF	RF	Arima-RNN
RMSE	0,2633	0,2394	0,2615	0,2453	0,2781	0,2763	0,3594	0,3468
MAE	0,176	0,1506	0,1506	0,165	0,2	0,1767	0,231	0,2145

Os resultados obtidos na análise dos modelos com a base de testes mais recente reforçam a importância da disponibilidade de dados para o desempenho preditivo. Interessante observar que ocorre uma tendência consistente de melhoria no desempenho à medida que mais dados estão disponíveis, conforme evidenciado pelos resultados de RMSE e MAE no segundo conjunto de dados.

Como mostra a Tabela 8 o *Random Forest*, ARIMA-RF e ARIMA-RNN demonstraram, mais uma vez, uma capacidade robusta de adaptação aos padrões presentes nos dados, revelando uma consistência em sua eficácia preditiva. Ressaltando que estes erros são expressos em metros.

Esses resultados reforçam a relevância de abordagens híbridas, como ARIMA-RF e ARIMA-RNN, que combinam métodos clássicos de séries temporais com técnicas mais avançadas de aprendizado de máquina. Essas combinações estratégicas permitem capturar tanto a estrutura temporal complexa quanto as relações não lineares

presentes nos dados, resultando em modelos mais poderosos e flexíveis.

Um aspecto crucial a ser considerado na avaliação do desempenho dos modelos preditivos é a sensibilidade a valores extremos, especialmente quando se trata de dados que estão muito acima da média. Notou-se que, em situações em que o conjunto de dados apresenta valores significativamente elevados em horizontes preditivos mais longos, o desempenho dos modelos tende a diminuir, esta constatação é a mesma evidenciada pela maioria dos autores na revisão de literatura utilizada.

A justificativa para essa queda no desempenho reside na dificuldade dos modelos em lidar com a imprevisibilidade e a variabilidade introduzida por valores extremos. Quando os dados apresentam picos ou quedas bruscas, especialmente em horizontes temporais mais distantes, os modelos podem ter dificuldades em capturar e generalizar esses padrões imprevisíveis.

Os modelos, por natureza, buscam aprender a partir dos padrões presentes nos dados de treinamento, e a presença de valores extremos pode desafiar a capacidade desses modelos em prever cenários fora da distribuição normal. Isso pode resultar em previsões menos precisas e maior erro, especialmente quando os valores extremos têm um impacto significativo no comportamento futuro da variável dependente.

O desempenho superior do modelo ARIMA-RNN, especialmente no primeiro conjunto de dados, quando lida com *outliers*, pode ser atribuído à natureza combinada desses dois métodos: ARIMA e RNN. A adição das Redes Neurais Recorrentes à estrutura do ARIMA traz uma vantagem significativa ao lidar com padrões mais complexos e não lineares. As RNNs conseguem aprender dependências de longo prazo nos dados, o que é especialmente valioso em situações onde a relação entre variáveis pode ser não linear ou envolver características temporais de longo alcance. Essa capacidade de adaptação das RNNs complementa a estabilidade proporcionada pelo ARIMA, contribuindo para uma modelagem mais robusta em cenários desafiadores.

No caso dos *outliers*, a combinação do ARIMA e RNN permite que o modelo compreenda e ajuste-se a padrões temporais mais complexos e irregulares, enquanto mantém a estabilidade oferecida pelo ARIMA para preservar a coerência temporal. Essa combinação entre métodos clássicos de séries temporais e técnicas avançadas de aprendizado profundo confere ao ARIMA-RNN uma vantagem significativa na capacidade de lidar com dados que contenham valores extremos ou anomalias.

Assim, a robustez do modelo ARIMA-RNN na presença de *outliers* no primeiro conjunto de dados pode ser entendida como resultado da combinação estratégica dessas duas abordagens, proporcionando uma modelagem mais adaptável e precisa diante de situações desafiadoras e complexas.

Para um melhor entendimento destes resultados na Figura 36, tem-se um gráfico de *boxplot* que apresenta uma visualização das variações preditivas dos modelos selecionados durante o teste com os dados atuais. Cada *boxplot* no gráfico representa a

distribuição das previsões feitas pelos diferentes modelos, permitindo uma comparação imediata de suas performaces. As caixas dos *boxplots* mostram a dispersão dos valores previstos, com a linha mediana indicando a tendência central. As *whiskers* (linhas que se estendem para fora da caixa) mostram a variabilidade dos dados além das quartis, enquanto os pontos fora das *whiskers* podem ser considerados *outliers*.

Ao analisar o gráfico de *boxplot*, é possível observar não apenas a tendência central das previsões feitas pelos modelos, mas também a dispersão e a variabilidade dos resultados. Isso permite uma avaliação abrangente da consistência e do desempenho relativo de cada modelo, fornecendo percepções para a seleção e otimização dos modelos de previsão para uso futuro.

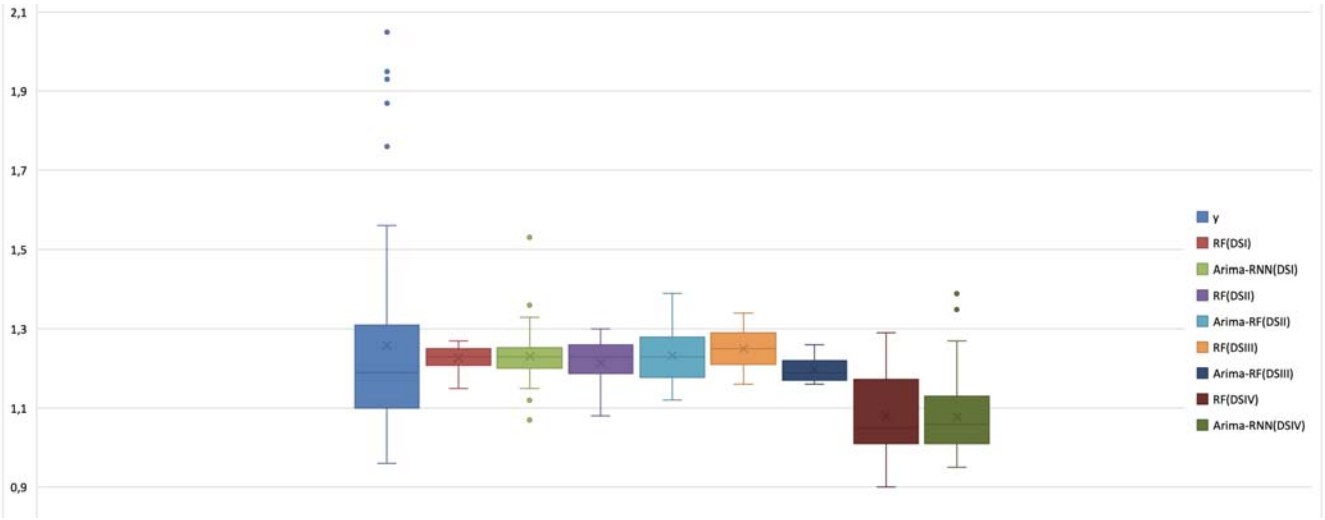

Figura 36 – Gráfico Boxplot de Comparação Preditiva

No caso dos três primeiros conjuntos de dados, observa-se que as previsões dos modelos estão localizadas na parte superior da mediana dos dados dependentes (y), indicando que, em geral, os modelos tendem a superestimar os valores de y. No entanto, é importante notar que essas previsões ainda se encontram dentro do intervalo interquartil (box) de y, o que sugere que os modelos estão capturando razoavelmente bem a tendência central dos dados.

Por outro lado, no quarto conjunto de dados, as previsões dos modelos estão localizadas no quartil inferior de y. Isso indica que os modelos estão subestimando os valores de y, sugerindo um viés sistemático em direção aos valores menores. Essa discrepância entre as previsões e os dados reais pode indicar uma limitação nos modelos em capturar a variabilidade dos dados ou uma inadequação na representação do fenômeno em questão.

Além disso, ao observar os *outliers* ou valores extremos, percebe-se que, de forma geral, os modelos têm dificuldade em prever esses pontos, o que é fator problemático mencionado nos trabalho relacionados. Esses valores estão distantes das previsões dos modelos, indicando uma baixa capacidade de generalização para situações ex-

tremas. No entanto, destaca-se que o modelo ARIMA-RNN parece se destacar nesse aspecto, sendo capaz de se adaptar melhor aos valores extremos em comparação com os outros modelos. Isso pode ser atribuído à capacidade do ARIMA-RNN de capturar tanto padrões temporais de curto prazo (ARIMA) quanto padrões complexos e não lineares (RNN), proporcionando uma maior flexibilidade e robustez na modelagem dos dados.

Em outro viés, em nosso estudo, considerando, contextos, diferentes ao aplicarmos o modelo ARIMA-RF para realizar previsões há cinco dias à frente utilizando o segundo conjunto de dados. O resultado obtido foi um RMSE de 0,1054. Comparativamente, ao revisarmos o trabalho de (Nguyen et al., 2020), notamos que, ao adotar uma abordagem híbrida envolvendo ARIMA-KNN, o autor alcançou um RMSE de 0,3668 para uma previsão de cinco dias. Essa diferença sugere uma performace consideravelmente superior do nosso modelo ARIMA-RF em termos de precisão preditiva, reiterando que são cenários distintos, mas as escalas dos dados são semelhantes, logo fornecendo assim uma percepção dos bons resultados.

Em outra comparação com o trabalho de (Xie; Lou, 2019), que empregou o modelo híbrido ARIMA-SVR para prever um dia à frente, observamos que o RMSE foi de 0,2666. No entanto, ao adotarmos o ARIMA-RNN em nosso estudo para a mesma janela preditiva de um dia à frente, conseguimos alcançar um RMSE significativamente menor, aferindo 0,0664. Essa discrepância indica que nosso modelo superou o desempenho apresentado no estudo de (Xie; Lou, 2019) em termos de acurácia nas previsões de curto prazo.

Vale ressaltar que diversos fatores podem influenciar nos resultados obtidos em diferentes abordagens de modelagem e conjuntos de dados. Um dos fatores cruciais é a quantidade de atributos considerados. Em nosso caso, dispomos de um conjunto de dados relativamente extenso, o que pode contribuir para aprimorar a capacidade preditiva do modelo. Além disso, a escolha adequada dos hiperparâmetros e a qualidade do pré-processamento dos dados também desempenham papéis essenciais nos resultados. Outro fator que pesa em relação a este trabalho é que tornamos mais rígidos na avaliação das predições, ou seja, dividindo os conjuntos não somente em treinamento e testes, mas sim em treinamento, validação e testes.

Assim, embora os resultados sejam promissores em nosso estudo, é importante destacar que a eficácia de um modelo depende de uma análise aprofundada de múltiplos fatores, e as variações nos métodos e dados podem levar a resultados distintos.

Em resumo, a seleção do conjunto de dados 2 para a implementação de um modelo preditivo em produção destaca-se pela sua capacidade de fornecer uma base sólida e abrangente para análises futuras. Embora a obtenção desses dados possa ser mais desafiadora, sua qualidade e quantidade superiores compensam essas dificuldades. O modelo ARIMA-RF, escolhido por sua consistência e desempenho com-

provado, adiciona uma camada de confiabilidade às previsões, garantindo resultados precisos e relevantes para orientar estratégias futuras com confiança. Assim, ao optar por esta abordagem, podemos garantir não apenas a eficácia das previsões, mas também a sustentabilidade e o sucesso contínuo das iniciativas baseadas em dados.

Para realizar previsões precisas e confiáveis sobre os níveis de água, é crucial utilizar todos os dados climatológicos disponíveis. Dentre esses dados, a velocidade do vento emerge como um fator essencial, embora sua relação com os níveis de água seja forte, porém indireta. Dada essa complexidade, foi necessário recorrer a modelos avançados, como o *Random Forest*, para capturar sua relevância e impacto na variação dos níveis de água ao longo do tempo ou do espaço.

Além da velocidade do vento, a temperatura também desempenha um papel significativo na determinação dos níveis de água. Sua relação direta com os níveis de água é uma evidência clara da influência das condições térmicas no comportamento dos corpos de água.

Outro aspecto a ser considerado é a direção do vento, que, apesar de apresentar uma relação importante com os níveis de água, possui menos observações disponíveis em comparação com outros preditores climáticos. Mesmo com essa limitação, as poucas observações foram suficientes para destacar sua relevância na modelagem dos níveis de água, como evidenciado no quarto conjunto de dados analisado.

Estudos anteriores e análises recentes têm apontado a velocidade do vento como o atributo mais significativo na região de Santa Isabel quando se trata de prever o nível de água do canal São Gonçalo. Essa constatação se deve não apenas à influência direta da velocidade do vento sobre a movimentação e a evaporação da água, mas também à sua capacidade de impactar outras variáveis ambientais relevantes. Autores anteriores já evidenciaram a importância crucial da velocidade do vento, destacando sua influência direta na dinâmica dos corpos de água e seu papel fundamental na formação de padrões climáticos locais.

Em segundo lugar, a temperatura média surge como um atributo relevante na modelagem do nível de água do canal São Gonçalo. Sua influência é percebida tanto na sazonalidade quanto nas tendências de longo prazo dos níveis de água, refletindo a importância das condições térmicas na regulação do ciclo hidrológico.

Em terceiro lugar, a precipitação é considerada um fator de influência significativo nos níveis de água do canal. A quantidade e a distribuição das chuvas desempenham um papel crucial na recarga dos aquíferos, no fluxo dos rios e no balanço hídrico da região, impactando diretamente os níveis de água do canal São Gonçalo.

Outros atributos, como a direção do vento, a umidade, a insolação e a evaporação, também são considerados importantes nessa sequência na modelagem do nível de água do canal, embora com menor influência em comparação com a velocidade do vento, a temperatura média e a precipitação. Essas variáveis desempenham pa-

péis complementares, contribuindo para uma compreensão abrangente e precisa dos padrões hidrológicos na região de Santa Isabel. Destaca-se que a direção do vento possui menos dados disponíveis para os testes com os modelos.

A velocidade do vento é destacada como o atributo mais significativo na região para a previsão do nível de água do canal São Gonçalo, seguida pela temperatura média e pela precipitação. Essas conclusões são fundamentais para o desenvolvimento de modelos preditivos robustos e eficazes, capazes de fornecer informações valiosas para a gestão e a tomada de decisões relacionadas aos recursos hídricos na região.

Em resumo, a utilização abrangente e criteriosa de todos os dados climatológicos disponíveis é essencial para uma predição precisa e acurada dos níveis de água. A combinação de diferentes variáveis, como velocidade do vento, temperatura e direção do vento, por meio de técnicas avançadas de modelagem, permite uma compreensão mais completa e detalhada dos padrões hidrológicos e ambientais, contribuindo para uma gestão eficaz dos recursos hídricos e para a mitigação de riscos relacionados à água.

4.3.5 Limitações dos modelos utilizados

Nesta seção, serão abordadas as limitações observadas nos modelos discutidos, destacando suas deficiências em lidar com desafios específicos encontrados em problemas de séries temporais. Serão analisadas questões como a capacidade de capturar padrões temporais complexos, a sensibilidade a *outliers*, a complexidade algorítmica e outros fatores que podem impactar a eficácia e a aplicabilidade desses modelos na previsão de séries temporais. Ao compreender e reconhecer essas limitações, os praticantes e pesquisadores podem tomar decisões informadas sobre a seleção e o desenvolvimento de modelos mais adequados para seus cenários de aplicação específicos.

Ao aplicar modelos como *Random Forest* e *Support Vector Regression* em problemas de regressão em séries temporais, uma das limitações significativas é a incapacidade de capturar padrões temporais. O *Random Forest* trata cada ponto de dados de forma independente, sem considerar a sequência temporal dos eventos. Isso significa que as dependências temporais, como tendências e sazonalidades cruciais em séries temporais, podem não ser adequadamente capturadas pelo modelo. Da mesma forma, o SVR não leva explicitamente em conta a ordem temporal dos dados, o que pode resultar em previsões que não refletem adequadamente as dinâmicas temporais subjacentes. Esses modelos, portanto, podem não identificar e aprender corretamente os padrões complexos ao longo do tempo, comprometendo a acurácia das previsões. Neste caso, para diminuir a influência desse problema geramos esses atrasos temporais de forma manual.

Além disso, tanto *Random Forest* quanto SVR podem ser sensíveis a *outliers*, o que representa outra limitação importante. No caso do SVR, *outliers* podem influenciar significativamente o modelo, especialmente se os parâmetros de regularização não forem bem ajustados. Isso pode levar a previsões distorcidas e menos precisas. O *Random Forest* é mais robusto aos *outliers*, mas ainda pode ser afetado se houver uma quantidade significativa de *outliers* nos dados de treinamento. A presença de *outliers* em séries temporais pode distorcer a distribuição dos dados, fazendo com que o modelo aprenda padrões incorretos ou dê peso excessivo a eventos atípicos, afetando negativamente a precisão das previsões.

Uma das limitações significativas dos modelos híbridos propostos para a predição em séries temporais é a complexidade algorítmica, que resulta em um aumento exponencial do tempo de processamento e previsão. Esses modelos combinam múltiplas técnicas, como redes neurais e métodos estatísticos, para tentar capturar padrões complexos nos dados. Embora possam oferecer maior precisão, a integração de diferentes abordagens aumenta a complexidade computacional, exigindo mais recursos de processamento e memória. Consequentemente, o tempo necessário para treinar e validar esses modelos é substancialmente maior, tornando-os menos práticos para aplicações em tempo real ou para conjuntos de dados muito grandes. Essa complexidade também pode dificultar a implementação e a manutenção dos modelos, exigindo maior domínio técnico e recursos computacionais avançados.

Em outro viés, os modelos híbridos propostos podem, de fato, ser aplicados em outras bacias hidrográficas, pois as metodologias utilizadas são adaptáveis a diversos cenários hidrológicos. No entanto, é importante destacar que esses algoritmos foram especificamente treinados e calibrados com dados do Canal São Gonçalo. Como resultado, o desempenho preditivo observado pode não ser replicado com a mesma precisão em outras bacias sem uma readequação do modelo. Para alcançar resultados confiáveis em diferentes contextos hidrológicos, seria necessário um novo processo de treinamento e validação com dados locais específicos, garantindo que o modelo capture as particularidades e variabilidades das novas áreas de estudo.

A implementação desses modelos em bacias hidrográficas compartilhadas enfrenta diversos desafios, sendo a obtenção e padronização dos dados de medição um dos principais. A variabilidade na frequência de monitoramento entre diferentes centros de medição, onde alguns registram dados a cada hora enquanto outros o fazem apenas uma vez por dia, dificulta a harmonização dos dados para uma escala temporal uniforme. Essa disparidade complica a integração dos dados em um modelo coerente, pois inconsistências na resolução temporal podem levar a resultados imprecisos e menos confiáveis. Portanto, para aplicar eficazmente esses modelos em bacias hidrográficas compartilhadas, é fundamental desenvolver métodos que possam normalizar diferentes frequências de monitoramento, garantindo assim a consistência

e a precisão das previsões.

Os modelos apresentados ainda necessitam de testes em cenários que envolvem eventos extremos de nível de água para avaliar seu desempenho preditivo em condições adversas. Embora tenham demonstrado bom desempenho preditivo em situações de normalidade, é crucial verificar sua precisão e robustez durante eventos extremos, como inundações ou secas severas. Testar os modelos nessas circunstâncias desafiadoras permitirá identificar possíveis limitações e ajustá-los para garantir que possam fornecer previsões confiáveis e úteis em todas as condições, assegurando uma gestão mais eficaz e segura das bacias hidrográficas.

5 CONCLUSÃO

A criação e análise de conjuntos de dados com abordagens distintas representam contribuições substanciais para esta tese, evidenciando a complexidade e os desafios inerentes à modelagem de problemas que envolvem séries temporais com dados climatológicos. O processo envolveu um investimento significativo de tempo para compreender as variantes e padrões presentes nos dados, estabelecendo uma base sólida para a investigação.

Uma contribuição adicional foi a concepção e desenvolvimento de dois modelos distintos. O primeiro modelo, voltado para séries temporais, que se destaca por fornecer resultados a partir do final da base temporal, possibilitando uma abordagem que considera a temporalidade das informações. Já o segundo modelo proposto, apresenta uma abordagem mais flexível, permitindo a inserção manual de dados e a realização de previsões em qualquer ponto no tempo, proporcionando uma aplicabilidade mais ampla e adaptável a cenários práticos.

A justificativa para a criação desses modelos distintos reside na necessidade de abordagens versáteis que atendam a diferentes contextos e demandas práticas. O primeiro modelo, ao considerar a série temporal até o final da base, atende a cenários onde a temporalidade é essencial para as previsões. Por outro lado, o segundo modelo oferece maior flexibilidade, adaptando-se a situações em que a inserção manual de dados e previsões em momentos específicos são cruciais.

A combinação desses elementos contribui para uma compreensão mais abrangente e aplicada da modelagem preditiva em cenários climatológicos. O desafio enfrentado e as soluções propostas refletem a complexidade inerente à interação entre dados climatológicos e a previsão de eventos futuros. Essa abordagem multidimensional enriquece não apenas o entendimento teórico, mas também a aplicabilidade prática das técnicas desenvolvidas.

A adoção de técnicas híbridas, combinando a modelagem linear do ARIMA para os dados de nível com modelos tradicionais de aprendizagem de máquina para os dados climatológicos, emergiu como uma estratégia eficaz e comprovadamente superior em relação aos modelos tradicionais de aprendizagem de máquina. A união entre a

capacidade do ARIMA em capturar padrões temporais em dados lineares e a flexibilidade dos modelos de aprendizagem de máquina em lidar com complexidades não lineares dos dados climatológicos resultou em desempenho aprimorado e previsões mais precisas. Essa abordagem híbrida proporciona uma compreensão mais abrangente das relações presentes nos dados climatológicos, destacando a importância de integrar métodos tradicionais e avançados para obter uma modelagem mais completa e eficiente em cenários desafiadores.

A tese não apenas avança no campo da modelagem de séries temporais climatológicas, mas também contribui com metodologias práticas e adaptáveis. Uma delas é destinada para a modelagem de séries temporais em si, enquanto a outra é voltada para a incorporação de aprendizado de máquina em séries temporais. Essas metodologias podem ser aplicadas em uma variedade de contextos, levando em consideração as especificidades e desafios associados às previsões climáticas. Essa abordagem integrada e flexível representa um avanço significativo na compreensão e aplicação eficaz de modelos preditivos em domínios climatológicos desafiadores.

Essas metodologias foram desenvolvidas com base em uma ampla gama de dados climatológicos provenientes da base do Instituto Nacional de Meteorologia. Além disso, foram considerados dados específicos de nível e direção do vento, fornecidos pela Agência da Lagoa Mirim. A inclusão desses dados detalhados e abrangentes permitiu uma análise mais aprofundada e precisa dos padrões climáticos locais, fortalecendo assim a robustez e a eficácia das metodologias propostas.

REFERÊNCIAS

AMBROSI, E. **VARIABILIDADE DA LINHA DE COSTA NA FOZ DO CANAL SÃO GONÇALO E ADJACÊNCIAS DA LAGOA DOS PATOS**. 2018. Dissertação (Mestrado em Ciência da Computação) — UNIVERSIDADE FEDERAL DO RIO GRANDE, Santa Maria.

BAKSHI, K.; BAKSHI, K. Considerations for artificial intelligence and machine learning: Approaches and use cases. In: IEEE AEROSPACE CONFERENCE, 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p.1–9.

BARROS, F. M. Lins-de. Análise integrada da vulnerabilidade costeira e riscos associados. In: VI CONGRESSO PLANEJAMENTO E GESTÃO DAS ZONAS COSTEIRAS DOS PAÍSES DE EXPRESSÃO PORTUGUESA. ILHA DE BOA VISTA, CABO VERDE, CD-ROM, 2011. **Anais...** [S.l.: s.n.], 2011.

BELTRAME, L. F. d. S. et al. Estudo para avaliação e gerenciamento da disponibilidade hídrica da Bacia da Lagoa Mirim. , [S.l.], 1998.

BIRYLO, M.; RZEPECKA, Z.; KUCZYNSKA-SIEHIEN, J.; NASTULA, J. Analysis of water budget prediction accuracy using ARIMA models. **Water Science and Technology: Water Supply**, [S.l.], v.18, n.3, p.819–830, 2018.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v.4, n.4.

BLAIN, G. C.; KAYANO, M. T.; CAMARGO, M. B. P. d.; LULU, J. Variabilidade amostral das séries mensais de precipitação pluvial em duas regiões do Brasil: Pelotas-RS e Campinas-SP. **Revista Brasileira de Meteorologia**, [S.l.], v.24, p.1–11, 2009.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time series analysis**: forecasting and control. [S.l.]: John Wiley & Sons, 2015.

BOX, M. A.; BOX, G. P. **Physics of radiation and climate**. [S.l.]: Crc Press, 2015.

BREIMAN, L. Bagging predictors. **Machine learning**, [S.l.], v.24, n.2, p.123–140, 1996.

BRERETON, R. G.; LLOYD, G. R. Support vector machines for classification and regression. **Analyst**, [S.l.], v.135, n.2, p.230–267, 2010.

CASTILLO, J. M. M.; CSPEDES, J. M. S.; CUCHANGO, H. E. E. Water level prediction using artificial neural network model. **J. Appl. Eng. Res**, [S.l.], v.13, p.14378–14381, 2018.

CAVALCANTE, R. B. L.; MENDES, C. A. B. Calibração e validação do módulo de correntologia do modelo IPH-A para a Laguna dos Patos (RS/Brasil). **Rbrh: revista brasileira de recursos hídricos. Porto Alegre, RS. Vol. 19, n. 3 (jul./set. 2014), p. 191-204**, [S.l.], 2014.

CHANSON, H. Environmental Hydraulics of Open Channel Flows. **JOURNAL OF FLUID MECHANICS**, [S.l.], v.557, n.1, p.473–473, 2006.

CHEN, S.; COWAN, C.; GRANT, P. Orthogonal Least Squares Learning Algorithm for Radial. **IEEE Transactions on neural networks**, [S.l.], v.2, n.2, p.303, 1991.

CHOI, C. et al. Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea. **Water**, [S.l.], v.12, n.1, p.93, 2019.

CLM, a. **Barragem do São Gonçalo. Estudo Preliminar de Viabilidade**. [S.l.]: Seção Brasileira da Comissão da Lagoa Mirim. Tomo I. Relatório. Março, 1970. p. 94, 1970.

COLLARES, G. L. **Lagoa Mirim e Canal São Gonçalo – Hidrovia Uruguai-Brasil – Considerações técnicas**: o território e suas estruturas. [S.l.: s.n.], 2022.

DINIZ, H.; ANDRADE, L.; CARVALHO, A.; ANDRADE, M. Previsão de séries temporais utilizando redes neurais artificiais e modelos de box e jenkins. **Available at: Available at: http://repository. usp. br/single. php**, [S.l.], 1998.

DUDA, R. O.; HART, P. E. et al. **Pattern classification**. [S.l.]: John Wiley & Sons, 2006.

ELMAN, J. L. Finding structure in time. **Cognitive science**, [S.l.], v.14, n.2, p.179–211, 1990.

FINCH, J. A comparison between measured and modelled open water evaporation from a reservoir in south-east England. **Hydrological Processes**, [S.l.], v.15, n.14, p.2771–2778, 2001.

GHIMIRE, B. N. Application of ARIMA model for river discharges analysis. **Journal of Nepal Physical Society**, [S.l.], v.4, n.1, p.27–32, 2017.

GONÇALVES, A. R. Máquina de vetores suporte. **Acesso em**, [S.l.], v.21, 2010.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

GOUVÊA, T. H. **Análise estatística da influência da precipitação e de características do solo na variação do nível d´ água em área de recarga do aquífero Guarani**. 2009. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

HANKE, J. E.; WICHERN, D. W. **Business forecasting**. [S.l.]: Pearson Educación, 2005.

HARTMANN, C.; HARKOT, P. F. C. Influência do canal São Gonçalo ao aporte de sedimentos para o estuário da Laguna dos Patos-RS. **Brazilian Journal of Geology**, [S.l.], v.20, n.1, p.329–332, 1990.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning**: data mining, inference, and prediction. [S.l.]: Springer, 2009. v.2.

HRNJICA, B.; BONACCI, O. Lake level prediction using feed forward and recurrent neural networks. **Water Resources Management**, [S.l.], v.33, n.7, p.2471–2484, 2019.

HUSS, M.; HOCK, R. A new model for global glacier change and sea-level rise. **Frontiers in Earth Science**, [S.l.], v.3, p.54, 2015.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting**: principles and practice. [S.l.]: OTexts, 2018.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v.112.

JOO, T. W.; KIM, S. B. Time series forecasting based on wavelet filtering. **Expert Systems with Applications**, [S.l.], v.42, n.8, p.3868–3874, 2015.

KARSBURG, R. M. **Precipitação e velocidade do vento na oscilação dos níveis d'água do canal São Gonçalo-RS**. 2016. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Pelotas.

KELLEY, H. J. Gradient theory of optimal flight paths. **Ars Journal**, [S.l.], v.30, n.10, p.947–954, 1960.

KIRCHGÄSSNER, G.; WOLTERS, J.; HASSLER, U. **Introduction to modern time series analysis**. [S.l.]: Springer Science & Business Media, 2012.

KUHN, M.; JOHNSON, K. et al. **Applied predictive modeling**. [S.l.]: Springer, 2013. v.26.

KUINCHTNER, A.; BURIOL, G. A. Clima do Estado do Rio Grande do Sul segundo a classificação climática de Köppen e Thornthwaite. **Disciplinarum Scientia| Naturais e Tecnológicas**, [S.l.], v.2, n.1, p.171–182, 2001.

LAWLER, G. F. **Introduction to stochastic processes**. [S.l.]: Chapman and Hall/CRC, 2018.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. Estatística: teoria e aplicações. **Rio de janeiro: LTC**, [S.l.], v.811, 2000.

LIMA, P. R. B. D.; MARQUES, F.; ORTH, S. Predição do nível de água utilizando os modelos ARIMA e Random Forest: Um Estudo de Caso da Barragem Eclusa do São Gonçalo. In: L SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 2023. **Anais...** [S.l.: s.n.], 2023. p.24–35.

MASLIN, M. **Climate change**: a very short introduction. [S.l.]: OUP Oxford, 2014.

MORETTIN, P. A.; TOLOI, C. M. **Análise de séries temporais**: modelos lineares univariados. [S.l.]: Editora Blucher, 2018.

MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python**: a guide for data scientists. [S.l.]: "O'Reilly Media, Inc.", 2016.

NGUYEN, X. H. et al. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. **Advances in Water Resources**, [S.l.], v.142, p.103656, 2020.

RASCHKA, S.; MIRJALILI, V. **Python machine learning**: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. [S.l.]: Packt Publishing Ltd, 2019.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, [S.l.], v.65, n.6, p.386, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, [S.l.], v.323, n.6088, p.533–536, 1986.

RUSSELL, S. J. **Artificial intelligence a modern approach**. [S.l.]: Pearson Education, Inc., 2010.

SANTOS, J. A. dos; RICCIARDI, T. R. Estudo sobre o potencial de aproveitamento de água de chuva na faculdade de engenharia mecânica (FEM). **Revista Ciências do Ambiente On-Line**, [S.l.], v.9, n.1, 2013.

SIMON, A. L. H.; SILVA, P. F. da. Análise geomorfológica da planície lagunar sob influência do canal São Gonçalo–Rio Grande do Sul–Brasil. **Geosciences= Geociências**, [S.l.], v.34, n.4, p.749–767, 2015.

SOLANDER, K. C. et al. Simulating human water regulation: The development of an optimal complexity, climate-adaptive reservoir management model for an LSM. **Journal of Hydrometeorology**, [S.l.], v.17, n.3, p.725–744, 2016.

SORJAMAA, A. et al. Methodology for long-term prediction of time series. **Neurocomputing**, [S.l.], v.70, n.16-18, p.2861–2869, 2007.

STAUDEMEYER, R. C.; MORRIS, E. R. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. **arXiv preprint arXiv:1909.09586**, [S.l.], 2019.

SUN, H.-P.; XIA, X.-L.; LE, J.-J.; HUANG, H. Water level prediction based on polyelement correlation analysis and improved bp neural network. In: INTERNATIONAL CONFERENCE ON NETWORK AND INFORMATION SYSTEMS FOR COMPUTERS (ICNISC), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p.250–253.

TANANAEV, N.; DEBOLSKIY, M. Turbidity observations in sediment flux studies: Examples from Russian rivers in cold environments. **Geomorphology**, [S.l.], v.218, p.63–71, 2014.

TIAO, G. C.; TSAY, R. S. A canonical correlation approach to modeling multivariate time series. In: BUSINESS AND ECONOMIC STATISTICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION', WASHINGTON, DC, 1985. **Proceedings...** [S.l.: s.n.], 1985. p.112–120.

TOMAZELLI, L. O regime dos ventos e a taxa de migração das dunas eólicas costeiras do Rio Grande do Sul, Brasil. **Pesquisas em Geociências**, [S.l.], v.20, n.1, p.18–26, 1993.

TRENBERTH, K. E.; FASULLO, J. T.; KIEHL, J. Earth's global energy budget. **Bulletin of the American Meteorological Society**, [S.l.], v.90, n.3, p.311–324, 2009.

TYUGU, E. Artificial intelligence in cyber defense. In: INTERNATIONAL CONFERENCE ON CYBER CONFLICT, 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p.1–11.

VAN BEEK, L.; WADA, Y.; BIERKENS, M. F. Global monthly water stress: 1. Water balance and water availability. **Water Resources Research**, [S.l.], v.47, n.7, 2011.

VAZ, A. C.; MÖLLER JUNIOR, O. O.; ALMEIDA, T. L. d. Análise quantitativa da descarga dos rios afluentes da Lagoa dos Patos. , [S.l.], 2006.

VERDÉRIO, A. et al. Sobre o uso de regressão por vetores suporte para a construção de modelos em um método de região de confiança sem derivadas. , [S.l.], 2015.

VÖRÖSMARTY, C. J. et al. Global threats to human water security and river biodiversity. **nature**, [S.l.], v.467, n.7315, p.555–561, 2010.

WANG, Q.; WANG, S. Machine learning-based water level prediction in Lake Erie. **Water**, [S.l.], v.12, n.10, p.2654, 2020.

WESCOAT JR, J. L. Books of note–Water in Crisis: A Guide to the World's Fresh Water Resources edited by Peter H. Gleick. **Environment**, [S.l.], v.36, n.4, p.26, 1994.

WURBS, R. A.; AYALA, R. A. Reservoir evaporation in Texas, USA. **Journal of Hydrology**, [S.l.], v.510, p.1–9, 2014.

XIE, Y.; LOU, Y. Hydrological Time Series Prediction by ARIMA-SVR Combined Model based on Wavelet Transform. In: INTERNATIONAL CONFERENCE ON INNOVATION IN ARTIFICIAL INTELLIGENCE, 2019., 2019. **Proceedings...** [S.l.: s.n.], 2019. p.243–247.

XU, G. et al. A water level prediction model based on ARIMA-RNN. In: IEEE FIFTH INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING SERVICE AND APPLICATIONS (BIGDATASERVICE), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p.221–226.

YU, Z.; LEI, G.; JIANG, Z.; LIU, F. ARIMA modelling and forecasting of water level in the middle reach of the Yangtze River. In: INTERNATIONAL CONFERENCE ON TRANSPORTATION INFORMATION AND SAFETY (ICTIS), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p.172–177.

ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O. Recurrent neural network regularization. **arXiv preprint arXiv:1409.2329**, [S.l.], 2014.

ZHANG, J. et al. Using Recurrent Neural Network for Intelligent Prediction of Water Level in Reservoirs. In: IEEE 44TH ANNUAL COMPUTERS, SOFTWARE, AND APPLICATIONS CONFERENCE (COMPSAC), 2020., 2020. **Anais...** [S.l.: s.n.], 2020. p.1125–1126.

ZHU, X. J. Semi-supervised learning literature survey. , [S.l.], 2005.