



Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Cherubin Cunha do Nascimento

Marcello Soares de Oliveira

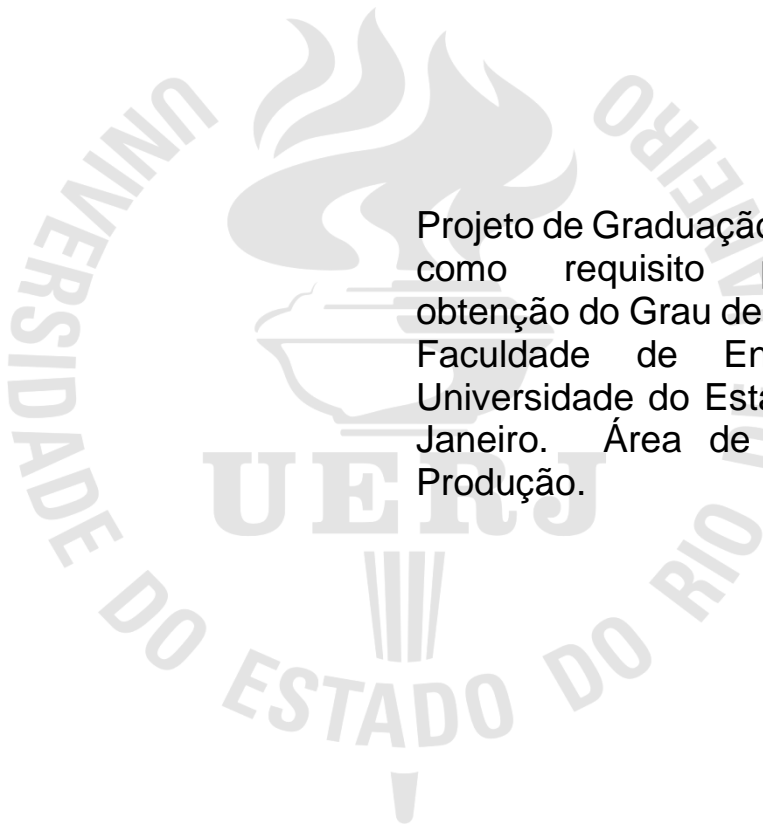
**Análise de Séries Temporais com Técnicas de *Machine Learning*
em Empresas de Saneamento**

Rio de Janeiro

2024

Cherubin Cunha do Nascimento
Marcello Soares de Oliveira

**Análise de Séries Temporais com Técnicas de *Machine Learning*
em Empresas de Saneamento**



Projeto de Graduação apresentado,
como requisito parcial para
obtenção do Grau de Engenheiro, à
Faculdade de Engenharia da
Universidade do Estado do Rio de
Janeiro. Área de concentração
Produção.

Orientador: Prof. Dr. Valter Moreno

Rio de Janeiro

2024

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

N244 Nascimento, Cherubin Cunha do.

Análise de séries temporais com técnicas de *Machine Learning* em empresas de saneamento / Cherubin Cunha do Nascimento, Marcello Soares de Oliveira. – 2024.

48 f.

Orientador: Valter Moreno.

Projeto Final (Graduação) - Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

Bibliografia: f. 46-48

1. Engenharia de produção - Monografias. 2. Análise de séries temporais - Monografias. 3. Abastecimento de água – Monografias. I. Oliveira, Marcello Soares de. II. Moreno, Valter. III. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. IV. Título.

CDU 658.5

Bibliotecário: Iremar Leal – CRB7/5728

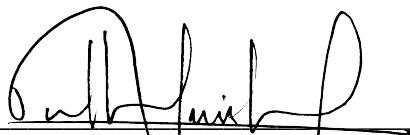
Cherubin Cunha do Nascimento
Marcello Soares de Oliveira

Análise de Séries Temporais com Técnicas de *Machine Learning* em Empresas de Saneamento

Projeto de Graduação apresentado, como requisito parcial para obtenção do Grau de Engenheiro, à Faculdade de Engenharia da Universidade do Estado do Rio de Janeiro. Área de concentração Produção.

Aprovado em: 27 de fevereiro de 2024.

Banca Examinadora:



Prof. Dr. Valter de Assis Moreno Junior (Orientador)
Universidade do Estado do Rio de Janeiro - UERJ



Prof. Msc. Helcio de Oliveira Rocha
Universidade do Estado do Rio de Janeiro - UERJ

Rio de Janeiro

2024

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelas incontáveis bênçãos que já me proporcionou e que ainda tem preparado para enviar sobre mim.

Agradeço a minha companheira Luana Carvalho e a minha filha Bela Nascimento, que são as pessoas mais importantes na minha vida, amo vocês.

Agradeço a minha mãe Clayr Cunha e meu irmão Camilo Nunes por todo amor concedido durante a minha criação e essencial na construção do meu caráter.

Aos colegas e professores da Universidade do Estado do Rio de Janeiro pela trajetória acadêmica, em especial aos grandes amigos Bruno Stürmer, Daniel Fabrizio e Leonardo Colares com quem dividi essa jornada desde o início.

Por fim, agradeço ao orientador Prof. Dr. Valter Moreno pela paciência, instruções e conhecimentos transmitidos. Ao Marcello Soares pela cooperação no desenvolvimento deste Projeto de Graduação.

Cherubin Cunha do Nascimento

Gostaria de em primeiro lugar agradecer a minha mãe, Arlete Soares, e minha querida irmã, Andrea, que aqui expresso minha profunda gratidão por seu amor, apoio e por sempre acreditarem na minha capacidade.

Agradecer ao meu orientador Prof. Dr. Valter Moreno, pela orientação, paciência e valiosos insights ao longo deste trabalho.

Agradeço também ao meu parceiro de trabalho, Cherubin Cunha, pela colaboração, troca de ideias e pelo trabalho em equipe que nos fizeram chegar ao final desta etapa.

Dedico uma parte especial deste trabalho ao meu sobrinho João Marcelo e à Thaina Silva, pela parceria e incentivo.

Minha gratidão à instituição UERJ e a Faculdade de Engenharia, em nome da Professora Maria Eugênia e do Professor Jorge Valério que além de mestres foram grandes amigos.

Agradeço aos amigos do CAENG e do IMB, cuja amizade e apoio foram fundamentais para superar desafios e alcançar este objetivo.

Marcello Soares de Oliveira

RESUMO

NASCIMENTO, Cherubin Cunha do; OLIVEIRA, Marcello Soares de, **Análise de Séries Temporais com Técnicas de *Machine Learning* em Empresas de Saneamento**. Rio de Janeiro, 2024. 48f. Projeto de graduação (Graduação) Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2024.

O objetivo deste projeto foi realizar uma análise de séries temporais utilizando técnicas de *Machine Learning* para previsão do Índice de Perdas de Água na Distribuição (IPD) em empresas de saneamento dos municípios brasileiros. Para atingir este propósito, foi adotada a abordagem de árvores de decisão através do algoritmo Extreme Gradient Boosting (XGBoost). Foram utilizados dados e indicadores dos sistemas de abastecimento de água oriundas do Sistema Nacional de Informações sobre Saneamento (SNIS) e informações socioeconômicas advindas Instituto Brasileiro de Geografia e Estatística (IBGE) e Atlas Brasil.

Ao coletar, tratar e modelar os dados temporais referentes as informações dos sistemas de abastecimento das empresas de saneamento brasileiras e indicadores socioeconômicos da população dos seus respectivos municípios, o projeto buscou desenvolver um modelo através de técnicas de *Machine Learning* que seja capaz de prever o indicador IN049 do SNIS, Índice de Perdas na Distribuição.

O objetivo central do nosso estudo foi alcançado. Esta afirmação possui o respaldo nos resultados obtidos através das métricas de desempenho utilizadas no processo de avaliação e interpretação do modelo desenvolvido.

Palavras-chave: Análise de Séries Temporais, *Machine Learning*, Árvore de Decisão, XGBoost, Redução de Perdas de Água, Sistemas de Abastecimento de Água, Saneamento, IPD, Eficiência Hídrica, SNIS, Modelagem Preditiva.

ABSTRACT

The objective of this project was to conduct a time series analysis using Machine Learning techniques to forecast the Water Loss Index in the Distribution (WLI) in sanitation companies across Brazilian municipalities. To achieve this purpose, the approach of decision trees by way of the Extreme Gradient Boosting (XGBoost) algorithm was adopted. Data and indicators from water supply systems were sourced from the Sistema Nacional de Informações sobre Saneamento (SNIS), while socioeconomic information was obtained from the Instituto Brasileiro de Geografia e Estatística (IBGE) e Atlas Brasil .

By collecting, processing, and modeling temporal data related to information from Brazilian sanitation companies' supply systems and socioeconomic indicators of their respective municipalities' populations, the project proposed to develop a model using machine learning techniques capable of predicting the IN049 indicator from the SNIS, the Water Loss Index.

The mean objective of our study was achieved. This statement is supported by the results obtained through performance metrics used in the evaluation and interpretation of the developed model.

Keywords: Time Series Forecasting, *Machine Learning*, Decision Tree, XGBoost, Water Loss Reduction, Water Supply Systems, Sanitation, WLI, Water Efficiency, SNIS, Predictive Modeling.

LISTA DE FIGURAS

Figura 1 – Fluxograma de procedimentos metodológicos	24
Figura 2 – Pré-processamento de dados: Informações e indicadores por ano	32
Figura 3 – Carregamento dos dados em python com a biblioteca pandas	33
Figura 4 – Análise exploratória: contagem de entradas e tipos de dados	34
Figura 5 – Conversão de variáveis	35
Figura 6 – Divisão em dados de treinamento e teste	36
Figura 7 – Melhores hiperparâmetros encontrados	37
Figura 8 – Resultado das métricas de desempenho	38
Figura 9 – Gráfico de dispersão entre previsões e valores reais	39
Figura 10 – Gráfico de diferença entre valores reais e previsões do ano 0	40
Figura 11 – Gráfico de diferença entre valores reais e previsões do ano 1	40
Figura 12 – Gráfico de diferença entre valores reais e previsões do ano 2	41
Figura 13 – Gráfico de diferença entre valores reais e previsões do ano 3	41

LISTA DE ABREVIATURAS E DE SIGLAS

ARIMA	AutoRegressive Integrated Moving Average
DMC	Distrito de Medição e Controle
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH-M	Índice de Desenvolvimento Humano Municipal
IPD	Índice de Perdas na Distribuição
IWA	International Water Association
MAE	Erro Absoluto Médio
ML	<i>Machine Learning</i>
MSE	Erro Quadrático Médio
PIB	Produto Interno Bruto
PIB_PC	Produto Interno Bruto per capita
R ²	Coefficiente de Determinação
RMSE	Erro Quadrático Médio Raiz
RNN	Redes Neurais Recorrentes
SNIS	Sistema Nacional de Informações sobre Saneamento
UF	Unidade Federativa
WLI	Water Loss Index
XGBOOST	Extreme Gradient Boosting

SUMÁRIO

INTRODUÇÃO	14
Contexto e Relevância da Pesquisa	14
Delimitação da Pesquisa	17
Estrutura do Trabalho	18
1. REFERENCIAL TEÓRICO.....	19
1.1 Aplicação de Análise de Séries Temporais.....	19
1.2 <i>Machine Learning</i>.....	20
1.3 Árvore de Decisão	21
1.4 O Algoritmo XGBoost	22
2. PROCEDIMENTOS METODOLÓGICOS.....	24
2.1 Definição do Escopo do Estudo	24
2.2 Definição do Problema.....	25
2.3 Referencial Teórico e Modelo Utilizado.....	26
2.4 Análise de Séries Temporais Utilizando <i>Machine Learning</i>	27
<u>2.4.1 Coleta de Dados Históricos</u>	<u>27</u>
<u>2.4.2 Pré-processamento dos Dados</u>	<u>27</u>
<u>2.4.3 Aplicação do Algoritmo de <i>Machine Learning</i></u>	<u>28</u>
<u>2.4.4 Treinamento e Teste do Modelo Aplicado</u>	<u>28</u>
2.5 Avaliação e Interpretação dos Resultados do Modelo Aplicado	28
3. DESENVOLVIMENTO DO MODELO DE <i>MACHINE LEARNING</i>	29
3.1 Coleta dos Dados	29
3.2 Análise Exploratória e Tratamento dos Dados	32
3.3 Divisão dos Dados em Treinamento e Teste	35
3.4 Otimização de Hiperparâmetros e Treinamento do Modelo	36
3.5 Avaliação e Interpretação do Modelo com as Métricas de Desempenho	37
3.6 Visualização das Previsões do Modelo	39
CONCLUSÃO	43
Atendimento aos Objetivos da Pesquisa	43
Considerações Finais	44
Sugestões de Trabalhos Futuros.....	44
REFERÊNCIAS.....	46

INTRODUÇÃO

Contexto e Relevância da Pesquisa

Este projeto utilizou abordagens de análise de séries temporais com *Machine Learning* para aprimorar o planejamento da distribuição de água, levando em consideração os resultados acumulados ao longo do horizonte de planejamento. Desta forma, pretendeu-se contribuir para o aperfeiçoamento dos controles do sistema de abastecimento de água brasileiro, para reduzir o volume perdido ao longo do percurso entre a saída para distribuição e o consumo da população. Com isso, esperou-se gerar resultados práticos que venham a mitigar a crise hídrica sofrida atualmente e, conseqüentemente, contribuir para a preservação do meio ambiente de forma geral.

O trabalho utilizou dados reais, coletados através do site do SNIS – Sistema Nacional de Informações sobre Saneamento (“SNIS - Diagnóstico anual de Água e Esgotos”, 2019). O SNIS consiste em um banco de dados administrado na esfera federal, e contém informações sobre a prestação de serviços de água e esgoto. As informações e indicadores disponibilizados pelo SNIS servem a múltiplos propósitos. No âmbito federal, elas se destinam ao planejamento e à execução das políticas públicas, visando orientar a aplicação de investimentos, a construção de estratégias de ação, o acompanhamento de programas, bem como a avaliação do desempenho dos serviços (“Definição SNIS”, 2021).

A água é um dos elementos essenciais à vida e às atividades do ser humano, tornando-se ao longo dos últimos anos tema de amplo debate. Se, por um lado, o acesso à água potável, por enquanto, não é igualitário, por outro, os níveis de desperdício e poluição crescem dia após dia. Atualmente, cerca de um bilhão de pessoas possuem alguma deficiência no acesso à água potável no mundo, e os diversos meios de captação d'água vêm sendo alvo de intensa exploração e degradação (ANDRADE SOBRINHO; BORJA, 2016). Tendo em vista o cenário atual de crise ambiental no planeta, é essencial que sejam desenvolvidas alternativas para a redução das perdas de água. Em particular, é de suma importância que as empresas, sejam elas públicas ou privadas, e seus respectivos sistemas de abastecimento de água, disponham de métodos eficazes para auxiliar na gestão das

perdas, visando a redução do volume perdido no percurso entre as unidades de distribuição e os locais de medição de consumo da população.

A perda de água é considerada um dos principais indicadores de desempenho operacional dos prestadores de serviço de abastecimento de água. Em todos os integrantes de um sistema de abastecimento de água podem ocorrer perdas, desde a captação até a distribuição (KUSTERKO et al., 2018). Entretanto, a dimensão dessas perdas varia de acordo com o perfil de cada unidade componente de um sistema de abastecimento de água, seja ele captação, tratamento, distribuição ou macro e micromedição.

A IWA – *The International Water Association* – fundada em 1998, classifica, levando em conta a sua natureza, as perdas de água como perdas reais (físicas) e perdas aparentes (comerciais) (BRASIL, 2020). As perdas reais são aquelas em que, de alguma forma, a água não chega ao cliente, ou seja, por mais que a ligação do indivíduo ou da empresa seja abastecida, determinado volume é perdido no trajeto, devido a, por exemplo, vazamentos em redes, adutoras, cavaletes ou ramais. As perdas aparentes são referentes à falta de medição do volume disponibilizado ao cliente. Ou seja, a água chega na ligação do indivíduo, porém o volume não é medido adequadamente pela empresa prestadora do serviço de abastecimento. Isso ocorre, por exemplo, devido à erros de medição, falta de hidrômetros, equipamentos mal aferidos, falhas do cadastro comercial, e, principalmente, ligações clandestinas, furto de água e demais fraudes (BRASIL, 2008).

Vale destacar que os volumes medidos pela empresa para limpeza de ruas, bombeiros, ou até mesmo a água utilizada para serviços operacionais, como descarga e limpeza de reservatórios, devem ser classificados como volume de serviço, de acordo com o SNIS. Já o volume medido pelas unidades e sedes para utilização dos seus funcionários, bem como àquele distribuído para comunidades de habitações populares onde haja hidrômetro, devem ser classificados como volume medido não faturado (BRASIL, 2018).

Conforme estudo realizado pelo Instituto Trata Brasil em 2020, quando os indicadores de perdas de água do Brasil foram comparados com patamares dos países desenvolvidos, verificou-se que ainda existe uma grande distância entre eles, até por motivos da diferença entre investimentos financeiros e equipamentos de alta tecnologia. Em termos de eficiência, a média dos índices de perdas no Brasil em 2018 foi de 39,02%, enquanto a média dos países desenvolvidos foi de 15%. Tendo em

vista o cenário do saneamento no Brasil, existe uma grande necessidade de se utilizar modelos de gestão eficientes que possam auxiliar na redução de perdas de água e o respectivo atingimento das metas contratuais propostas em momento de concessão deste serviço ao prestador.

Nos últimos anos, os esforços na área de saneamento foram concentrados cada vez mais na eficiência e na gestão otimizada dos sistemas de abastecimento de água. Uma extensa gama de trabalhos já foi conduzida, abordando os desafios e as soluções encontradas nesse contexto específico. Em particular, a gestão de sistemas de abastecimento de água tem sido objeto de estudo em diversas pesquisas, visando melhorias operacionais, ações para redução de perdas e garantia de fornecimento contínuo de água tratada para população. Neste contexto, temos o exemplo de Gouveia (2022) , que explorou técnicas inovadoras de *Machine Learning* para mitigar as perdas de água, com foco na predição de vazamentos em redes e ramais.

Diante do cenário supracitado, este trabalho buscou explorar, expandir os conhecimentos adquiridos, que contribuiu para o aprimoramento das estratégias de gestão de sistemas de abastecimento de água, com ênfase na aplicação dos conceitos de análise de séries temporais com *Machine Learning*. Utilizou-se técnicas de árvore de decisão para previsão do Indicador de Perdas na Distribuição (IPD), de forma similar a trabalhos que foram realizados para previsão de indicadores de consumo de informações relevantes para gestão de recursos naturais da sociedade (KAMTZIRIDIS, 2023).

De forma geral, a abordagem baseada em análise de séries temporais com *Machine Learning* foi de suma relevância para muitos problemas de previsão que envolvem um componente de tempo. As previsões foram feitas para novos dados quando o resultado real pode não ser conhecido até alguma data futura (BROWNLEE, 2020).

No contexto empresarial do saneamento, em que é possível obter informações de desempenho histórico, porém a incerteza, a complexidade operacional e as restrições de recursos são desafios constantes, a análise de séries temporais com *Machine Learning* surgiu como uma ferramenta poderosa. Ao proporcionar a capacidade de compreender e modelar os mecanismos estocásticos que dão origem a uma série observada e prever os valores futuros com base na história dessa série, essa abordagem possibilitou tomadas de decisões mais informadas e estratégicas,

considerando as variáveis envolvidas de maneira adaptável e flexível (CRYER; CHAN, 2009).

Neste trabalho, desenvolveu-se um modelo de análise de séries temporais com aplicações de *Machine Learning* utilizando da técnica de árvore de decisão, para que sirva de insumo para decisões de planejamento voltadas à redução de perdas na distribuição de água em um contexto de empresas brasileiras de saneamento, medidas por meio do indicador IN049 do SNIS, Índice de Perdas na Distribuição.

Delimitação da Pesquisa

A proposta deste trabalho foi realizar uma análise de séries temporais utilizando técnicas de *Machine Learning* para prever e otimizar a gestão do Índice de Perdas na Distribuição (IPD) em sistemas de abastecimento de água. O foco principal foi a análise de dados reais provenientes do Sistema Nacional de Informações sobre Saneamento (SNIS), combinados com os Índices de Desenvolvimento Humano Municipal (IDH-M) e Produto Interno Bruto (PIB) dos municípios do país.

A disponibilidade limitada de dados históricos no Sistema Nacional de Informações sobre Saneamento (SNIS) impactou a inclusão de alguns municípios na análise. Uma parte dos municípios não possuiu dados históricos para todos os indicadores presentes na base do SNIS, mesmo para aqueles considerados mais relevantes e integrados ao modelo proposto. Essa limitação implicou que a análise se concentrou nos municípios para os quais houve dados históricos suficientes nos indicadores selecionados. A ausência de informações para alguns municípios pôde influenciar a abrangência da pesquisa, e foi essencial reconhecer essa restrição ao interpretar os resultados.

Além da limitação proveniente de dados insuficientes, destacou-se a granularidade das informações contidas na base do SNIS, sendo dados oriundos de municípios e para uma melhor gestão para redução de perdas de água recomenda-se a divisão da região em setores menores para maior eficiência de medição e controle.

Apesar dessas limitações, a pesquisa prosseguiu com a análise de séries temporais utilizando técnicas de *Machine Learning* com base nos dados disponíveis. O modelo resultante teve uma aplicação mais específica, mas ainda assim, buscou

oferecer insights relevantes para a gestão do Índice de Perdas na Distribuição (IPD) em sistemas de abastecimento de água.

Recomendou-se, para trabalhos futuros, explorar maneiras de superar essas limitações, como estratégias para coleta e integração de dados em municípios com registros históricos incompletos e informações dos volumes e índices dos Distritos de Medição e Controle (DMC) de cada município. Essa iniciativa poderia contribuir para uma análise mais abrangente e representativa no contexto do saneamento brasileiro.

Estrutura do Trabalho

Este trabalho foi organizado em cinco capítulos, incluindo a Introdução e Conclusão.

Na Introdução foram abordadas as principais características, objetivos, relevância e delimitações do projeto.

O Capítulo 1 apresentou o Referencial Teórico. Neste capítulo foram explorados os fundamentos da análise de séries temporais, conceitos sobre aplicação de técnicas de *Machine Learning* e árvore de decisão, além de breve descrição referente ao algoritmo que será utilizado, o Extreme Gradiente Boosting (XGBoost).

O Capítulo 2 detalhou os procedimentos metodológicos adotados para conduzir a análise de séries temporais com *Machine Learning* neste projeto. Foram apresentadas as etapas desde a seleção e preparação dos dados até a implementação e avaliação dos modelos.

No Capítulo 3 foram realizadas simulações para explorar o comportamento das séries temporais ao longo do tempo. Essas simulações visaram extrair insights cruciais para a análise, permitindo uma compreensão mais profunda das tendências, padrões e relações presentes nos dados temporais. Neste capítulo foram detalhados todos os procedimentos realizados desde a coleta dos dados até a avaliação e visualização do resultado do modelo proposto.

O último capítulo dedicou-se às conclusões, abordando a aplicabilidade do método utilizado, o alcance dos objetivos estabelecidos e considerações finais. Também foram fornecidas sugestões para pesquisas futuras, visando a continuidade e aprimoramento do tema explorado.

1. REFERENCIAL TEÓRICO

Este capítulo teve como propósito discutir os conceitos de análise de séries temporais, *Machine Learning*, árvore de decisão e o algoritmo XGBoost, que foram utilizados neste trabalho, identificou as principais abordagens existentes e analisou suas vantagens e desvantagens ao lidar com cenários complexos e de grande escala. Dentro das técnicas abordadas de análise de séries temporais, foi dada ênfase aos métodos baseados em *Machine Learning*, destacando suas aplicações e implicações na tomada de decisão.

O pressuposto deste capítulo foi que o leitor possuía conhecimentos básicos sobre *Machine Learning*, análise de dados temporais, árvore de decisão e programação, além de métodos de treinamento e avaliação de modelos de *Machine Learning*.

A organização do capítulo seguiu da seguinte forma: na seção 1, foram descritos e discutidos os conceitos sobre a aplicação de análise de séries temporais, com base nas pesquisas realizadas até o momento; na seção 2, apresentaram-se os conceitos sobre a aplicação de técnicas de *Machine Learning*, na seção 3 foram abordados os conceitos sobre árvore de decisão e, por fim, na seção 4 uma breve explicação sobre o algoritmo de XGBoost.

1.1 Aplicação de Análise de Séries Temporais

O conteúdo desta seção, que contemplou os principais conceitos sobre a análise de séries temporais, tem por base os trabalhos de (2021), (2023), e (2004).

As séries temporais são caracterizadas por conjuntos sequenciais de dados, coletados em intervalos regulares ou irregulares ao longo do tempo. Formalmente, são representadas por vetores $x(t)$, onde t simboliza o tempo decorrido. Cada $x(t)$ é tratado como uma variável aleatória, conferindo propriedades estocásticas a esses conjuntos de dados.

Dentro do contexto das séries temporais, três características fundamentais surgem como elementos específicos: (1) a tendência, indicadora da direção geral dos dados ao longo do tempo, que proporciona *insights* sobre o comportamento

ascendente, descendente ou estacionário da série; (2) a sazonalidade, que revela padrões cíclicos que se repetem em intervalos específicos; e (3) a aleatoriedade, que reflete a variabilidade nos dados não explicada por tendência ou sazonalidade, representando o componente inesperado intrínseco à série temporal.

As aplicações práticas das séries temporais são vastas e abrangem diversas áreas de atuação. No âmbito econômico e financeiro, essas séries podem ser empregadas para prever preços de ações e analisar tendências econômicas. No setor de saúde, são utilizadas no monitoramento de pacientes e na previsão de possíveis surtos epidemiológicos. Além disso, no que tange à climatologia, contribuem para a análise de padrões climáticos e previsões meteorológicas.

Para extrair e analisar informações de séries temporais, alguns métodos tradicionais, como médias móveis, suavização exponencial e decomposição de séries temporais, são usualmente adotados. Tais técnicas fazem parte de ferramentas essenciais para compreender os padrões existentes nas séries temporais, proporcionando uma base para a análise exploratória e a aplicação de modelos preditivos.

Projetos de análise de séries temporais têm geralmente como etapas a definição do problema, a coleta de informações, a análise exploratória dos dados, a escolha e desenvolvimento de modelos, e a avaliação e utilização do modelo de previsão desenvolvido.

1.2 *Machine Learning*

Os conceitos abordados nesta seção sobre a aplicação de técnicas de *Machine Learning* foram baseados nos trabalhos de (2012), (2016), (2016) e (2015).

O progresso tecnológico de maneira geral, atrelado ao crescente surgimento de métodos inovadores, vem enriquecendo as alternativas de desenvolvimento de modelos utilizados nas análises e previsões de séries temporais. As abordagens modernas, em particular, as técnicas de *Machine Learning* (ML) vêm promovendo melhorias relevantes na compreensão e prognóstico desse tipo de dados dinâmicos.

A utilização de algoritmos de ML, como as Redes Neurais Recorrentes (RNNs), representa um avanço considerável em relação a outros métodos mais tradicionais

(ex., AutoRegressive Integrated Moving Average – ARIMA). Tais abordagens são capazes de capturar padrões complexos e não lineares presentes em séries temporais, proporcionando uma modelagem mais flexível e adaptável às variações intrínsecas dos dados.

Diferentemente dos métodos tradicionais, que muitas vezes impõem premissas específicas sobre a estrutura dos dados, os algoritmos de *Machine Learning* são capazes de aprender a partir dos padrões presentes numa série temporal. Isso os torna úteis em cenários em que a relação entre as variáveis é complexa ou sujeita a grandes mudanças ao longo do tempo.

Outro benefício das técnicas de *Machine Learning* é a capacidade de lidar com grandes volumes de dados e integrar informações de fontes variadas. Além disso, esses métodos oferecem a possibilidade de adaptação contínua, sendo capazes de reajustar seus modelos à medida que novos dados são obtidos.

No entanto, vale destacar que a aplicação de técnicas de ML em séries temporais envolve desafios. A escolha adequada do algoritmo, de acordo com o problema a ser solucionado, a configuração dos parâmetros, e a interpretação dos resultados são aspectos cruciais que demandam uma compreensão profunda das características específicas da série temporal e dos princípios associados aos algoritmos utilizados.

1.3 Árvore de Decisão

Os conceitos sobre árvores de decisão tratados nesta seção adveio dos trabalhos de (2002) e (2008).

Árvores de decisão são modelos de *Machine Learning* amplamente utilizados em casos de classificação e regressão. Elas consistem em estruturas hierárquicas que refletem testes sequenciais baseados nas características específicas dos dados, separando-os em subconjuntos distintos. Cada nó interno da árvore representa um teste sobre uma dada característica, enquanto os ramos conduzem a subárvores subsequentes ou a folhas que contêm as decisões finais, ou seja, o valor previsto para dados que têm as características associadas aos resultados dos testes anteriores. As árvores de decisão são modelos com grande interpretabilidade, possibilitando que os

usuários e analistas compreendam facilmente o processo de tomada de decisão definido por uma árvore.

As árvores de decisão evoluíram com o passar dos anos, dando origem a variantes, como florestas aleatórias (*Random Forests*) e técnicas baseadas em *gradient boosting*. Tais modelos combinam múltiplas árvores para otimizar o desempenho preditivo e lidar com questões complexas, como, por exemplo, o *overfitting* ou sobreajuste. O processo para construção de uma árvore de decisão envolve a seleção de características relevantes, a definição de critérios de divisão dos dados, e a poda da árvore para evitar uma complexidade excessiva e mitigar o *overfitting*.

As árvores de decisão são bastante utilizadas, com aplicações em áreas tão diversas quanto diagnósticos médicos e sistemas de recomendação. As técnicas baseadas em árvores são capazes de lidar com dados categóricos e numéricos, sendo uma poderosa ferramenta para a resolução de problemas de previsão complexos.

1.4 O Algoritmo XGBoost

O conteúdo desta seção foram baseados nos trabalhos de (2016) e (2024).

O algoritmo Extreme Gradient Boosting (XGBoost) representa um grande avanço dentro das técnicas de *Machine Learning*, em especial nos problemas relacionados a classificação e regressão, e vem sendo bastante utilizado na resolução de tarefas preditivas complexas. O XGBoost tem sua base no método de impulsionamento de gradientes (*gradiente boosting*). Ele combina diversas árvores de decisão fracas (com poucas quebras) em um modelo com melhor desempenho, destacando-se por sua capacidade de lidar com dados desbalanceados, com alta dimensionalidade, e com ruído.

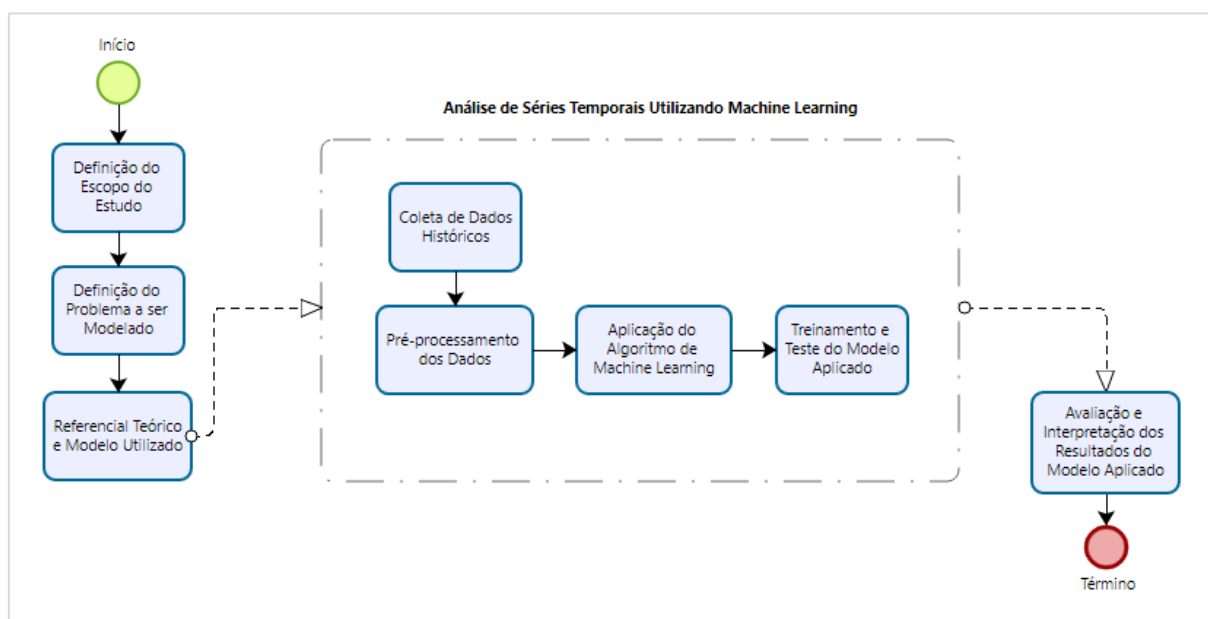
Uma característica do XGBoost que o distingue das demais técnicas é a inclusão de métodos de regularização, que ajudam a prevenir o sobreajuste e a aprimorar a generalização do modelo. Além disso, o algoritmo integra funções de perda customizáveis, permitindo uma otimização orientada a tipos de problemas diversos.

A eficiência computacional do XGBoost, aliada à sua capacidade de lidar com uma grande variedade de métricas de desempenho, o posiciona como uma ferramenta extremamente relevante dentro do conjunto de técnicas de *Machine Learning*.

2. PROCEDIMENTOS METODOLÓGICOS

Este capítulo teve por objetivo detalhar os procedimentos metodológicos adotados no desenvolvimento deste trabalho. A figura a seguir mostrou na forma de um fluxograma a metodologia que foi aplicada. Suas etapas foram descritas nas próximas seções.

Figura 1 – Fluxograma de procedimentos metodológicos



Fonte: O autor, 2024.

2.1 Definição do Escopo do Estudo

Para definir o escopo do estudo foi necessário responder as seguintes perguntas:

- i. Qual problema foi analisado?

Além do notório problema referente a crise hídrica mundial e a necessidade de modelos, teóricos e práticos, que auxiliem na redução de perdas de água, o principal problema que foi analisado nesse estudo foi a capacidade de previsão do Índice de Perdas na Distribuição (IPD) das empresas de saneamento.

ii. Como o estudo foi desenvolvido?

O estudo foi realizado por meio da coleta e tratamento dos dados históricos do SNIS dos municípios brasileiros, fonte oficial no âmbito federal para informações de saneamento, IBGE e Atlas Brasil, e seu posterior tratamento e utilização, com a aplicação do algoritmo XGBoost de *Machine Learning*, onde buscou-se prever os índices de perdas de cada município.

iii. Quais foram os objetivos e as limitações do estudo?

O estudo possuiu o objetivo principal de utilizar o algoritmo XGBoost de *Machine Learning* para prever o indicador IN049 do SNIS (IPD), levando em consideração as limitações da base histórica extraída do SNIS citadas no capítulo de introdução.

iv. Qual foi a importância desse estudo? Como pudemos associar este trabalho com a prática?

O trabalho possuiu grande relevância, pois teve como tema de estudo a distribuição de água, um recurso fundamental para sociedade, que foi atrelado a técnicas de aprendizado de máquina para apoiar no planejamento dos prestadores do serviço de saneamento. Além das empresas de saneamento brasileira, este estudo poderá ser replicado na análise de outros indicadores que dizem respeito ao consumo e utilização de recursos naturais.

2.2 Definição do Problema

Quando pensamos nos principais recursos indispensáveis à vida humana, logo nos vem à mente a água. Tendo em vista que este é um recurso finito do meio ambiente, nós, seres humanos, temos a obrigação de sempre criar alternativas para garantir que este recurso tão valioso seja utilizado de forma correta. Dessa maneira, as empresas que prestam serviço de saneamento devem estar sempre atentas quanto ao volume perdido no seu sistema de distribuição e medição. O Índice de Perdas na

Distribuição (IPD) é, portanto, considerado um dos principais indicadores de desempenho operacional desses tipos de prestadores, sejam eles privados ou públicos.

As áreas abastecidas pelos prestadores podem ser divididas em municípios, distritos e outras subdivisões. Definem-se para essas áreas objetivos e ações para redução do índice perdas de água, seja na distribuição ou na medição dos consumidores finais.

Diversos fatores influenciam dentro de um sistema de abastecimento de água. O IPD é o índice principal para metrificar a eficiência do sistema e até mesmo avaliar a própria empresa prestadora de serviço.

Anualmente, as empresas devem reportar ao SNIS as informações do seu respectivo sistema de abastecimento de água. Esses dados poderiam ser utilizados para planejar ou até mesmo prever cenários futuros. Este trabalho buscou tratar este problema, utilizando de técnicas de *Machine Learning* para prever o indicador IPD, com base na análise de séries temporais, de acordo com a base histórica proveniente do SNIS.

2.3 Referencial Teórico e Modelo Utilizado

As referências encontradas que foram utilizadas neste estudo abordaram conceitos fundamentais sobre a análise de séries temporais, técnicas de *Machine Learning* e árvores de decisão, além do algoritmo XGBoost que foi utilizado. Tais abordagens estão sendo cada vez mais estudadas e aplicadas na previsão de informações em diversos segmentos do mercado. Ao compreender tendências, sazonalidades e padrões complexos presentes nos dados ao longo do tempo, essas técnicas proporcionaram uma base sólida para análises exploratórias e modelos preditivos. A interpretabilidade das árvores de decisão e a adaptabilidade do XGBoost foram aproveitadas em setores complexos como o saneamento, em específico o sistema de abastecimento de água de um município, o que permitiu uma abordagem consistente na tomada de decisões para a gestão eficiente dos indicadores desse sistema.

2.4 Análise de Séries Temporais Utilizando *Machine Learning*

A fim de melhor organizar, separamos esta seção nos tópicos já ilustrados no fluxograma acima.

2.4.1 Coleta de Dados Históricos

Contemplou uma das etapas fundamentais desse trabalho, obter os dados e traduzi-los em informações úteis para criar um algoritmo de *Machine Learning* capaz de prever cenários futuros. A coleta dessas informações foi possível de ser realizada através do levantamento histórico obtido em relatórios divulgados no site do SNIS.

A base de dados do SNIS é uma fonte essencial para análise e monitoramento do panorama do saneamento básico no Brasil. Para este estudo, foi realizada uma consulta ao site do SNIS, com filtragem dos anos de 1998 a 2021, visando obter um conjunto abrangente de dados ao longo do tempo.

Além das informações e indicadores que foram obtidos no site do SNIS, também foram coletados dados complementares do Instituto Brasileiro de Geografia e Estatística (IBGE), para obter o Produto Interno Bruto (PIB) e no site Atlas Brasil para o Índice de Desenvolvimento Humano Municipal (IDHM), com intuito de fornecer uma visão mais abrangente do contexto socioeconômico dos municípios analisados.

2.4.2 Pré-processamento dos Dados

No processo de pré-processamento de dados foram realizados ajustes na base de dados, além da criação de um dicionário de dados para descrever cada informação nela contida, visando uma melhor organização e compreensão das variáveis.

Além disso, foram identificados e realizados possíveis tratamentos na base de dados que possam impactar a análise, como por exemplo municípios que não possuem massa de dados significativa, valores ausentes e dados inconsistentes referente aos indicadores deste estudo. As informações de alguns municípios foram excluídas da base de dados para garantir a qualidade e a consistência dos resultados.

2.4.3 Aplicação do Algoritmo de *Machine Learning*

Para a análise de séries temporais, foi escolhida a técnica XGBoost, que é um poderoso algoritmo de *Machine Learning*, se destacando em várias tarefas de modelagem preditiva, incluindo previsão de séries temporais.

Conforme já mencionado, o XGBoost é um método de aprendizagem conjunto que combina as previsões de vários modelos fracos (árvores de decisão) para criar um forte modelo preditivo. O XGBoost é conhecido por sua escalabilidade, velocidade e capacidade de lidar com relacionamentos complexos nos dados (SHARMA, 2024).

2.4.4 Treinamento e Teste do Modelo Aplicado

Para avaliar o desempenho do modelo XGBoost, foi necessário particionar os dados da série temporal em conjuntos de treinamento e teste. O conjunto de treinamento foi utilizado para o treinamento do modelo, e o conjunto de testes, para a avaliação e estimativa de seu desempenho para previsões baseadas em novos dados coletados. Preservamos a ordem temporal das observações que foi crucial ao dividir os dados.

2.5 Avaliação e Interpretação dos Resultados do Modelo Aplicado

Após obter o resultado do modelo aplicado, foram empregadas métricas fundamentais para avaliação de desempenho. Detalhamos minuciosamente essas métricas, contextualizando sua aplicação no nosso modelo de previsão. Essa análise aprofundada foi fundamental para compreender o quão bem o modelo se ajusta aos dados e contribui para a consecução dos objetivos propostos nesta pesquisa.

3. DESENVOLVIMENTO DO MODELO DE *MACHINE LEARNING*

No capítulo de desenvolvimento do modelo, aplicamos um passo a passo na construção do nosso algoritmo. Nesta parte do projeto, cada seção representou as etapas da elaboração do modelo, desde a coleta de dados, até a abordagem específica na interpretação e visualização dos resultados obtidos.

Como mencionado anteriormente, iniciamos com a Coleta dos Dados, extraindo as informações que consideramos relevantes e que serviram como base para nosso modelo. Em seguida, a seção de Análise Exploratória e Tratamento dos Dados identificamos padrões, outliers e tratamos eventuais lacunas nos dados. A Divisão dos Dados em Treinamento e Teste foi realizada em seguida.

Dando continuidade, foi feita a Codificação de Variáveis Categóricas e o Treinamento do Modelo. Na etapa de treinamento, realizou-se o Ajuste do Modelo e Otimização de Hiperparâmetros do XGBoost. Esta etapa buscou os melhores hiperparâmetros para a construção do modelo. Dedicamos uma seção à Avaliação do Desempenho do Modelo, onde utilizamos métricas fundamentais, como o erro absoluto médio e a raiz do erro quadrático médio, para analisar a precisão do modelo. Finalmente, a seção de Visualização das Previsões do Modelo comparou de forma gráfica os valores previstos e os valores reais, e avaliamos o desempenho do modelo.

3.1 Coleta dos Dados

O processo de coleta de dados para o desenvolvimento do modelo de análise de séries temporais reuniu informações de diferentes fontes, sendo a principal delas a base histórica do Sistema Nacional de Informações sobre Saneamento (SNIS). Utilizamos o site oficial do SNIS, aplicando o filtro dos anos de 1998 a 2021. Além disso, foram selecionados apenas os indicadores específicos relacionados à população e abastecimento de água (ex., POP_TOT, AG002, AG003 etc.). Abaixo descrevemos o significado de cada indicador extraído do site (“SNIS - Série Histórica”, 2024):

- **POP_TOT** (população total do município do ano de referência): Refere-se à população total do município no ano de referência, ou seja, o número total de habitantes na área em questão.
- **AG002** (quantidade de ligações ativas de água): Representa a quantidade de ligações ativas de água no período atual, indicando o número de conexões de água em funcionamento no município.
- **AG003** (quantidade de economias ativas de água): Indica a quantidade de economias ativas de água no período anterior ao anterior, ou seja, o número de unidades consumidoras de água em operação no período passado.
- **AG005** (extensão da rede de água): Refere-se à extensão da rede de água no período atual, representando o comprimento total da rede de distribuição de água na área.
- **AG006** (volume de água produzido): Indica o volume de água produzido no período atual, ou seja, a quantidade total de água fornecida pela empresa de abastecimento no período considerado.
- **AG010** (volume de água consumido): Representa o volume de água consumido no período atual, indicando a quantidade total de água utilizada pelos consumidores durante o período em questão.
- **AG011** (volume de água faturado): Indica o volume de água faturado no período atual, ou seja, a quantidade total de água que foi cobrada dos consumidores durante o período considerado.
- **AG021** (quantidade de ligações totais de água): Refere-se à quantidade total de ligações de água no período atual, representando o número total de conexões de água existentes no município.
- **IN013** (índice de perdas no faturamento): Representa o índice de perdas no faturamento no período atual, indicando a porcentagem de água que é perdida durante o processo de faturamento e cobrança.
- **IN049** (índice de perdas na distribuição): Indica o índice de perdas na distribuição no período atual, representando a porcentagem de água perdida durante o processo de distribuição.

- **IN050** (Índice bruto de perdas lineares): Refere-se ao índice bruto de perdas lineares no período atual, representando a porcentagem de água perdida devido a vazamentos e problemas na rede de distribuição.
- **IN051** (índice de perdas por ligação): Indica o índice de perdas por ligação no período atual, representando a porcentagem de água perdida por conexão de água durante o período considerado.
- **IN052** (índice de consumo de água): Representa o índice de consumo de água no período atual, indicando a eficiência no uso da água pelos consumidores durante o período em questão.
- **IN053** (Consumo médio de água por economia): Refere-se ao consumo médio de água por economia no período atual, representando a quantidade média de água utilizada por unidade consumidora durante o período considerado.

Adicionalmente, para enriquecer a base de dados, adicionamos informações socioeconômicas relevantes. O Produto Interno Bruto (PIB) dos municípios foi obtido diretamente do site do Instituto Brasileiro de Geografia e Estatística (IBGE). Da mesma maneira, extraímos o PIB per capita (“Produto Interno Bruto dos Municípios - IBGE”, 2024):

- **PIB** (Produto Interno Bruto): Indica o Produto Interno Bruto (PIB) do município no período atual, representando o valor total de todos os bens e serviços produzidos na área durante o período em questão.
- **PIB_PC** (Produto Interno Bruto per capita): Representa o Produto Interno Bruto per capita do município no período atual, ou seja, o valor médio do PIB por habitante na área.

Complementando esses dados, incluímos o Índice de Desenvolvimento Humano de cada município (IDH-M), adquirido de o site Atlas Brasil (“Índice de Desenvolvimento Humano Município - Atlas BR”, 2024).

- **IDH-M** (Índice de Desenvolvimento Humano Municipal): Refere-se ao Índice de Desenvolvimento Humano Municipal, um indicador composto que

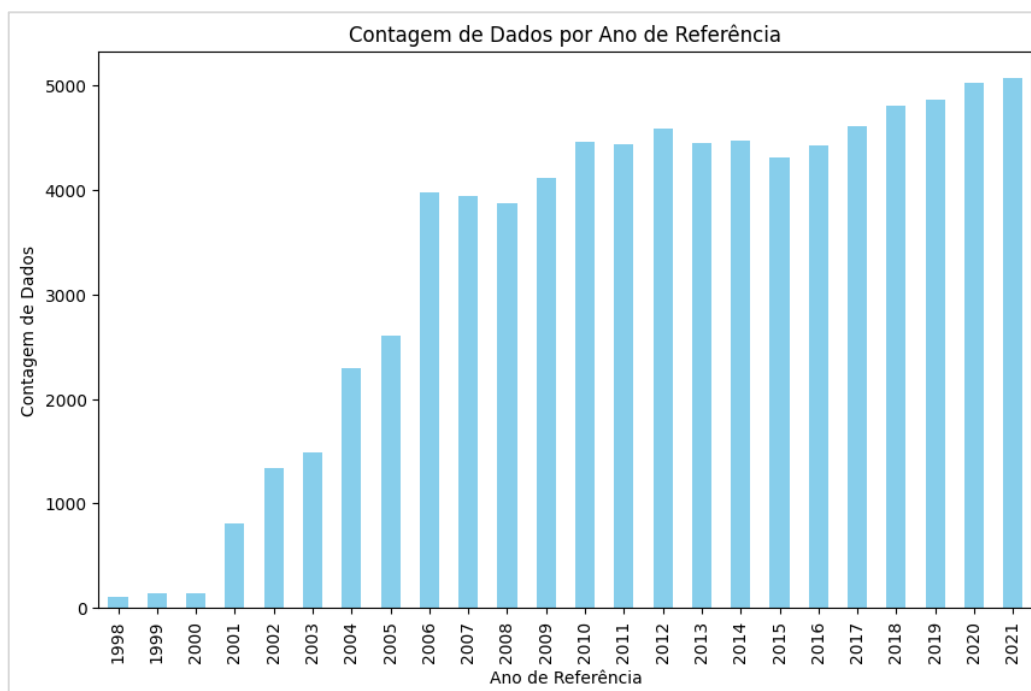
mede o desenvolvimento humano com base em três dimensões: saúde, educação e padrão de vida.

3.2 Análise Exploratória e Tratamento dos Dados

Uma etapa crucial no desenvolvimento de modelos de análise de séries temporais utilizando *Machine Learning* é a análise exploratória inicial seguida do tratamento dos dados coletados. Essa fase compreendemos e interpretamos as características dos dados, identificando possíveis padrões e anomalias, e garantindo que os dados estejam prontos para serem utilizados no processo de modelagem.

Inicialmente, realizou-se a exclusão de anos com dados insuficientes, ou seja, dados de indicadores que representaram números abaixo 3500 itens em nossa base para análise. Este processo envolveu a remoção de anos em que a quantidade de informações disponíveis não era suficientemente para os critérios estabelecidos neste estudo, considerando apenas os dados a partir do ano de 2006.

Figura 2 – Pré-processamento de dados: Informações e indicadores por ano



Fonte: O autor, 2024.

Após a exclusão dos anos com dados insuficientes, procedeu-se com a inclusão de três anos anteriores aos dados disponíveis. Cabe ressaltar que esses dados já estavam disponíveis na própria base de dados. Essa decisão foi tomada com o intuito de ampliar a série temporal e permitir uma análise mais precisa ampliando a granularidade de dados no estudo.

Outro passo importante foi a exclusão de dados ausentes, excluindo casos em que municípios não apresentavam informações disponíveis ou indicadores apresentavam valores negativos ou acima de 100%. Essa filtragem foi essencial para garantir a integridade dos dados e evitar distorções nos resultados.

Por fim, as variáveis da base foram renomeadas de acordo com as suas descrições. Essa harmonização foi realizada visando facilitar a compreensão e interpretação dos dados por parte dos pesquisadores e leitores. Os novos nomes foram criados de forma a tornar as informações mais claras e consistentes, contribuindo para uma análise mais precisa e coerente.

Para continuar a análise exploratória, o conjunto de dados consolidado foi carregado em um ambiente de programação Python, com a biblioteca Pandas.

O código a seguir ilustra como essa operação foi realizada:

Figura 3 – Carregamento dos dados em python com a biblioteca pandas

```
# Caminho para o arquivo CSV
caminho_arquivo = caminho_arquivo = r'C:\Users\55219\OneDrive - Universidade do Estado do Rio de Janeiro\PROJETO DE GRADUAÇÃO\Pr

# Carregue os dados em um DataFrame do Pandas, especificando a vírgula como separador decimal
df = pd.read_csv(caminho_arquivo, sep=',')

# Exiba as primeiras linhas dos dados para verificar se foram carregados corretamente
df.head()
```

	ano_0	ano_1	ano_2	ano_3	Mun	UF	pop_tot_0	pop_tot_1	pop_tot_2	pop_tot_3	...	pi_b_2	pi_b_3	pi_b_pc_0	pi_b_pc_1	pi_b_pc_2	pi_b_pc_3	idm
0	2021	2020	2019	2018	Acrelândia	AC	15721	15490	15258	15020	...	253152	253138	19525	19525	18594	18853	
1	2020	2019	2018	2017	Acrelândia	AC	15490	15258	15020	14386	...	253138	229844	19525	18594	18853	15985	
2	2019	2018	2017	2016	Acrelândia	AC	15258	15020	14386	14120	...	229844	239810	18594	18853	15985	18970	
3	2018	2017	2016	2015	Acrelândia	AC	15020	14386	14120	13889	...	239810	212981	18853	15985	18970	15355	
4	2017	2016	2015	2014	Acrelândia	AC	14386	14120	13889	13813	...	212981	207822	15985	18970	15355	15288	

5 rows x 74 columns

Fonte: O autor, 2024.

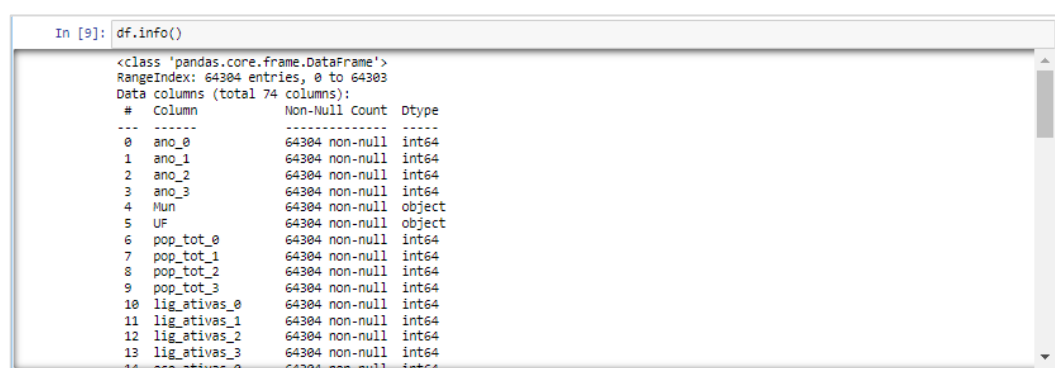
Após a execução desse código, os cinco primeiros registros do conjunto de dados foram exibidos, fornecendo uma visão inicial das variáveis e seus valores.

Após a exibição das primeiras linhas do conjunto de dados, foi realizada uma análise mais detalhada da sua estrutura, utilizando o método info() da biblioteca

Pandas. Este método fornece informações essenciais sobre o conjunto de dados, incluindo o número total de entradas, o tipo de dados de cada coluna e a quantidade de valores não nulos.

O conjunto de dados consiste em 74 colunas, representando diferentes métricas e indicadores ao longo de quatro anos consecutivos.

Figura 4 – Análise exploratória: contagem de entradas e tipos de dados



```
In [9]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64304 entries, 0 to 64303
Data columns (total 74 columns):
#   Column              Non-Null Count  Dtype  
---  --
0   ano_0                64304 non-null  int64  
1   ano_1                64304 non-null  int64  
2   ano_2                64304 non-null  int64  
3   ano_3                64304 non-null  int64  
4   Mun                  64304 non-null  object  
5   UF                   64304 non-null  object  
6   pop_tot_0            64304 non-null  int64  
7   pop_tot_1            64304 non-null  int64  
8   pop_tot_2            64304 non-null  int64  
9   pop_tot_3            64304 non-null  int64  
10  lig_ativas_0         64304 non-null  int64  
11  lig_ativas_1         64304 non-null  int64  
12  lig_ativas_2         64304 non-null  int64  
13  lig_ativas_3         64304 non-null  int64  
14  lig_ativas_4         64304 non-null  int64
```

Fonte: O autor, 2024.

Aqui estão alguns pontos importantes sobre os resultados gerados:

- **Formato do DataFrame:** O DataFrame é uma estrutura de dados tabular do tipo pandas, com um total de 64.304 entradas (ou observações) distribuídas em 74 colunas.
- **Colunas e Tipos de Dados:** As colunas representam diferentes variáveis, incluindo indicadores econômicos, sociais e demográficos, bem como dados de consumo e produção de água. Os tipos de dados das colunas incluem int64 (para variáveis inteiras), float64 (para variáveis numéricas com casas decimais) e object (para variáveis de texto).
- **Contagem de Valores Não Nulos:** Para cada coluna, a contagem de valores não nulos indica o número de observações válidas presentes. Essa informação é crucial para identificar a presença de dados ausentes e orientar as estratégias de tratamento posterior, se necessário.

Além da análise inicial dos dados, também foi realizada uma etapa de preparação dos dados, que inclui a conversão de algumas variáveis para o tipo de

dados categórico. Isso foi realizado de modo a garantir que seja possível a leitura desse tipo de variável pelo algoritmo XGBoost.

As seguintes operações foram aplicadas para converter as variáveis 'Mun' (município) e 'UF' (unidade federativa) que eram do tipo “object” para o tipo de dados categórico:

Figura 5 – Conversão de variáveis

```
In [10]: df['Mun'] = df['Mun'].astype('category')
         df['UF'] = df['UF'].astype('category')
```

Fonte: O autor, 2024.

Após serem analisados, pré-processados e tratados de acordo com os procedimentos estabelecidos, os dados agora estão prontos para serem utilizados na construção do modelo analítico.

3.3 Divisão dos Dados em Treinamento e Teste

Após a análise exploratória, pré-processamento e tratamento dos dados, o próximo passo foi dividir o conjunto de dados em conjuntos para treinamento e teste.

Para cada conjunto (treinamento e teste), os dados foram separados em conjuntos de recursos (x), que incluem todas as variáveis independentes utilizadas para fazer previsões, e conjuntos de alvo (y), que representam a variável dependente que está sendo prevista.

Foi importante verificar o tamanho dos conjuntos de treinamento e teste para garantir que a divisão tenha sido realizada corretamente e que a proporção entre eles tenha sido consistente com o planejado.

Essa divisão em conjuntos de treinamento e teste permitiu uma avaliação do desempenho do modelo e ajudou a evitar problemas de *overfitting*, garantindo que o modelo seja capaz de generalizar bem para novos dados.

Figura 6 – Divisão em dados de treinamento e teste

```

In [24]: # Separar os dados em conjuntos de treinamento e teste (80% treinamento, 20% teste)
train_size = int(len(df) * 0.8)
train_data, test_data = df[:train_size], df[train_size:]

# Separar os conjuntos de recursos (X) e alvo (y) para treinamento e teste
colunas_recursos = ['ano_0', 'ano_1', 'ano_2', 'ano_3', 'Mun', 'UF',
                    'pop_tot_0', 'pop_tot_1', 'pop_tot_2', 'pop_tot_3',
                    'lig_ativas_0', 'lig_ativas_1', 'lig_ativas_2', 'lig_ativas_3',
                    'eco_ativas_0', 'eco_ativas_1', 'eco_ativas_2', 'eco_ativas_3',
                    'ext_rede_0', 'ext_rede_1', 'ext_rede_2', 'ext_rede_3',
                    'vol_prod_0', 'vol_prod_1', 'vol_prod_2', 'vol_prod_3',
                    'vol_cons_0', 'vol_cons_1', 'vol_cons_2', 'vol_cons_3',
                    'vol_fatu_0', 'vol_fatu_1', 'vol_fatu_2', 'vol_fatu_3',
                    'lig_tot_ativ_0', 'lig_tot_ativ_1', 'lig_tot_ativ_2', 'lig_tot_ativ_3',
                    'ind_perd_fatu_0', 'ind_perd_fatu_1', 'ind_perd_fatu_2', 'ind_perd_fatu_3',
                    'ind_perd_distr_0', 'ind_perd_distr_1', 'ind_perd_distr_2', 'ind_perd_distr_3',
                    'ind_perd_lin_0', 'ind_perd_lin_1', 'ind_perd_lin_2', 'ind_perd_lin_3',
                    'ind_perd_lig_0', 'ind_perd_lig_1', 'ind_perd_lig_2', 'ind_perd_lig_3',
                    'ind_cons_0', 'ind_cons_1', 'ind_cons_2', 'ind_cons_3',
                    'cons_med_eco_0', 'cons_med_eco_1', 'cons_med_eco_2', 'cons_med_eco_3',
                    'pib_0', 'pib_1', 'pib_2', 'pib_3',
                    'pib_pc_0', 'pib_pc_1', 'pib_pc_2', 'pib_pc_3',
                    'idhm_0', 'idhm_1', 'idhm_2', 'idhm_3']

colunas_alvo = ['ind_perd_distr_0', 'ind_perd_distr_1', 'ind_perd_distr_2', 'ind_perd_distr_3']

X_train = train_data[colunas_recursos]
y_train = train_data[colunas_alvo]

X_test = test_data[colunas_recursos]
y_test = test_data[colunas_alvo]

# Verificar o tamanho dos conjuntos de treinamento e teste
print("Número de amostras no conjunto de treinamento:", len(X_train))
print("Número de amostras no conjunto de teste:", len(X_test))

Número de amostras no conjunto de treinamento: 51443
Número de amostras no conjunto de teste: 12861

```

Fonte: O autor, 2024.

Após a divisão dos dados em conjuntos de treinamento e teste, estávamos prontos para avançar no próximo passo da nossa análise, que envolveu o ajuste do modelo e otimização de hiperparâmetros do problema em questão.

3.4 Otimização de Hiperparâmetros e Treinamento do Modelo

Com os conjuntos de treinamento e teste prontos, agora realizamos o treinamento do modelo, juntamente com a otimização de seus hiperparâmetros. Neste caso, utilizamos o algoritmo XGBoost, conhecido por sua eficácia em problemas de regressão.

Em seguida, definimos a grade de hiperparâmetros a serem testados e configuramos a estratégia de validação cruzada. Criamos o objeto GridSearchCV para realizar a busca pelos melhores hiperparâmetros. Finalmente, ajustamos o melhor modelo aos dados de treinamento completos, e avaliamos seu desempenho com os dados de teste.

Figura 7 – Melhores hiperparâmetros encontrados

```

In [20]: xgb_model = xgb.XGBRegressor(enable_categorical=True)

# Definindo a grade de hiperparâmetros a serem testados
param_grid = {
    'n_estimators': [100, 200, 300], # Número de árvores de decisão no modelo
    'max_depth': [3, 6, 9], # Profundidade máxima da árvore
    'learning_rate': [0.01, 0.1, 0.3] # Taxa de aprendizado do modelo
}

# Definindo a estratégia de validação cruzada (neste caso, usando KFold com 5 folds)
kfold = KFold(n_splits=5, shuffle=True, random_state=42)

# Criando o objeto GridSearchCV
grid_search = GridSearchCV(estimator=xgb_model, param_grid=param_grid, cv=kfold, scoring='neg_mean_squared_error', refit=True)

# Ajustando o modelo
grid_result = grid_search.fit(X_train, y_train)

# Obtendo os melhores hiperparâmetros
best_params = grid_result.best_params_
print("Melhores hiperparâmetros encontrados:", best_params)

# Avaliando o desempenho do modelo com os melhores hiperparâmetros nos dados de teste
best_model = grid_result.best_estimator_
test_score = best_model.score(X_test, y_test)
print("Desempenho do modelo nos dados de teste:", test_score)

Melhores hiperparâmetros encontrados: {'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 300}
Desempenho do modelo nos dados de teste: 0.9999305751059115

```

Fonte: O autor, 2024.

Este processo de treinamento ajustou os parâmetros do modelo aos dados de treinamento e produziu um modelo capaz de fazer previsões sobre novos conjuntos de dados e então, foi possível realizar a avaliação e interpretação do modelo de acordo com as métricas de desempenho exploradas na próxima seção.

3.5 Avaliação e Interpretação do Modelo com as Métricas de Desempenho

Após o treinamento do modelo de regressão XGBoost, foi importante avaliar o desempenho do modelo com os dados de teste usando métricas apropriadas. As principais métricas utilizadas foram o Erro Quadrático Médio (MSE), o Coeficiente de Determinação (R^2), Raiz do Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE).

O Erro Quadrático Médio (MSE) é uma medida da média dos quadrados dos erros entre os valores previstos e os valores reais. Quanto menor o valor do MSE, melhor é o desempenho do modelo. No caso deste modelo, o MSE foi calculado como 0.02203, indicando que as previsões do modelo estão muito próximas dos valores reais.

O Coeficiente de Determinação (R^2), também conhecido como R-squared, indica a proporção da variância na variável dependente que é explicada pelo modelo, a partir das variáveis preditoras (x). Um valor mais próximo de 1,0 indica um melhor

ajuste do modelo aos dados. Neste caso, o R^2 foi calculado como 0.99993, o que sugere que o modelo explica quase toda a variabilidade dos dados de teste.

A Raiz do Erro Quadrático Médio (RMSE) é a raiz quadrada do MSE e fornece uma interpretação mais intuitiva dos erros na unidade original da variável-alvo (IPD). Um RMSE mais baixo indica um melhor desempenho do modelo. No caso deste modelo, o RMSE foi calculado como 0.14842, o que confirma a precisão das previsões.

O Erro Médio Absoluto (MAE), é uma métrica de avaliação de desempenho de um modelo de regressão que mede a média das diferenças absolutas entre os valores observados (reais) e os valores previstos pelo modelo. O MAE foi calculado como 0.0698 sugerindo que o modelo tem um bom desempenho em relação à precisão das previsões.

Figura 8 – Resultado das métricas de desempenho

```
In [23]: # Fazer previsões nos dados de teste
y_pred = best_model.predict(X_test)

# Calcular o MSE
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)

# Calcular o RMSE
rmse = np.sqrt(mse)
print("Root Mean Squared Error (RMSE):", rmse)

# Calcular o R²
r2 = r2_score(y_test, y_pred)
print("R² Score:", r2)

# Calcular o MAE
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error (MAE):", mae)

Mean Squared Error (MSE): 0.0220299649517921
Root Mean Squared Error (RMSE): 0.14842494720158098
R² Score: 0.9999305751059115
Mean Absolute Error (MAE): 0.0698206787157185
```

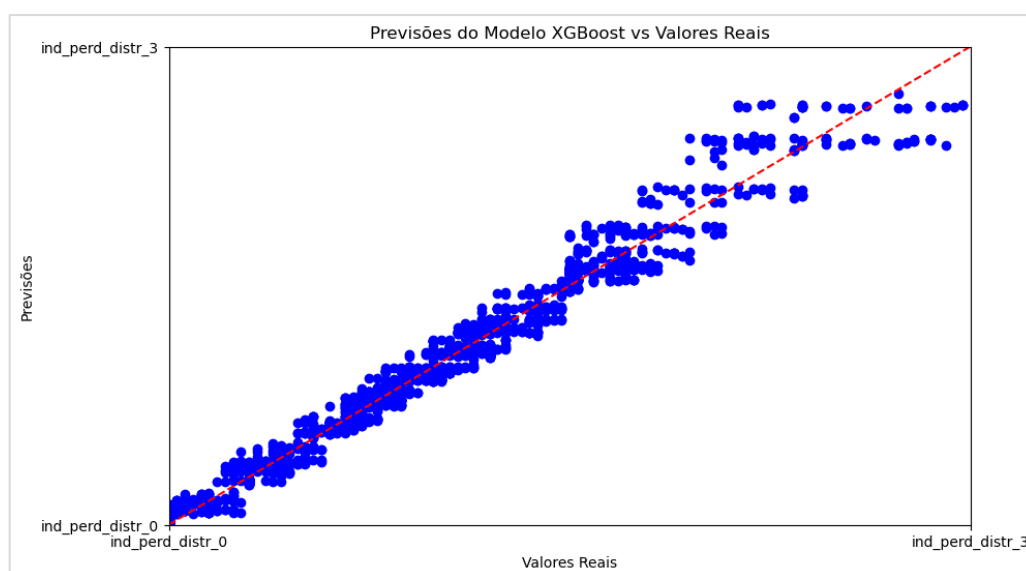
Fonte: O autor, 2024.

Esses resultados indicaram que o modelo de regressão XGBoost treinado apresentou um desempenho excepcionalmente bom na tarefa de previsão com os dados de teste fornecidos. As previsões foram muito próximas dos valores reais, e o modelo explicou quase toda a variabilidade dos dados. Em seguida utilizamos visualizações gráficas para corroborar com esta avaliação do resultado obtido no modelo.

3.6 Visualização das Previsões do Modelo

A visualização das previsões do modelo em comparação com os valores reais foi uma maneira de entender de forma ilustrativa o desempenho do modelo. No gráfico abaixo, os valores reais estão representados no eixo x, enquanto as previsões do modelo estão no eixo y. Cada ponto no gráfico representou uma observação nos dados de teste. Idealmente, os pontos deviam se alinhar em torno da linha de referência (linha vermelha tracejada), o que indicaria que as previsões do modelo estão muito próximas dos valores reais.

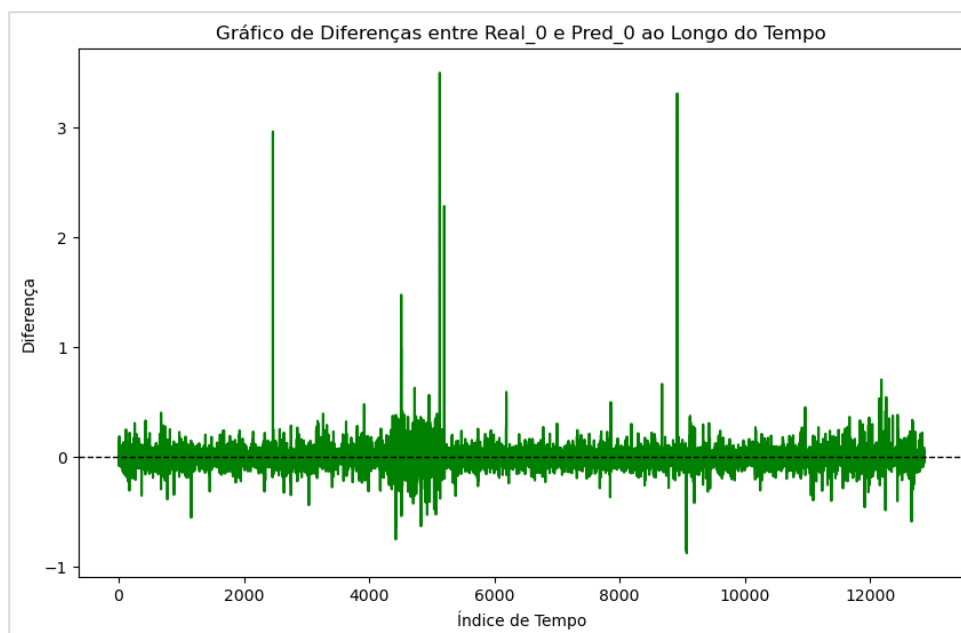
Figura 9 – Gráfico de dispersão entre previsões e valores reais



Fonte: O autor, 2024.

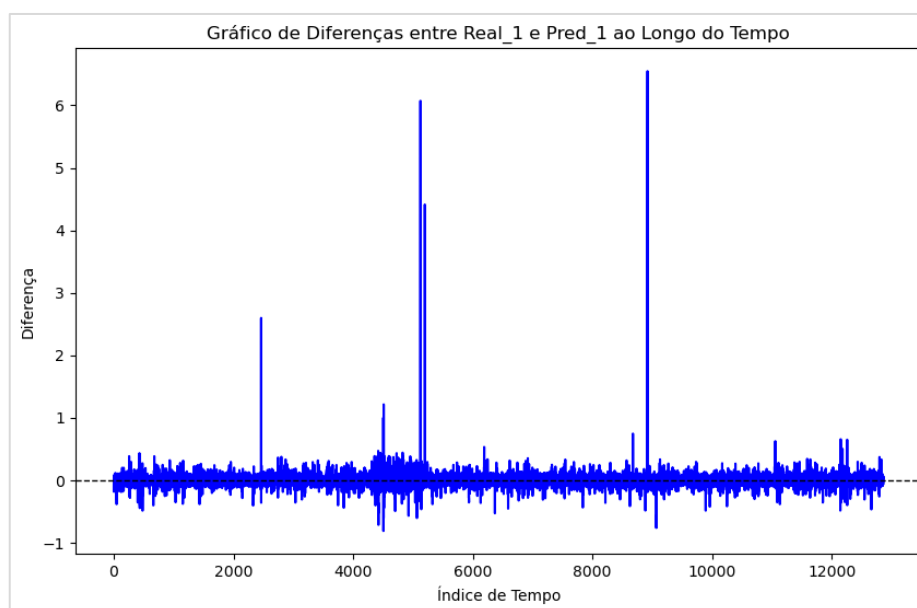
Já nos gráficos a seguir, foi fornecido a diferença entre os valores reais e as previsões em cada ano de previsão. Ele foi útil para visualizar os erros do modelo em diferentes pontos. A diferença entre os valores reais e as previsões foi plotada ao longo do tempo. Se os valores estivessem próximos de zero, indicaria que o modelo fez previsões precisas. Por outro lado, valores positivos indicariam que o modelo subestimou os valores reais, enquanto valores negativos indicariam uma superestimação.

Figura 10 – Gráfico de diferença entre valores reais e previsões do ano 0



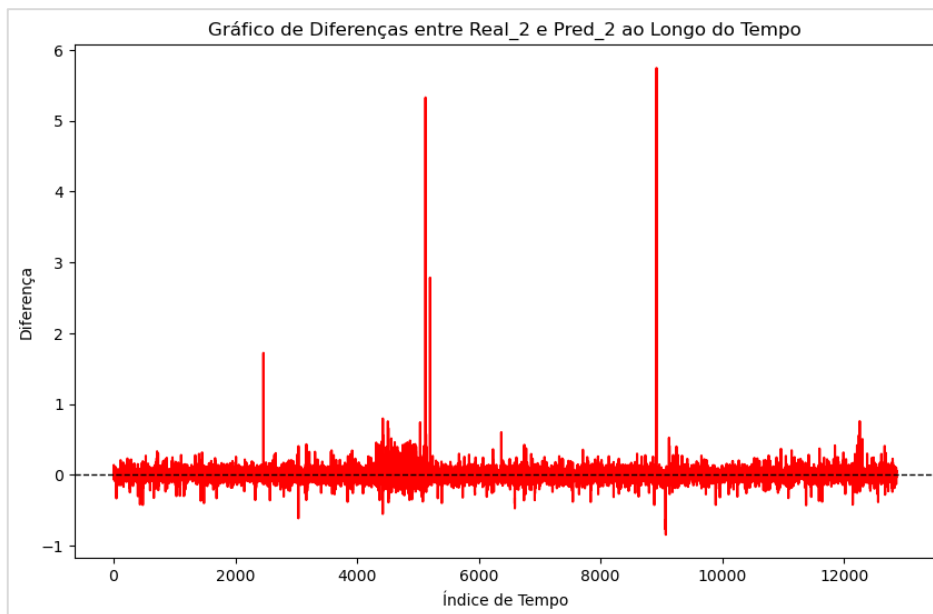
Fonte: O autor, 2024.

Figura 11 – Gráfico de diferença entre valores reais e previsões do ano 1



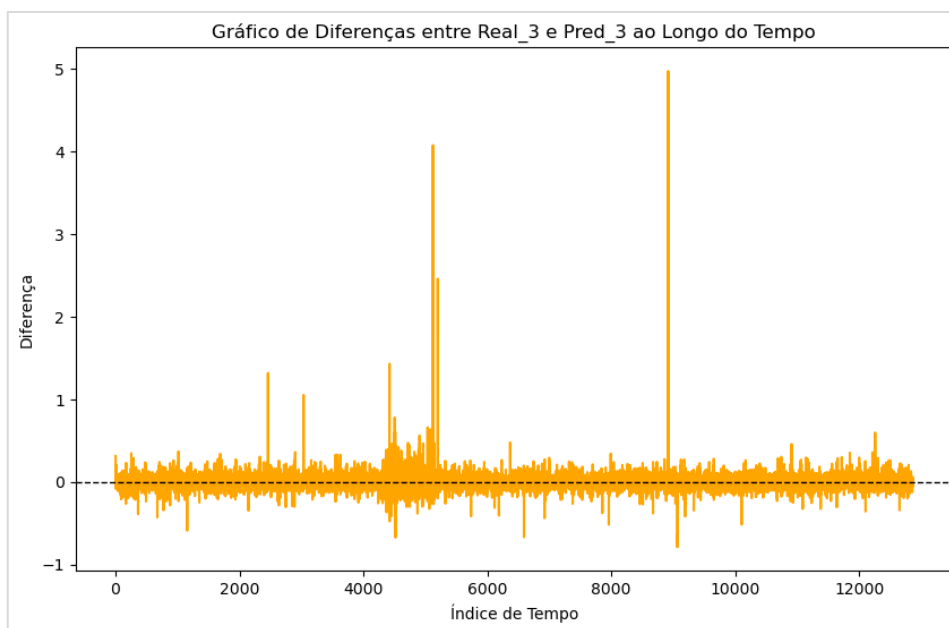
Fonte: O autor, 2024.

Figura 12 – Gráfico de diferença entre valores reais e previsões do ano 2



Fonte: O autor, 2024.

Figura 13 – Gráfico de diferença entre valores reais e previsões do ano 3



Fonte: O autor, 2024.

Com isso, finalizamos o desenvolvimento do modelo de XGBoost para previsão do indicador IN049 do SNIS que se refere ao Índice de Perdas na Distribuição (IPD) ao longo do tempo, utilizando os indicadores fornecidos pelo site SNIS, IBGE e Atlas Brasil. Verificamos que o modelo é capaz de gerar boas estimativas a partir dos indicadores de entrada. Deste modo, é possível utilizá-lo na prática para fazer

previsões do indicador de perdas ao longo do tempo, e embasar um processo de planejamento estratégico de uma empresa de fornecimento de água.

CONCLUSÃO

Neste capítulo apresentamos a conclusão deste trabalho. Inicialmente recordamos quais objetivos haviam sido propostos no início do trabalho e apresentamos o modo como os objetivos foram atingidos. Em seguida elaboramos as considerações finais e proposições de trabalhos futuros que poderiam dar continuidade ao presente estudo.

Atendimento aos Objetivos da Pesquisa

Nesta seção, exploramos em maiores detalhes o atendimento aos objetivos propostos no início deste estudo, em específico o sucesso na conquista do nosso objetivo central: desenvolver um modelo baseado em técnicas de *Machine Learning* que seja capaz de prever o indicador IN049 do SNIS, Índice de Perdas na Distribuição (IPD).

O objetivo central do nosso estudo foi plenamente alcançado. Esta afirmação possui o respaldo nos resultados obtidos através das métricas de desempenho utilizadas no processo de avaliação e interpretação do modelo desenvolvido, o Erro Quadrático Médio (MSE), o Coeficiente de Determinação (R^2), a Raiz do Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE).

Como já mencionado anteriormente, o MSE calculado teve um resultado de 0.02203, indicando que as previsões do modelo estão bem próximas dos valores reais. Já o R^2 calculado foi igual a 0.99993, sugerindo que o modelo explica praticamente toda a variabilidade dos dados de teste. O RMSE calculado foi de 0.14842, confirmando a precisão das previsões. Já o MAE calculado igual a 0.06982, sugere que o modelo tem um bom desempenho em relação à precisão das previsões.

Além dos resultados obtidos através das métricas de desempenho, as visualizações das previsões do modelo se mostraram satisfatórias, conforme ilustrado nas figuras 9 a 13 da seção 3.6.

Considerações Finais

Este trabalho ressalta o grande benefício de utilizar técnicas avançadas de *Machine Learning* para previsão de séries temporais, especificamente propor soluções de acordo com a técnica de árvore de decisão com o algoritmo XGBoost.

É de suma importância reforçar que todas as etapas realizadas neste trabalho foram cruciais para obtermos o sucesso relevante no resultado ao final do desenvolvimento e atendimento aos objetivos propostos. Desde a seleção das informações e indicadores que seriam objeto de estudo, a coleta dos dados em fontes oficiais e confiáveis, o trabalho de tratamento dos dados, bem como os ajustes necessários no modelo de treinamento e teste, além da utilização de métricas adequadas para a avaliação e interpretação dos resultados, incluindo as análises gráficas que corroboram o parecer final.

Durante todo o trabalho buscou-se utilizar referências confiáveis, ferramentas adequadas, e as boas práticas definidas pela comunidade de *Machine Learning*, aumentando a confiabilidade do modelo desenvolvido e dos resultados obtidos.

Sugestões de Trabalhos Futuros

A partir dos conhecimentos e resultados obtidos neste trabalho será possível desenvolver novos trabalhos futuros.

Algumas sugestões para potencializar e refinar a análise desenvolvida neste estudo envolvem obter uma maior granularidade dos dados. Por exemplo, pode-se levantar informações em níveis menores do que o município, entrando no detalhamento de cada setor de abastecimento ou até mesmo nos Distritos de Medição e Controle (DMCs) que já tenham sido definidos. Um outro recurso que pode ser útil é a inclusão de novas variáveis relevantes, como as ações realizadas dentro das empresas de saneamento para a redução de perdas de água, que incluem os consertos de vazamentos em redes, ramais, cavaletes e adutoras, bem como as atividades de combate à fraude e irregularidades. Um ponto interessante seria a integração de indicadores operacionais e comerciais, buscando-se obter um modelo mais abrangente do que o elaborado neste trabalho.

Nossos resultados trabalho também sugerem que conceito e técnicas de *Machine Learning* possam ser aplicados a séries temporais de outros indicadores, não somente na gestão eficiente do sistema de abastecimento de água em empresas de saneamento, mas em toda a gestão dos recursos hídricos ou até mesmo prestadores de outros serviços, principalmente aqueles que envolvem a utilização e consumo de recursos naturais.

REFERÊNCIAS

ANDRADE SOBRINHO, R.; BORJA, P. C. Gestão das perdas de água e energia em sistema de abastecimento de água da Embasa: um estudo dos fatores intervenientes na RMS. **Engenharia Sanitaria e Ambiental**, v. 21, n. 4, p. 783–795, dez. 2016.

BRASIL, M. DO D. R. **Glossário de Informações - Água e Esgotos**. Brasil: SNIS, 2018.

BRASIL, S. N. DE S. A. (ORG). (ED.). **Abastecimento de água: gerenciamento de perdas de água e energia elétrica em sistemas de abastecimento: guia do profissional em treinamento: nível 2**. Salvador, BA: ReCESA, 2008.

BRASIL, T. PERDAS DE ÁGUA 2020 (SNIS 2018): DESAFIOS PARA DISPONIBILIDADE HÍDRICA E AVANÇO DA EFICIÊNCIA DO SANEAMENTO BÁSICO. n. 1ª, p. 68, 2020.

Definição SNIS. Disponível em: <<http://www2.ibam.org.br/rcidades/snis.html>>.

KUSTERKO, S. et al. Gestão de perdas em sistemas de abastecimento de água: uma abordagem construtivista. **Engenharia Sanitaria e Ambiental**, v. 23, n. 3, p. 615–626, jun. 2018.

SNIS - Diagnóstico anual de Água e Esgotos. Disponível em: <<http://www.snis.gov.br/diagnosticos/agua-e-esgotos>>. Acesso em: 11 maio. 2021.

ANDRADE SOBRINHO, R.; BORJA, P. C. Gestão das perdas de água e energia em sistema de abastecimento de água da Embasa: um estudo dos fatores intervenientes na RMS. **Engenharia Sanitaria e Ambiental**, v. 21, n. 4, p. 783–795, dez. 2016.

ARMSTRONG, J. S. (ED.). **Principles of forecasting: a handbook for researchers and practitioners**. 4. printing ed. Boston: Kluwer Academic, 2004.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009) ed. New York, NY: Springer New York, 2016.

BRASIL, M. DO D. R. **Glossário de Informações - Água e Esgotos**. Brasil: SNIS, 2018.

BRASIL, S. N. DE S. A. (ORG). (ED.). **Abastecimento de água: gerenciamento de perdas de água e energia elétrica em sistemas de abastecimento: guia do profissional em treinamento: nível 2**. Salvador, BA: ReCESA, 2008.

BRASIL, T. PERDAS DE ÁGUA 2020 (SNIS 2018): DESAFIOS PARA DISPONIBILIDADE HÍDRICA E AVANÇO DA EFICIÊNCIA DO SANEAMENTO BÁSICO. n. 1ª, p. 68, 2020.

BROWNLEE, J. **What Is Time Series Forecasting?** , 15 ago. 2020. Disponível em: <<https://machinelearningmastery.com/time-series-forecasting/>>. Acesso em: 12 fev. 2024

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais...**13 ago. 2016. Disponível em: <<http://arxiv.org/abs/1603.02754>>. Acesso em: 12 fev. 2024

CRYER, J. D.; CHAN, K.-S. **Time series analysis: with applications in R**. 2. ed., corr. print ed. New York, NY: Springer, 2009.

Definição SNIS. Disponível em: <<http://www2.ibam.org.br/rcidades/snis.html>>.

DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, p. 78–87, out. 2012.

GOUVEIA, C. **TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À PREDIÇÃO DE VAZAMENTOS EM RAMAIS DE REDES DE DISTRIBUIÇÃO DE ÁGUA**. Tese de Mestrado—[s.l.] UNIVERSIDADE DE BRASÍLIA, 11 fev. 2022.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. Third print edition ed. Melbourne, Australia: Otexts, Online Open-Access Textbooks, 2021.

Índice de Desenvolvimento Humano Município - Atlas BR. , 2024. Disponível em: <<http://www.atlasbrasil.org.br/consulta>>

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 17 jul. 2015.

KAMTZIRIDIS, G. **Time Series Forecasting with XGBoost and LightGBM: Predicting Energy Consumption**. , 27 fev. 2023. Disponível em: <<https://medium.com/mlearning-ai/time-series-forecasting-with-xgboost-and-lightgbm-predicting-energy-consumption-460b675a9cee>>. Acesso em: 12 fev. 2024

KAUR, H. A Comprehensive Review on Time Series Forecasting Techniques. p. p103–p111, maio 2023.

KUSTERKO, S. et al. Gestão de perdas em sistemas de abastecimento de água: uma abordagem construtivista. **Engenharia Sanitaria e Ambiental**, v. 23, n. 3, p. 615–626, jun. 2018.

LEO, C. **The Math Behind XGBoost**. , 10 jan. 2024. Disponível em: <<https://medium.com/p/3068c78aad9d>>

LONGARAY, A. A. et al. Emprego de métodos multicritério em decisões gerenciais: uma análise bibliométrica da produção científica brasileira. **Revista Contemporânea de Contabilidade**, v. 13, n. 29, p. 113, 26 ago. 2016.

MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python: a guide for data scientists**. First edition ed. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2016.

PODGORELEC, V. et al. Decision Trees: An Overview and Their Use in Medicine. **Journal of Medical Systems**, v. 26, n. 5, p. 445–463, 2002.

Produto Interno Bruto dos Municípios - IBGE. , 2024. Disponível em: <<https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?edicao=18021>>

ROKACH, L.; MAIMON, O. **Data mining with decision trees: theory and applications**. Hackensack (NJ): World Scientific, 2008.

SHARMA, N. **How to Use XGBoost for Time-Series Forecasting?** , 4 jan. 2024. Disponível em: <<https://www.analyticsvidhya.com/blog/2024/01/xgboost-for-time-series-forecasting/>>. Acesso em: 11 fev. 2024

SNIS - Diagnóstico anual de Água e Esgotos. Disponível em: <<http://www.snis.gov.br/diagnosticos/agua-e-esgotos>>. Acesso em: 11 maio. 2021.

SNIS - Série Histórica. , 2024. Disponível em: <<http://app4.mdr.gov.br/serieHistorica/>>