



PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO E
SISTEMAS (PPGEPS)

FRANCHESCO SANCHES DOS SANTOS

HYDROFLOW: PREVISÃO PRECISA DA DEMANDA DE ÁGUA COM MÉTODOS
DE GRADIENTE, REGRESSÃO E ARIMA

CURITIBA
2023

FRANCHESCO SANCHES DOS SANTOS

**HYDROFLOW: PREVISÃO PRECISA DA DEMANDA DE ÁGUA COM MÉTODOS
DE GRADIENTE, REGRESSÃO E ARIMA**

Projeto de Pesquisa de Mestrado apresentado ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS). Área de concentração: Automação e Controle de Sistemas, da Escola Politécnica, da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Orientador: Dr. Leandro dos Santos Coelho
Coorientadora: Dr. Viviana Cocco Mariani

CURITIBA
2023

FRANCHESCO SANCHES DOS SANTOS

**HYDROFLOW: PREVISÃO PRECISA DA DEMANDA DE ÁGUA COM
MÉTODOS DE GRADIENTE, REGRESSÃO E ARIMA**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS). Área de concentração: Gerência de Produção e Logística, da Escola Politécnica, da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

COMISSÃO EXAMINADORA

Dr. Leandro dos Santos Coelho
Orientador

Pontifícia Universidade Católica do Paraná

Dr. Viviana Cocco Mariani
Coorientadora

Pontifícia Universidade Católica do Paraná

Convidado A
Membro Externo
Instituição A

Convidado B
Banca
Instituição B

Curitiba, 22 de maio de 2023

Com gratidão, dedico este trabalho a Deus.
Devo a ele tudo o que sou.

Agradecimentos

Primeiramente, expresso minha gratidão a Deus por todas as bênçãos recebidas, pois foi Ele quem abriu caminhos e me deu forças para superar esse desafio, tornando-o possível.

À minha família, sou grato pelo apoio incondicional e pelo estímulo constante para seguir em frente com determinação, buscando sempre alcançar novos patamares.

Agradeço ao professor Leandro dos Santos Coelho pela oportunidade de trabalhar ao seu lado e compartilhar seus conhecimentos e experiências ao longo do meu mestrado. Sua orientação contribuiu significativamente para o meu crescimento profissional e pessoal, tornando este trabalho uma realidade.

À professora Viviana Cocco Mariani, agradeço pela disponibilidade e paciência em me auxiliar nas minhas dificuldades, utilizando seu conhecimento para contribuir com o desenvolvimento da pesquisa.

Quero expressar minha gratidão à equipe da Pontifícia Universidade Católica do Paraná (PUCPR) e aos demais professores, especialmente à secretária Denise da Mata Medeiros (PPGEPS), pela paciência, carinho e apoio prestados em diversas ocasiões, sem medir esforços.

Aos meus amigos, que sempre torceram por mim, e aos novos amigos que conquistei ao longo dessa jornada, agradeço por compartilharmos momentos de alegria nessa batalha.

Sou grato ao investimento em bolsas de estudo concedidas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que possibilitou a conclusão dessa etapa da minha carreira profissional e acadêmica.

“A matemática é o alfabeto no qual Deus escreveu o universo.”

Galileu Galilei

Resumo

Em um cenário competitivo, a previsão assertiva de demanda tem se tornado cada vez mais uma ferramenta estratégica para diversas áreas organizacionais. Nesse contexto, a previsão de séries temporais desempenha um papel importante na tomada de decisões. Recentemente, a capital paranaense enfrentou uma grave crise na área da saúde, com períodos de desabastecimento que geraram instabilidade na oferta de moradia para muitas famílias. Na abordagem para solucionar esse problema, foram utilizados métodos encontrados na revisão bibliográfica realizada durante este trabalho. Os métodos escolhidos são utilizados como uma forma de tomada de decisão para a demanda de água. Cada método escolhido tem a capacidade de lidar com o problema de maneira diferente e oferecer soluções viáveis para a tomada de decisão. Com o objetivo de mitigar e tomar as melhores decisões para o problema enfrentado pela Companhia de Saneamento do Paraná (SANEPAR) em 2020 e evitar surpresas no futuro, este trabalho visa aprimorar o uso da água. Embora o evento isolado ocorrido nesse ano possa não se repetir nos próximos anos, é importante buscar melhorias na gestão dos recursos hídricos. Os métodos derivados do modelo autorregressivo integrado de médias móveis (ARIMA), incluindo aqueles com variáveis exógenas e considerando a sazonalidade dos dados, são os modelos de previsão mais eficazes para modelar os dados com variações temporais, embora cada método tenha suas peculiaridades, todos são baseados no modelo ARIMA inicial. Os modelos de boosting, como XGBoost (*Impulso Extremo de Gradiente*), seguidos pelo modelo de regressão linear simples (LR), são considerados os melhores modelos para séries temporais devido ao uso da abordagem de gradient boosting para previsões. Essa escolha é baseada em métricas de erro, em que um menor valor indica uma melhor capacidade de tomada de decisão. As métricas adotadas neste artigo são o erro percentual absoluto médio (MAPE), o erro absoluto médio (MAE) e a raiz quadrada do erro quadrático médio (RMSE). Em séries temporais, essas métricas são comumente utilizadas para avaliar a eficácia dos modelos de previsão em diferentes circunstâncias e horizontes de previsão. O modelo XGBoost apresentou um erro de 0.264% no MAPE, o menor entre os modelos avaliados, enquanto o modelo LR obteve o maior erro de 5% no horizonte de previsão mais longo (um mês). O modelo de médias móveis (MA) obteve um erro de 0.113% no MAPE, enquanto o modelo LR apresentou um erro de 5%. Assim, o modelo LR pode ser mais eficiente para conjuntos de dados menores, trabalhando com um volume reduzido de dados, enquanto os erros aumentam à medida que o horizonte de previsão aumenta.

Palavras-chave: Previsão, Economia de água, Séries temporais, Revisão sistemática da literatura.

Abstract

In a competitive scenario, accurate demand forecasting has become an increasingly strategic tool for various organizational sectors. In this context, time series forecasting plays a significant role in decision-making. Recently, the capital city of Paraná faced a severe healthcare crisis, with periods of shortages that caused serious instability in housing supply for many families. In addressing this problem, the methods found through the literature review conducted in this work were used for prediction purposes. The chosen methods serve as a decision-making approach for water demand. Each selected method is capable of dealing with the problem in a different way and providing viable solutions for decision-making. With the objective of mitigating the issue faced by the Paraná Sanitation Company (SANEPAR) in 2020 and avoiding surprises in the near future, this work aims to enhance water usage. Although the isolated event that occurred in that year may not repeat itself in subsequent years, it is important to seek improvements in water resource management. Methods derived from the autoregressive integrated moving average (ARIMA) model, including those with exogenous variables and considering data seasonality, prove to be the most effective forecasting models for modeling data with temporal variations. Although each method has its own characteristics, they are all based on the initial ARIMA model. Boosting models, such as XGBoost (Extreme Gradient Boosting), followed by the simple linear regression (LR) model, are considered the best models for time series due to their use of the gradient boosting approach for predictions. This choice is based on error metrics, where a lower value indicates better decision-making capability. The metrics used in this article include mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean squared error (RMSE). In time series, these metrics are commonly used to evaluate the effectiveness of prediction models under different circumstances and forecast horizons. The XGBoost model achieved an MAPE error of 0.264 %, the lowest among the evaluated models, while the LR model had the highest error of 5 % in the longest forecast horizon (one month). The moving averages (MA) model obtained an MAPE error of 0.113 %, while the LR model had an error of 5 %. Therefore, the LR model may be more efficient for smaller datasets, working with a reduced volume of data, while errors increase as the forecast horizon lengthens.

Keywords: Forecasting, Water savings, Time series, Systematic literature review.

Lista de Abreviaturas e Siglas

AdaBoost	Impulso ou Estímulo adaptativo (do inglês <i>Adaptive Boosting</i>)
AR	Auto-Regressivo
ARIMA	Média Móvel Integrada Auto-Regressiva (do inglês <i>autoregressive integrated moving average</i>)
ARIMAX	Média Móvel Integrada Auto-Regressiva com regressores eXogenous (do inglês <i>autoregressive integrated moving average with eXogenous regressors</i>)
ARMA	Média Móvel Auto-Regressivo (do inglês <i>autoregressive moving average</i>)
ARX	Auto-Regressivo com variável Exógena (do inglês <i>autoregressive with exogeneous inputs</i>)
BrownBoost	Algoritmo de aumento
CNN	Rede Neural Convolucional (do inglês <i>Convolutional Neural network ou ConvNet</i>)
DBN	Rede de Crenças Profundas (do inglês <i>Deep Belief Network</i>)
EFB	Pacote de características exclusivas (do inglês <i>Exclusive Feature Bundling</i>)
FT	flow transmitter (Transmissor de fluxo)
Hz	Hertz
INMET	Instituto Nacional de Meteorologia
LGBMRegressor	Regressão Light GBM
Light GBM	Máquina de Impulso de Gradiente Leve (do inglês <i>Light Gradient Boosting Machine</i>)
LogitBoost	Representa uma aplicação de técnicas de regressão logísticas
LPBoost	Reforço da Programação Linear (do inglês <i>Linear Programming Boosting</i>)
LR	Regressão linear (do inglês <i>Linear Regression</i>)
LSTM	Memória de longo curto prazo (do inglês <i>Long short-term memory</i>)
m^3	Metros cúbicos
m^3/h	Metros cúbicos por hora

MA	Média Móvel (do inglês <i>moving average</i>)
MadaBoost	Modificando o sistema de ponderação da AdaBoost
MAE	Erro Médio Absoluto (do inglês <i>Mean Absolute Error</i>)
MAPE	Erro Percentual Médio Absoluto (do inglês <i>Mean Absolute Percentage Error</i>)
mca	Metros coluna de água
ML	Aprendizado de máquina (do inglês <i>machine learning</i>)
mm	Milímetros
MSE	Erro médio quadrático (do inglês <i>Mean Squared Error</i>)
PR	Estado do Paraná
RBAL	Recalque Bairro Alto
RFR	Regressão de floresta aleatória (do inglês <i>Random Forest Regression</i>)
RMSE	Erro de Raiz Média Quadrática (do inglês <i>Root Mean Squared Error</i>)
RNN	Rede Neural Recorrente (do inglês <i>Recurrent Neural Network</i>)
SANEPAR	Companhia de Saneamento do Paraná
SARIMA	Auto-Regressivos Integrados de Médias Móveis com Sazonalidade (do inglês <i>Integrated Auto-Regressive Moving Averages with Seasonality</i>)
SARIMAX	Média Móvel Integrada Auto-Regressiva Sazonal com regressores exogenos (do inglês <i>Seasonal Auto-Regressive Integrated Moving Average with exogenous regressors</i>)
SVM-VAR	Máquinas de vetor de suporte - Vetores Auto-Regressivos
Totalboost	Impulso total
XGBoost	Impulso Gradiente Extremo (do inglês <i>eXtreme Gradient Boosting</i>)
XGBRegressor	Regressão XGBoost

Lista de Tabelas

1	Cruzamento de palavras-chave através da aplicação de filtros de ano e de linguagem	20
2	Fator de impacto	22
3	Áreas e seus valores respetivos de artigos em cada área.	26
4	Descrição estatística dos dados com o filtro aplicado das 18h às 21h	55
5	Teste Nemenyi	64
6	Comparação dos modelos com 1 dia de antecedência 24h Treinamento	74
7	Comparação dos modelos com 1 dia de antecedência 24h Validação	75
8	Comparação dos modelos com 1 dia de antecedência 24h Teste	75
9	Comparação dos modelos com 1 dia de antecedência 24h Completo	76
10	Comparação dos modelos com 7 dias de antecedência 24h Treinamento	76
11	Comparação dos modelos com 7 dias de antecedência 24h Validação	77
12	Comparação dos modelos com 7 dias de antecedência 24h Teste	77
13	Comparação dos modelos com 7 dias de antecedência 24h Completo	78
14	Comparação dos modelos com 14 dias de antecedência 24h Treinamento	78
15	Comparação dos modelos com 14 dias de antecedência 24h Validação	79
16	Comparação dos modelos com 14 dias de antecedência 24h Teste	79
17	Comparação dos modelos com 14 dias de antecedência 24h Completo	80
18	Comparação dos modelos com 30 dias de antecedência 24h Treinamento	80
19	Comparação dos modelos com 30 dias de antecedência 24h Validação	81
20	Comparação dos modelos com 30 dias de antecedência 24h Teste	81
21	Comparação dos modelos com 30 dias de antecedência 24h Completo	82
22	Comparação dos modelos Ljung Box Treinamento	82
23	Comparação dos modelos Ljung Box Validação	83
24	Comparação dos modelos Ljung Box Teste	83
25	Comparação dos modelos Ljung Box Completo	83

Lista de Figuras

1	Paradigma de aprendizado de máquina	2
2	Mapa das Etapas	5
3	Estrutura da dissertação	9
4	Dados completos com uma frequência média de 24 horas	12
5	Plotagem de dados para o ano de 2020	12
6	Exemplo de séries temporais	13
7	Processo estocástico	14
8	Mapa conceitual do problema de pesquisa	15
9	Etapas da Revisão.	16
10	Palavras-chave mais populares na Scopus.	18
11	Palavras-chave mais populares na WoS	19
12	Analise das quantidades de artigos em relação aos anos.	20
13	Relação de autores entre artigos publicados	22
14	Ligaçāo bibliográfica entre os autores	23
15	Mapa mundial da publicação de artigos em todo o mundo	24
16	Áreas de aplicāção do tema	25
17	Modelo AR(7)	33
18	ARX (7)	33
19	Modelo MA(7)	36
20	ARMA (7,7)	37
21	ARIMA (7,1,7)	38
22	SARIMA (7, 1, 7)(2, 1, 1) ₁₂	39
23	ARIMAX (7, 1, 7)	40
24	SARIMAX (7, 1, 7)(2, 1, 1) ₁₂	40
25	Corelação de Pearson	42
26	Régressāo linear LT01 vs PT01 correlação 98%	43
27	Régressāo linear (LR) um passo a frente	44
28	Régressāo da Floresta Aleatória (RFR)	44
29	Esquema da Floresta Aleatória	45
30	Impulsionando gradiente com XGBoost e LightGBM	46
31	Crescimento em folha versus crescimento em nível	48
32	XGBoost e LighGBM régressāo	49
33	Decomposição STL aditiva dos dados coletados	57
34	Decomposição STL multiplicativa dos dados coletados	57
35	Violino no nível do reservatório	58

36	Violino da vazão de recalque	59
37	Autocorrelação e Autocorrelação parcial	61
38	Ruído branco	62
39	Comparação dos modelos ARIMAS	65
40	Comparação de modelos de regressão	66
41	Comparação dos modelos AR, ARX e MA, 1 dia à frente	84
42	Comparação dos modelos AR, ARX e MA, 7 dias à frente	84
43	Comparação dos modelos AR, ARX e MA, 14 dias à frente	85
44	Comparação dos modelos AR, ARX e MA, 30 dias à frente	85
45	Comparação dos modelos ARMA e ARIMA, 1 dia à frente	86
46	Comparação dos modelos ARMA e ARIMA, 7 dias à frente	86
47	Comparação dos modelos ARMA e ARIMA, 14 dias à frente	87
48	Comparação dos modelos ARMA e ARIMA, 30 dias à frente	87
49	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente	88
50	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente	88
51	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à frente	89
52	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à frente	89

Sumário

1	Introdução	1
1.1	Contexto da pesquisa	1
1.1.1	Motivação da pesquisa	2
1.2	Objetivo geral	3
1.2.1	Objetivos específicos e questão de pesquisa	3
1.3	Descrição do problema	4
1.4	Procedimentos metodológicos	4
1.4.1	Etapas da pesquisa	5
1.5	Justificativa da pesquisa	8
1.5.1	Contribuições	8
1.6	Estrutura do trabalho	9
2	Referencial	11
2.1	Detecção de anomalias	11
2.2	Revisão sistemática da literatura	13
2.3	Problematização da Revisão	15
2.4	Metodologia	16
2.5	Resultados da busca de revisão	18
2.6	Principais conclusão	27
3	Base Teórica	29
3.1	Métricas de Erros	29
3.1.1	Erro quadrático médio raiz (RMSE)	29
3.1.2	Erro Absoluto Médio (MAE)	30
3.1.3	Erro Percentual Absoluto Médio (MAPE)	31
3.2	Modelos de séries temporais univariados	32
3.2.1	Componente Autorregressivo	32
3.2.2	AR(0): Ruído branco	34
3.2.3	AR(1): Caminhadas aleatórias e Oscilações	34
3.2.4	AR(p): Termos de ordem superior	35
3.2.5	Média Móvel	35
3.2.6	Modelos ARMA e ARIMA	37
3.2.7	ARIMA	37
3.2.8	SARIMA	38
3.3	Modelos de série temporal multivariada	39

3.3.1	ARIMAX e SARIMAX	39
3.4	Modelos de aprendizado de máquina supervisionados	41
3.4.1	Regressão Linear (LR)	41
3.4.2	Definição do modelo	42
3.4.3	Floresta Aleatória	44
3.4.4	LightGBM e XGboost	45
3.4.5	O Gradiente em Gradiente de Boosting (Reforço)	46
3.4.6	Algoritmos de boosting de gradiente	47
3.4.7	A diferença entre XGBoost e LightGBM	47
3.5	Estudo de Caso	49
3.5.1	Definição do problema	50
3.5.2	Coleta de dados	50
3.5.3	Análise exploratória dos dados	50
3.5.4	Escolha do modelo	51
3.5.5	Divisão dos dados	51
3.5.6	Ajuste do modelo	51
3.5.7	Avaliação do modelo	53
3.5.8	Previsões futuras	53
3.5.9	Monitoramento e ajuste contínuo	53
3.5.10	Principais Conclusão	53
4	Resultados	54
4.1	Planejamento do Problema	54
4.1.1	Análise Exploratória dos dados (EDA)	54
4.1.2	Múltiplas entradas e saída única (MISO)	56
4.1.3	Decomposição STL	56
4.1.4	Separação dos dados	62
4.1.5	Estratégia de Previsão	63
4.1.6	Horizonte	63
4.1.7	Modelos de previsão e métricas de desempenho	63
4.1.8	Teste de Significância	64
4.1.9	Comparação dos modelos	65
4.2	Estudo de caso	66
5	Conclusões	67
5.1	Limitações da pesquisa e propostas futuras	67
Referências		69

A Apêndice - Comparaçāo dos modelos de previsão de series temporais média de 24h	74
B Apêndice - Comparaçāo dos modelos de previsão com o método Ljung Box	82
C Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h	84
D Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h	86
E Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e SARIMAX (7,1,7) (2,1,0,12) 24h	88

1 Introdução

Este capítulo apresenta o conteúdo abordado nesta dissertação, que se concentra na utilização de modelos de Aprendizado de Máquina (ML) para prever futuramente os dados coletados pela SANEPAR. Os dados coletados referem-se ao abastecimento de água no bairro alto durante o período de 2018 a 2020, quando ocorreu uma escassez que afetou toda a população da capital paranaense.

Dentro do contexto de análise de séries temporais e tomada de decisão, foram explorados modelos de ML para aplicação nesses dados. Por meio de uma revisão sistemática da literatura, foram identificados e tabulados os modelos clássicos mais comumente utilizados para análise de séries temporais.

1.1 Contexto da pesquisa

Ribeiro et al. (2021) A necessidade de desenvolvimento do planejamento estratégico no mundo corporativo e no dia-a-dia torna a análise de séries temporais e previsões valiosas ferramentas para apoiar o processo de tomada de decisão a curto, médio e longo prazo. Devido a não linearidades, sazonalidade, tendência e ciclicidade nos dados temporais, o desenvolvimento de modelos de previsão eficientes é uma tarefa desafiadora.

No conjunto de dados da SANEPAR, há um volume significativo no consumo de água e, com as interrupções que a cidade tem enfrentado, é necessário analisar os dados para compreender melhor os padrões de interrupção no abastecimento e os picos de consumo ao longo das horas e dias.

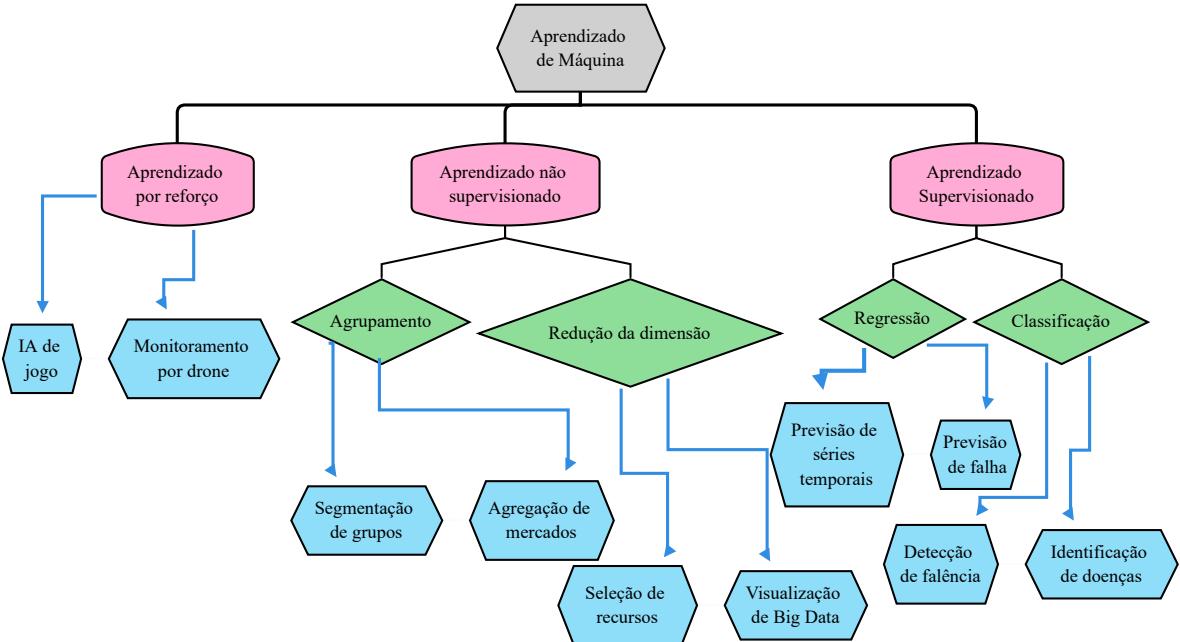
Nesta dissertação, será realizada uma revisão sistemática de modelos preditivos para avaliar o melhor modelo que pode ser utilizado e como ele pode ser validado para prever a escassez de água. Essas análises serão feitas utilizando a linguagem de programação *Python*.

A abordagem deste trabalho consiste em explorar o conceito de séries temporais e sua aplicação no campo do aprendizado de máquina. Os dados de séries temporais referem-se a dados coletados e armazenados ao longo do tempo, permitindo que observadores identifiquem anomalias nos dados. A classificação dos dados por ano ou dia é essencial na análise de séries temporais, e se os dados forem atribuídos aleatoriamente, pode ser mais desafiador fazer previsões e tomar decisões com base nos dados coletados.

É importante destacar que a análise de médias pode ser enganosa se não forem excluídos os valores discrepantes, também conhecidos como “*outliers*”. Esses valores discrepantes podem levar a resultados extremamente altos ou baixos que não refletem a realidade.

O campo do aprendizado de máquina abrange várias áreas, conforme ilustrado na Figura 1. Serão explorados os diferentes componentes do aprendizado de máquina e como eles podem ser aplicados em diversos contextos.

Figura 1: Paradigma de aprendizado de máquina



Fonte: Elaboração própria

1.1.1 Motivação da pesquisa

De acordo com (VASCONCELOS, 2020) Curitiba e região metropolitana enfrentou um rodízio com 36 horas com água e 36 horas sem abastecimento. A média geral dos reservatórios da região está em 27,96% da capacidade. Assim em medida a isso essa pesquisa tem como a abordagem da falta de água, essa falta que pode ser vista como uma seca, em média nos anos anteriores de 2020 a chuva tem marcado a quantia de 1.704 mm. (VASCONCELOS, 2020) Desde 2016, quando registrou 1.704 mm de chuva, Curitiba não atingiu mais a média anual de precipitação, que é de 1.490 mm, com base em dados da estação pluviométrica do IBMET. Apesar de abaixo da média, o mínimo registrado desde então ocorreu em 2020, com 1.158 mm.

Em meio a essa motivação, é possível realizar uma análise mais aprofundada dos dados fornecidos pela SANEPAR, a fim de prever e evitar a ocorrência de escassez de água, que foi registrada juntamente com a anomalia detectada em 2020. Com o retorno das chuvas, houve um aumento no nível dos reservatórios.

1.2 Objetivo geral

O objetivo desta pesquisa é identificar o melhor modelo de séries temporais para abordar o problema da escassez de água que ocorreu em Curitiba. Ao longo da dissertação, foram avaliados diversos modelos de regressão, com foco especial nos modelos baseados em *gradient boosting*, considerados eficazes na literatura para a previsão de séries temporais. Os principais modelos explorados incluem o ARIMA e suas variantes atualizadas. Além da previsão, também serão realizadas análises de anomalias nos dados, visando compreender as causas subjacentes a essas ocorrências.

1.2.1 Objetivos específicos e questão de pesquisa

Neste estudo, busca-se identificar e compreender possíveis anomalias nos dados, bem como investigar as causas por trás dessas ocorrências. O objetivo é responder às perguntas de pesquisa relacionadas a essas anomalias.

Q 1 A pressão é suficiente para a demanda diária?

Q 2 Quanta água deve ter no reservatório para evitar o acionamento das bombas no horário de pico (18 às 21 h)? Quanto maior a frequência de funcionamento da bomba maior a demanda. Valor máximo 60 Hz.

Q 3 Qual a vazão ótima para atender a demanda? Quanta pressão para atender a demanda?

Q 4 Ponto de equilíbrio entre demanda e vazão e ter um armazenamento sem necessidade de acionar as bombas no período do custo energético mais caro (18 às 21 horas).

Q 5 Se a SANEPAR ativar as bombas de sucção das 18 às 21 horas ela tem o maior custo energético, isto é, ela paga mais caro pela energia neste período.

- a. Qual o nível que deve estar no reservatório para não ser necessário a SANEPAR ativar as bombas das 18 às 21 horas sem faltar água para a população? Verificar a média das vazões nos horários críticos (onde tem a maior demanda 18 às 21 horas) para as diferentes estações do ano (Outono, Inverno, Primavera, Verão).
- b. Existe tendência, padrão, sazonalidade para os dados destes 3 anos do Bairro Alto?
- c. Identificar quais os horários de maior demanda das 18 às 21?
- d. Quanto tenho que armazenar previamente no reservatório para não acionar as bombas no horário de pico?

- e. Se a vazão cresce e a pressão decresce temos uma ANOMALIA na rede (com base no histórico).

1.3 Descrição do problema

A descrição do problema é fundamental para obter uma compreensão mais precisa do que está sendo abordado neste trabalho. É por meio dessa descrição que as variáveis-chave são expostas e o objetivo da previsão é estabelecido de forma clara. Sem um plano estruturado para determinar o que deve ser previsto, torna-se difícil justificar o uso de modelos de previsão de dados. Portanto, é essencial estabelecer um propósito claro e definir as metas da previsão antes de aplicar os modelos adequados.

- Bombas de sucção (B1, B2 e B3) – valor máximo da frequência 60 Hz

Variáveis importantes: Fluxo, pressão e nível

- Nível do Reservatório (Câmara 1) LT01 (m^3) - **PREVER**
- Vazão de entrada (FT01) (m^3/h)
- Vazão de gravidade (FT02) (m^3/h)
- Vazão de recalque (FT03) (m^3/h)
- Pressão de Sucção (PT01SU) (mca)
- Pressão de Recalque (PT02RBAL) (mca)

A pesquisa fará uso da variável LT01, que representa o nível do reservatório e desempenha um papel de extrema importância, como evidenciado pelas Figuras 4 e 5. Essas figuras retratam as anomalias ocorridas durante o período em que a capital paranaense foi afetada pela escassez de chuvas, resultando na redução do nível dos reservatórios e na implementação de rodízios periódicos, conforme discutido na subseção 1.1.1. Assim, tais observações permitem uma compreensão mais aprofundada das perspectivas futuras.

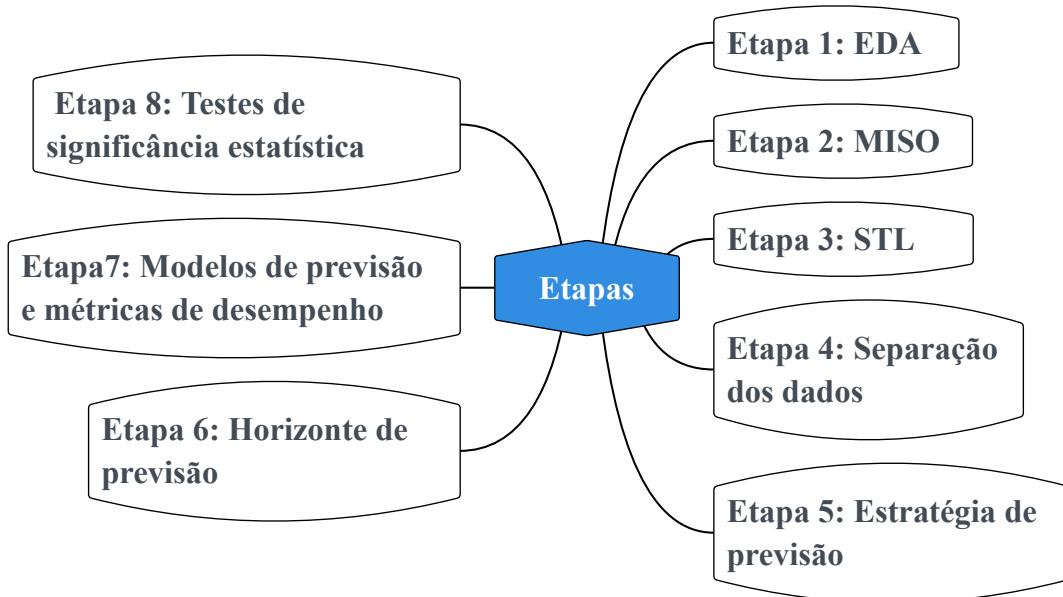
1.4 Procedimentos metodológicos

Com o intuito de realizar previsões e fazer comparações entre os modelos obtidos na revisão sistemática, será adotado um processo metodológico bem definido. Tal processo está detalhado na subseção 1.4.1 desta seção, onde foram estabelecidas as etapas a serem seguidas desde o início. Isso inclui a definição do que será previsto, bem como a seleção dos métodos a serem utilizados na Análise Exploratória de Dados (EDA), seguindo uma sequência lógica e coerente.

1.4.1 Etapas da pesquisa

A pesquisa seguiu as seguintes etapas:

Figura 2: Mapa das Etapas



Fonte: Elaboração própria

Etapa 1 Análise exploratória dos dados – EDA (do inglês *Exploratory Data Analysis*).

A exploração de dados na EDA é fundamental para entender melhor os dados que estão sendo trabalhados, como, por exemplo, excluir valores ausentes, saber como os dados estão separados em horas ou dias e, assim, tomar a melhor decisão a ser trabalhada com os dados, usar gráficos de linha na análise para observar a convergência dos dados e as anomalias que podem ocorrer.

Etapa 2 O que vai ser usado como variáveis previsoras e qual será a variável a ser predita (MISO).

Nessa etapa, tem o papel de relacionar as variáveis ao que será previsto, como os modelos de variáveis exógenas que são usados aqui nos modelos SARIMAX, ARX e ARIMAX do tipo ARIMA. Cada modelo tem a interação de mais variáveis do que o modelo ARIMA básico ou seus derivados AR, MA e SARIMA. O conhecimento de quais variáveis estão incluídas na modelagem do problema torna a modelagem mais abrangente quando o horizonte de previsão é estendido além dos dados.

Etapa 3 Fazer a decomposição STL (do inglês *Seasonal-Trend Decomposition*) Sazonalidade, Tendência e Resíduo.

O algoritmo STL executa suavização na série de tempo usando LOESS em dois loops; o loop interno itera entre a suavização sazonal e de tendência e o loop externo minimiza o efeito de valores atípicos. Durante o loop interno, o componente sazonal é calculado primeiro e removido para calcular o componente de tendência. O restante é calculado subtraindo os componentes sazonais e de tendência da série de tempo.

Os três componentes da análise STL se relacionam com a série de tempo bruta da seguinte forma:

$$y_i = s_i + t_i + r_i \quad (1)$$

Onde:

- y_i = O valor da série de tempo no ponto i .
- s_i = O valor do componente sazonal no ponto i .
- t_i = O valor do componente de tendência no ponto i .
- r_i = O valor do componente restante no ponto i .

Etapa 4 Separação dos dados.

A fim de obter uma divisão mais adequada dos dados, é realizado um estudo das medidas de tendência central e dispersão de cada conjunto. O conjunto de dados é então dividido em três partes distintas: treinamento, validação e teste. Nessa divisão, inicialmente, 70% dos dados são utilizados para o treinamento e validação, enquanto os 30% restantes são reservados para o conjunto de teste. Em seguida, a porção destinada ao treinamento e validação é subdividida em uma proporção de 80% para treinamento e 20% para validação.

Etapa 5 Estratégia de previsão (recursiva e iterada-método direto).

A estratégia recursiva consiste em utilizar um modelo de previsão de um passo de tempo várias vezes, onde a previsão obtida no passo anterior é utilizada como entrada para realizar a previsão do próximo passo de tempo.

No contexto da previsão da demanda de água para os próximos dias, seria desenvolvido um modelo de previsão de um único passo. Esse modelo seria aplicado para prever a demanda no primeiro dia e, em seguida, essa previsão seria utilizada como dado de entrada para prever a demanda do segundo dia. Esse processo se repetiria para os demais dias, permitindo a previsão da demanda ao longo do tempo.

Por Exemplo:

$$preditivo(t+1) = model_1(obs(t-1), obs(t-2), \dots, obs(t-n)) \quad (2)$$

$$preditivo(t+2) = model_2(obs(t-2), obs(t-3), \dots, obs(t-n)) \quad (3)$$

Brownlee (2016) como as previsões são usadas no lugar das observações, a estratégia recursiva permite que os erros de previsão se acumulem de tal forma que o desempenho possa se degradar rapidamente à medida que o horizonte de tempo de previsão aumenta.

Etapa 6 Horizonte de previsão (1 passo ou n passos à frente).

Para abordar a diversidade de horizontes de previsão, optou-se por considerar diferentes intervalos de tempo. Isso permitirá a realização de previsões para um passo à frente, uma semana, duas semanas e um mês, de forma a abranger distintas perspectivas de curto e médio prazo. Essa abordagem proporciona uma análise abrangente da capacidade dos modelos em lidar com horizontes de previsão variados, contribuindo para uma avaliação mais completa e precisa do desempenho dos mesmos.

Etapa 7 Modelos de previsão e métricas de desempenho.

Os modelos abordados nesta pesquisa são tanto os modelos clássicos de previsão quanto os modelos de regressão por gradiente. Entre os modelos clássicos, incluem-se o AR, ARX, ARMA, ARIMA, SARIMA, SARIMAX e ARIMAX, enquanto os modelos de regressão por gradiente englobam o LR, XGBRegressor, RFR e LGBMRegressor. A seleção desses modelos foi baseada em uma revisão sistemática realizada durante a dissertação, buscando identificar os modelos mais eficazes e amplamente utilizados na literatura.

Ao longo da pesquisa, foram adotadas três métricas principais para avaliar o desempenho dos modelos: RMSE, MAE e MAPE. Essas métricas foram escolhidas com base na revisão sistemática e são amplamente reconhecidas como medidas de qualidade de previsão. Cada métrica tem sua própria interpretação e importância, sendo detalhada na subseção 3.1 para um melhor entendimento de como são aplicadas e interpretadas na pesquisa.

Etapa 8 Aplicar os modelos de previsão e fazer comparativo baseado em testes de significância estatística (*Friedman* e *Nemenyi*).

O teste de Friedman é o teste não paramétrico usado para comparar dados de amostras vinculadas, ou seja, quando o mesmo indivíduo é avaliado mais de uma vez.

ou seja, quando o mesmo indivíduo é avaliado mais de uma vez. O teste de Friedman não usa os dados numéricos diretamente, mas sim as classificações ocupadas pelos dados após a classificação de cada grupo separadamente. Após a classificação, a hipótese de igualdade da soma das classificações de cada grupo é testada.

O teste consiste em fazer comparações em pares com o intuito de verificar qual dos fatores que diferem entre si. No entanto, o teste de Nemenyi é muito conservador e pode não encontrar diferença significativa entre os pares testados.

1.5 Justificativa da pesquisa

Ao longo desta dissertação, os seguintes aspectos são abordados visando a previsão e tomada de decisões adequadas para evitar a ocorrência futura de escassez de água.

1.5.1 Contribuições

Após as perguntas de pesquisa apresentadas na subseção 1.2.1, surgem duas contribuições significativas nesta dissertação. A primeira diz respeito à previsão da demanda de água na cidade de Curitiba, abordando aspectos como consumo e gasto de energia durante períodos de pico (conforme mencionado em **Q 5a.** a **Q 5e.**).

Nesse sentido, foram utilizados métodos de previsão de séries temporais, como os modelos ARIMA, ARMA, SARIMA, ARIMAX e SARIMAX, bem como modelos mais simples derivados do ARIMA, como AR, ARX e MA. Além disso, foram explorados modelos regressivos, como LR e RFR, e modelos baseados em gradientes, como XGBoost e LightGBM. Essa variedade de modelos foi selecionada visando uma previsão precisa e eficiente, levando em consideração as demandas relacionadas ao consumo de energia e água pela empresa SANEPAR, com o objetivo de minimizar os gastos associados.

As previsões foram realizadas tanto para o curto prazo (1 a 7 dias) quanto para o longo prazo (14 a 30 dias), a fim de embasar a tomada de decisões estratégicas em relação à demanda de água. Os resultados destacaram que, no longo prazo, os modelos ARIMA tiveram um desempenho superior em comparação aos modelos baseados em gradientes. Por outro lado, os modelos de gradiente mostraram-se mais eficazes nas previsões de curto prazo, como para um dia ou uma semana. Ainda assim, os modelos ARIMA e seus derivados superaram os modelos baseados em gradientes.

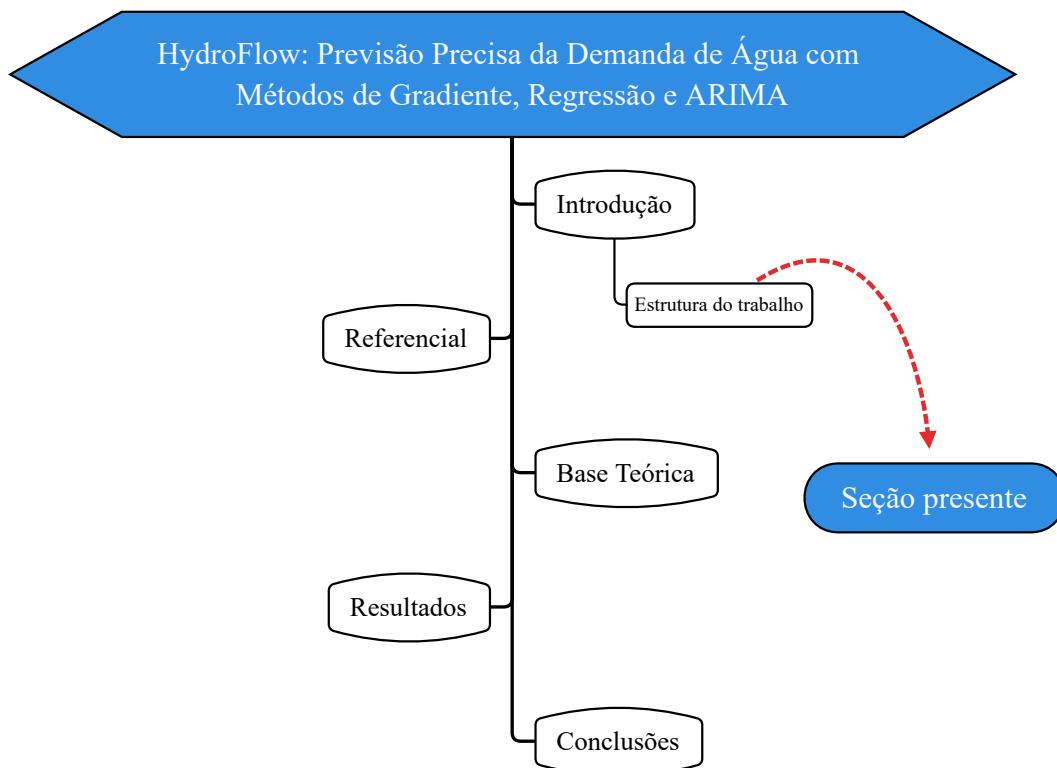
A comparação entre os modelos de previsão desempenha um papel central nesta dissertação. Através do teste estatístico Ljung-Box, é possível avaliar o desempenho de cada modelo ARIMA tanto no curto prazo quanto no longo prazo. No Apêndice B, apresenta-se a comparação dos modelos por meio desse teste estatístico. Além disso, nas

Figuras 39 e 40 do Apêndice A, é realizada a comparação dos modelos regressivos com os modelos ARIMA. Essas análises comparativas são cruciais para a seleção do modelo mais adequado, permitindo uma tomada de decisão embasada para enfrentar o problema em questão.

1.6 Estrutura do trabalho

Este documento está estruturado em 5 capítulos, divididos da seguinte forma:

Figura 3: Estrutura da dissertação



Fonte: Elaboração própria

O trabalho está estruturado em diferentes capítulos, cada um abordando aspectos específicos da pesquisa. O Capítulo 1, Introdução, apresenta a introdução do trabalho, fornecendo uma contextualização do estudo, destacando a motivação e os objetivos a serem alcançados. Também são apresentados o problema em questão, a metodologia utilizada, a justificativa da pesquisa, as contribuições esperadas e a organização do trabalho.

O Capítulo 2, Revisão Teórica, oferece uma visão geral das principais pesquisas e estudos relacionados às questões abordadas na pesquisa. Esse capítulo proporciona uma base teórica sólida para fundamentar a análise e interpretação dos resultados.

No Capítulo 3, Modelos, são apresentados os modelos que serão utilizados para trabalhar com os dados coletados. Essa seção detalha os modelos escolhidos, destacando

suas características e fundamentos teóricos.

O Capítulo 4, Resultados, é dedicado à apresentação dos resultados obtidos ao longo da pesquisa. Além disso, são realizadas análises e interpretações dos resultados, fornecendo insights relevantes para o entendimento do problema em estudo.

Por fim, o Capítulo 5, Conclusões, traz as considerações finais da pesquisa, abordando os principais achados e conclusões alcançadas. Também são apresentadas propostas para pesquisas futuras, visando expandir e aprofundar o conhecimento na área.

Essa estrutura organizada em capítulos permite uma apresentação clara e coerente do trabalho, abrangendo desde a introdução e fundamentação teórica até os resultados e conclusões finais.

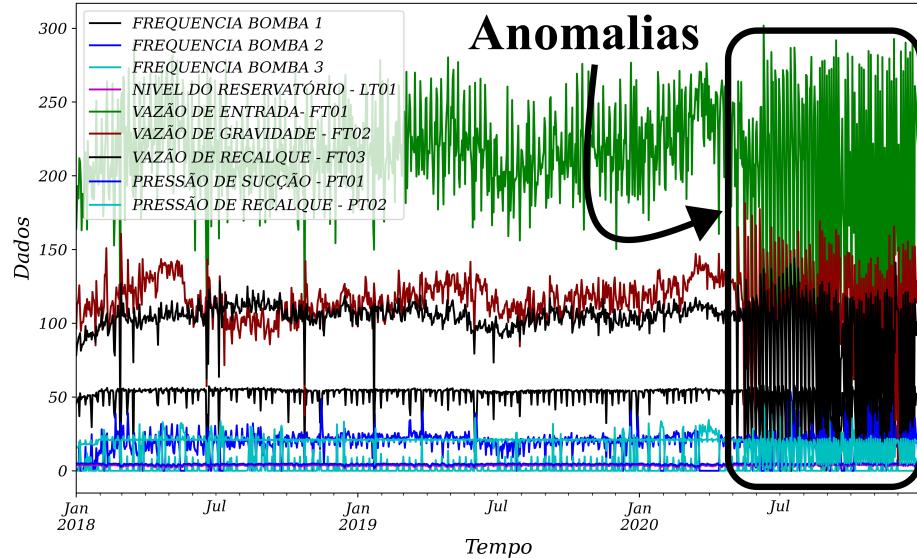
2 Referencial

Este capítulo apresenta a base da literatura coletada durante a preparação desta dissertação. Embora os resultados obtidos sejam mais modestos em comparação a uma tese, eles ainda são relevantes para o trabalho realizado aqui. A revisão bibliográfica realizada consiste em uma análise abrangente e crítica das principais fontes de literatura relacionadas ao tema em questão. Por meio dessa revisão, busca-se obter uma compreensão aprofundada do estado atual do conhecimento na área e identificar lacunas ou oportunidades de pesquisa. Os insights e informações extraídos da literatura são fundamentais para embasar a fundamentação teórica, metodologia e análise dos resultados desta dissertação. Dessa forma, a revisão bibliográfica desempenha um papel crucial no embasamento teórico e na contextualização do trabalho, fornecendo um sólido alicerce para o desenvolvimento e contribuição desta pesquisa.

2.1 Detecção de anomalias

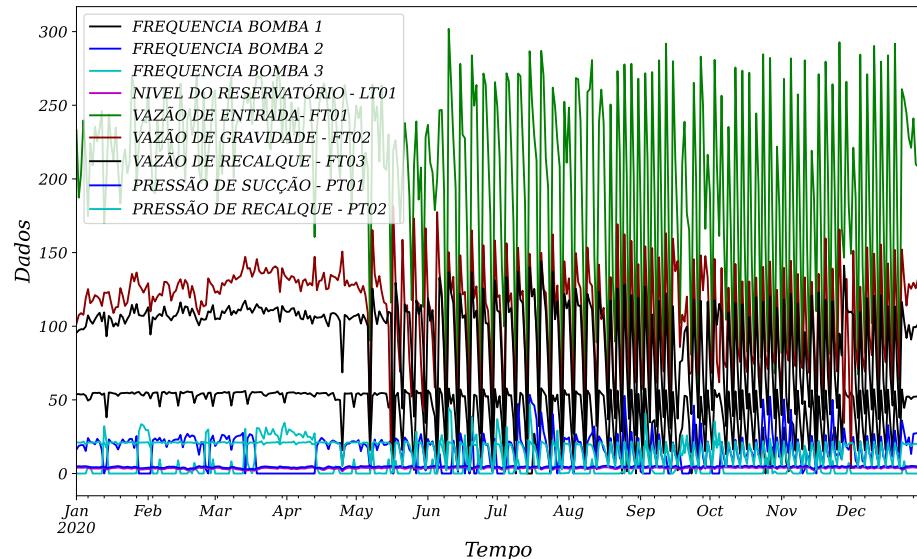
Detectar anomalias em séries temporais representa um desafio significativo para os previsores, pois requer habilidade em identificar mudanças nos dados mesmo quando não estão claramente evidentes. Nesse contexto, a coleta de dados realizada ao longo do tempo pela empresa SANEPAR revela anomalias mais expressivas do que inicialmente imaginado. A escassez de água que afetou a cidade de Curitiba se prolongou por vários dias, como evidenciado pelos gráficos de linha utilizados na etapa de trabalho mencionada (**Etapa 1**). Esses gráficos oferecem uma representação visual clara das variações nos níveis de água ao longo do tempo, auxiliando na compreensão da extensão do problema e na necessidade de uma abordagem adequada.

Figura 4: Dados completos com uma frequência média de 24 horas



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 5: Plotagem de dados para o ano de 2020



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Os dados coletados possuem uma dimensão de 26306 linhas 9 colunas. Essa ampla quantidade de dados será utilizada nos modelos descritos na subseção 1.4 para que seja possível prever e analisar as anomalias evidenciadas nas Figuras 4 e 5. Essas figuras ilustram visualmente as variações e padrões observados nos dados ao longo do tempo, destacando a importância de explorá-los de maneira apropriada a fim de compreender as

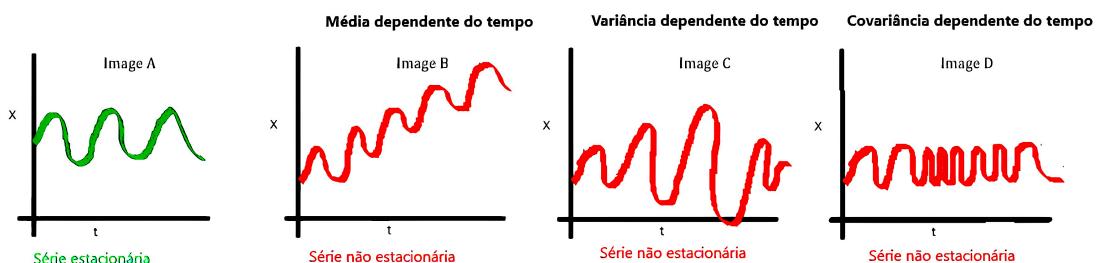
anomalias e embasar a tomada de decisões.

2.2 Revisão sistemática da literatura

As séries temporais desempenham um papel fundamental em diversos campos do conhecimento, como Economia (preços diários de estoques, taxa de desemprego mensal, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (chuvas, temperatura diária, velocidade do vento), entre outros. Ao longo dos anos, têm sido empregadas ferramentas computacionais para tornar a previsão em séries temporais mais eficiente, especialmente com o uso de técnicas de aprendizado de máquina e linguagens de programação como *Python* e *R*, que se destacam por sua capacidade de manipular e analisar dados temporais de forma eficaz.

Para compreender melhor o conceito de série temporal, é possível considerar o exemplo de um maratonista que pratica corrida regularmente ao longo de vários anos e uma pessoa sedentária que decide participar de uma corrida com uma distância máxima de 5 km. Ambos realizam a corrida ao mesmo tempo, utilizando monitores de frequência cardíaca que permitem o acompanhamento médico. Ao analisar os dados desde o início até o final da corrida, é possível observar que a série temporal do maratonista apresentará um comportamento mais estacionário, devido ao seu hábito regular de corrida. Por outro lado, a série temporal da pessoa sedentária será mais não estacionária, como ilustrado na Figura 6. Essa diferença ocorre devido à falta de regularidade na prática de exercícios físicos por parte da pessoa sedentária.

Figura 6: Exemplo de séries temporais

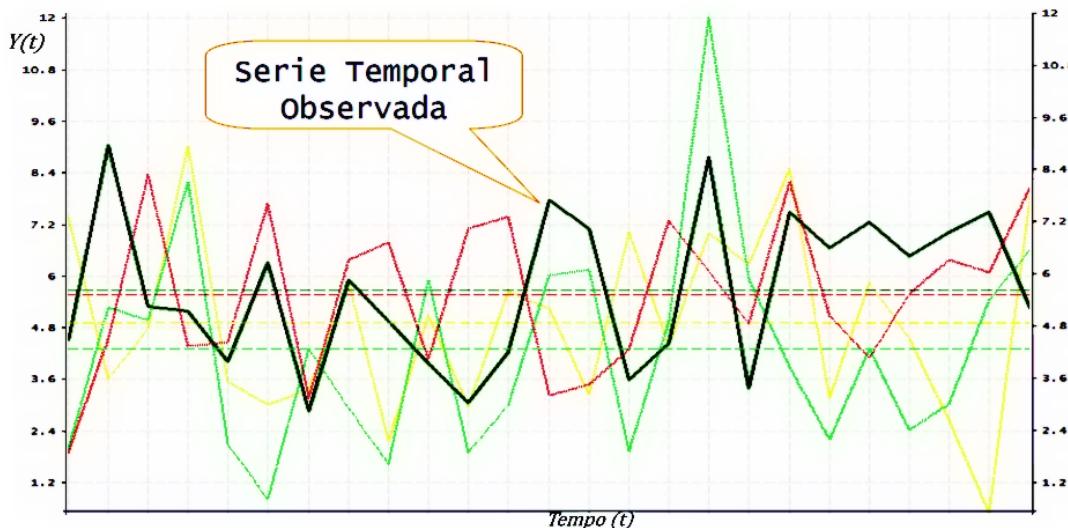


Fonte: (BRANDÃO, 2020)

Na Figura 6, é possível observar que o eixo x representa os dados observados ao longo do tempo, enquanto o eixo t representa o tempo decorrido. Além disso, as séries temporais são caracterizadas como processos estocásticos regidos por leis probabilísticas. Isso implica que elas podem ser concebidas como um conjunto de todas as possíveis trajetórias que uma variável alvo pode seguir, como ilustrado na Figura 6. No entanto,

somente uma dessas trajetórias será observada, de acordo com as características que se manifestaram durante o período analisado. Por exemplo, ao lançar um dado, existem seis possibilidades, mas apenas um número será obtido. Da mesma forma, em séries temporais, há uma infinidade de possibilidades, mas somente uma delas ocorrerá, de acordo com as características que se apresentaram nesse determinado período.

Figura 7: Processo estocástico



Fonte: (PINHEIRO, 2022)

Com $Y(t)$ representando os dados fictícios e $\text{Tempo} (t)$ representando a linha do tempo na Figura 6.

É possível pensar nisso como um conjunto de todas as trajetórias possíveis que poderiam ser observadas para uma variável.

Esta revisão sistemática da literatura aborda o tema das séries temporais, que é de grande relevância em diversas áreas, como ilustrado na Figura 16. Foi realizada uma análise das últimas seis anos para identificar as principais realizações nesse campo dentro desse curto período de tempo disponível. A seleção dos artigos foi baseada em critérios específicos, levando em consideração a relevância dos autores, os anos de atividade, os países com maior número de publicações e as palavras-chave mais frequentes.

O objetivo dessa revisão é analisar uma literatura selecionada, porém altamente relevante. Embora a série temporal tenha como foco a análise e modelagem da dependência temporal, considerando a ordem apresentada nas bases de dados, os artigos revisados também exploram o uso de técnicas de aprendizado de máquina em aplicações relacionadas.

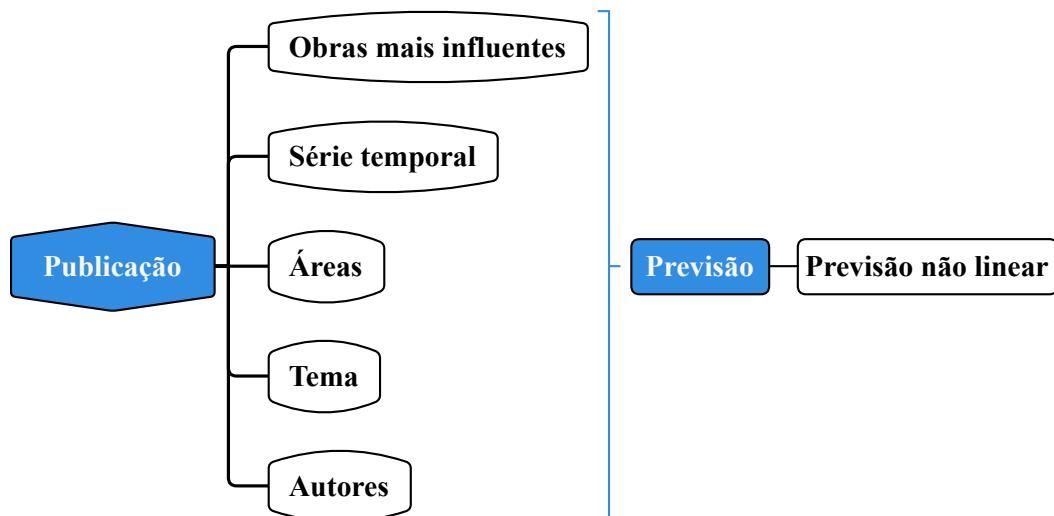
Embora nem todos os artigos revisados tenham uma forte relação com aprendizado de máquina, eles contribuem cientificamente para este trabalho e podem servir como base para outros pesquisadores. Essas análises fornecem uma visão básica para alguns

leitores que ainda não estão familiarizados com o conceito de séries temporais ou revisões sistemáticas da literatura.

2.3 Problematização da Revisão

Nesta subseção, é discutido um problema de pesquisa que pode ser compreendido por diversos leitores. A Figura 8 apresenta um mapa conceitual das publicações, destacando a importância dos autores como base para esta revisão. Os modelos propostos por esses autores são fundamentais para abordar o problema em questão, uma vez que a previsão em séries temporais é um desafio de grande significado por si só.

Figura 8: Mapa conceitual do problema de pesquisa



Fonte: Elaboração própria

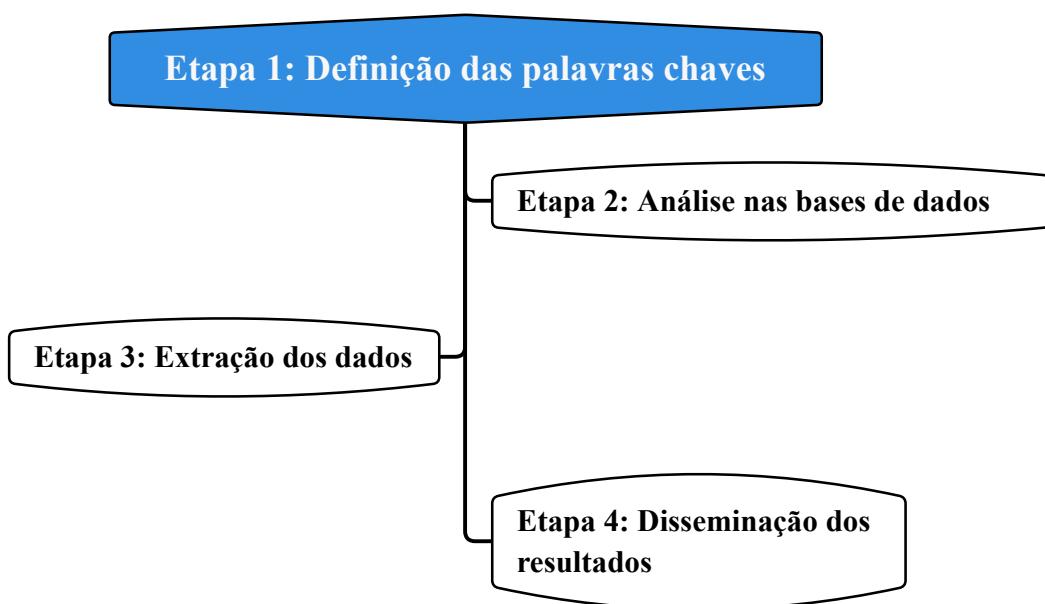
O mapa conceitual apresentado na Figura 8 ilustra a relação entre as palavras-chave que estão relacionadas ao problema em questão, proporcionando uma visão clara do que será abordado ao longo do trabalho. Esse mapa contribui para a identificação dos principais tópicos de pesquisa e das questões que serão exploradas posteriormente.

- Q 1** Quais os autores que mais publicam sobre o assunto de séries temporais?
- Q 2** Quais os países que mais publicam sobre o assunto?
- Q 3** Quais as áreas que mais publicam sobre o tema?
- Q 4** Quais são as obras mais influentes na análise de séries temporais?

2.4 Metodologia

Nesta subseção, é fornecida uma explicação detalhada de como a revisão foi conduzida, abrangendo desde a análise do banco de dados até a conclusão final da revisão. São apresentados os passos e critérios adotados para a seleção dos artigos, bem como os procedimentos utilizados para a extração e análise dos dados. A subseção visa esclarecer de forma clara e objetiva todo o processo metodológico empregado durante a realização da revisão.

Figura 9: Etapas da Revisão.



Fonte: Adaptado de Martins e Gorscheck (2016)

Etapa 1 A Figura 9 apresenta uma adaptação da metodologia proposta por Martins e Gorscheck (2016) para a realização desta revisão sistemática. Inicialmente, foram realizadas buscas nos bancos de dados Scopus, Web of Science e Lens, selecionando algumas bases relevantes para o tema da pesquisa.

Campo de pesquisa Scopus

TITLE-ABS-KEY ("time series forecasting") AND **TITLE-ABS-KEY** ("time series analysis") AND (**LIMIT-TO** (**DOCTYPE** , "ar")) AND (**LIMIT-TO** (**LANGUAGE** , "English")) AND (**LIMIT-TO** (**PUBYEAR** , 2022) OR **LIMIT-TO** (**PUBYEAR** , 2021) OR **LIMIT-TO** (**PUBYEAR** , 2020) OR **LIMIT-TO** (**PUBYEAR** , 2019) OR **LIMIT-TO** (**PUBYEAR** , 2018) OR **LIMIT-TO** (**PUBYEAR** , 2017))

Campo de pesquisa na Web of Science

“times series forecasting”(All Fields) and “time series analysis”(All Fields) (Publication Years: 2022 or 2021 or 2020 or 2019 or 2018 or 2017) (Document Types: Articles) (Languages: English)

Campo de pesquisa de Lens

Scholarly Works (11) = (“time series forecasting”) AND ((“time series analysis”) AND (“nonlinear forecasting”)) Filters: Year Published = (2016 - 2022) Publication Type = (journal article)

Para todas as bases de busca, foram considerados os últimos 6 anos, com exceção do Lens, que retornava poucos artigos. Nesta etapa, foram utilizadas palavras-chave que se adequam melhor à pesquisa, como *time series forecasting and time series analysis and nonlinear forecasting*.

Etapa 2 No cruzamento das palavras-chave, obteve-se um número considerável de artigos, sem restringir a área em que cada um pode ser publicado. A Tabela 1 apresenta a tabulação dos resultados obtidos, sem excluir duplicatas, que serão tratadas na seção 2.5.

Etapa 3 Nesta etapa, é realizada uma avaliação preliminar de cada artigo obtido, sem aplicar nenhum filtro anual nas buscas. Analisar todos os artigos dessa maneira resultaria em um número elevado, por exemplo, no banco de dados Scopus seriam 498 artigos, na Web of Science seriam 140 artigos e no Lens, que retorna poucos artigos, seriam 11 artigos, totalizando 649 artigos sem remover duplicatas. É importante ressaltar que esses artigos passaram apenas pelo filtro de idioma inglês e de serem artigos, visando aprimorar a busca e a tomada de decisões. Ao aplicar o filtro dos últimos 6 anos, obteve-se um número mais gerenciável de artigos para análise. Levando em consideração a diferença entre essa estimativa apresentada na Tabela 1 e a quantidade de artigos restantes após a remoção de duplicatas, temos menos de 356 artigos para análise. É válido lembrar que, ao remover as duplicatas, esse número pode diminuir ainda mais, atingindo o objetivo proposto neste trabalho.

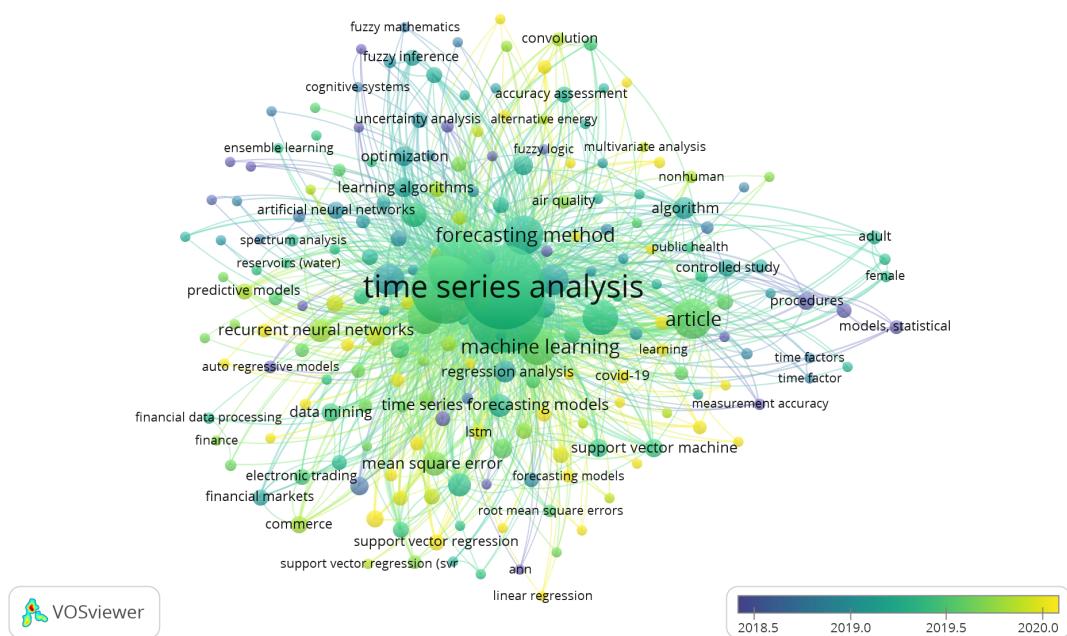
Etapa 4 Nesta etapa, é realizada uma análise mais aprofundada do conteúdo dos artigos selecionados, levando em consideração as áreas de especialização e correlação com séries temporais. Como esta revisão está inserida no contexto de um programa de mestrado em Engenharia de Produção e Sistemas, vale a pena analisar a correlação com áreas como Matemática. A Figura 16 mostra que as áreas mais relevantes para a pesquisa são **Informática, Engenharia e Matemática**, representando 50% das publicações. Portanto, a pesquisa está alinhada com a utilização de conceitos

matemáticos básicos para realizar uma estimativa do número de artigos que podem ser eliminados. Estima-se que cerca de 481 artigos possam ser excluídos, porém essa estimativa não possui uma base sólida. Utilizando o software Mendeley Desktop para obter o número exato de artigos sem duplicatas, chegou-se a um total de 308 artigos.

2.5 Resultados da busca de revisão

Nesta seção, serão apresentados os resultados da pesquisa, utilizando um software para melhor aproveitamento de cada banco de dados utilizado no trabalho. Inicialmente, foi realizada uma análise no software VOSviewer.

Figura 10: Palavras-chave mais populares na Scopus.

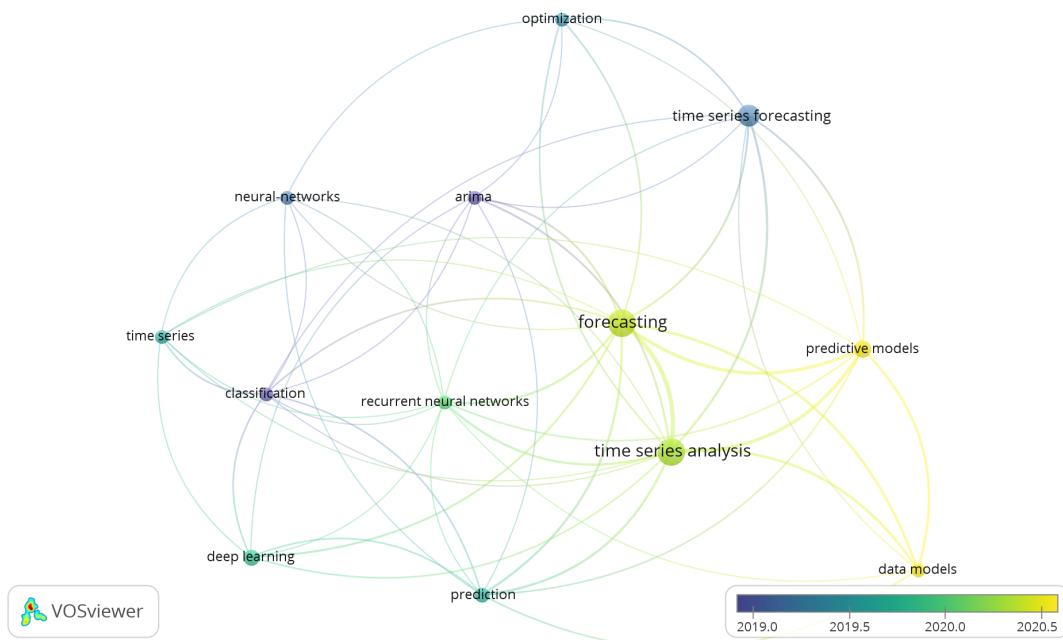


Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

A Figura 10 mostra uma lista das palavras mais frequentemente utilizadas como sinônimos ou em conjunto com *time series analysis* nos artigos. A análise da base de dados do Scopus foi feita com uma ferramenta que exibe as palavras-chave relacionadas em cada campo de busca, proporcionando uma visão abrangente das correlações com as palavras-chave principais.

Nesse primeiro momento, foram obtidas 3.484 palavras-chave, sendo que 212 delas atingiram o limite estabelecido. É importante destacar que as palavras-chave base utilizadas foram “*time series forecasting and time series analysis*” no Scopus.

Figura 11: Palavras-chave mais populares na WoS



Fonte: Elaboração própria a partir de dados da Web of Science (2018 a 2020)

A análise do banco de dados Web of Science, apresentada na Figura 11, também foi realizada por meio de uma ferramenta que mostra as palavras-chave relacionadas em cada campo de busca. Mais uma vez, é possível obter uma visão ampla das correlações com as palavras-chave principais.

Nesse primeiro momento, foram obtidas 305 palavras-chave, sendo que 13 delas atingiram o limite estabelecido. É importante ressaltar que as palavras-chave base utilizadas foram “*time series forecasting and time series analysis*” na Web of Science.

O banco de dados Lens não será apresentado aqui, pois, embora seja uma excelente fonte, não retornou muitos resultados na pesquisa realizada. O site do Lens retornou apenas 11 artigos com os filtros aplicados. Na **Etapa 1** apresenta o campo de busca utilizado nessa pesquisa, resultando nos 11 artigos encontrados.

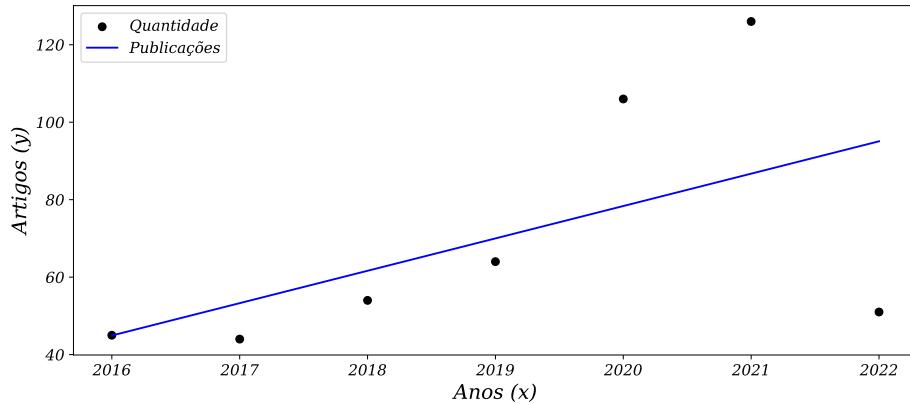
A Tabela 1 apresenta as palavras-chave utilizadas em cada base de dados, juntamente com o número de artigos encontrados inicialmente. No entanto, é importante ressaltar que esses dados ainda não foram processados para remover duplicatas. Após a utilização do software *Mendeley* para eliminar as duplicações, restaram 308 artigos únicos, os quais serão considerados nesta revisão.

Tabela 1: Cruzamento de palavras-chave através da aplicação de filtros de ano e de linguagem

Bases	Palavras Chaves				Resultado
Scopus	time series	AND	time series		490
	forecasting		analysis		
Web of Science	nonlinear	AND	time series		8
	forecasting		forecasting		
Lens	time series	AND	time series		126
	forecasting		analysis		
	nonlinear	AND	time series		14
	forecasting		forecasting		
Lens	time series	AND	time series	AND	11
	forecasting		analysis		
		Total			649

Fonte: Elaboração própria

Figura 12: Analise das quantidades de artigos em relação aos anos.



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

A Figura 12 apresenta um gráfico que ilustra a relação entre o número de artigos publicados e os anos correspondentes. Foi realizada uma análise utilizando regressão linear para examinar essa relação ao longo do tempo.

A equação de regressão linear obtida foi a seguinte:

$$y(x) = 8,3571x - 16.803 \quad \text{com } R^2 = 0,3062 \quad (4)$$

Na equação (4), $y(x)$ representa a equação da reta, onde x é a variável independente que corresponde aos anos. O coeficiente angular da reta é de 8,3571, e o coeficiente linear é de -16.803, que indica o ponto de intersecção com o eixo y.

O coeficiente de determinação, R^2 , é utilizado para avaliar a proporção da variação na variável dependente (número de artigos) que pode ser explicada pela variação na variável independente (anos). Nesse caso, o valor de R^2 foi de 0.3062, o que indica que aproximadamente 30,62% da variação nos números de artigos pode ser explicada pela passagem do tempo.

O coeficiente de determinação mede a relação entre a variável dependente e as variáveis independentes, representando a porcentagem da variação explicada pela regressão em relação à variação total. Quando o R^2 é igual a 1, todos os pontos observados estão exatamente na reta de regressão, indicando um ajuste perfeito, ou seja, todas as variações em y são totalmente explicadas pela variação em x_n através da função especificada, sem desvios em torno da função estimada. Por outro lado, quando o R^2 é igual a 0, conclui-se que as variações em y são exclusivamente aleatórias e a inclusão das variáveis x_n no modelo não fornece nenhuma informação sobre as variações em y .

A fórmula do coeficiente de determinação R^2 é dada pela equação:

$$R^2 = \frac{\left(\sum X \cdot Y - \frac{\sum X \cdot \sum Y}{n} \right)^2}{\left[\sum X^2 - \frac{(\sum X)^2}{n} \right] \cdot \left[\sum Y^2 - \frac{(\sum Y)^2}{n} \right]} = (r)^2 \quad (5)$$

Na equação (5), X e Y representam as coordenadas no plano cartesiano, como por exemplo, o par ordenado (x, y) . Na análise realizada com a relação entre o número de artigos e os anos, obteve-se um valor de $R^2 = 30\%$, o que implica que a linha de regressão é influenciada pelo valor encontrado de R^2 .

Embora seja uma análise simples da relação entre o número de artigos e os anos, essa é uma validação significativa para observar o teste F de significância, que deve ser sempre inferior a 5%, também conhecido como valor-p. Com base nesses valores, é possível analisar o significado da linha de regressão e observar que o ano de 2021 foi o ano em que a maioria dos artigos foi publicada sobre o tema das séries temporais.

Tabela 2: Fator de impacto

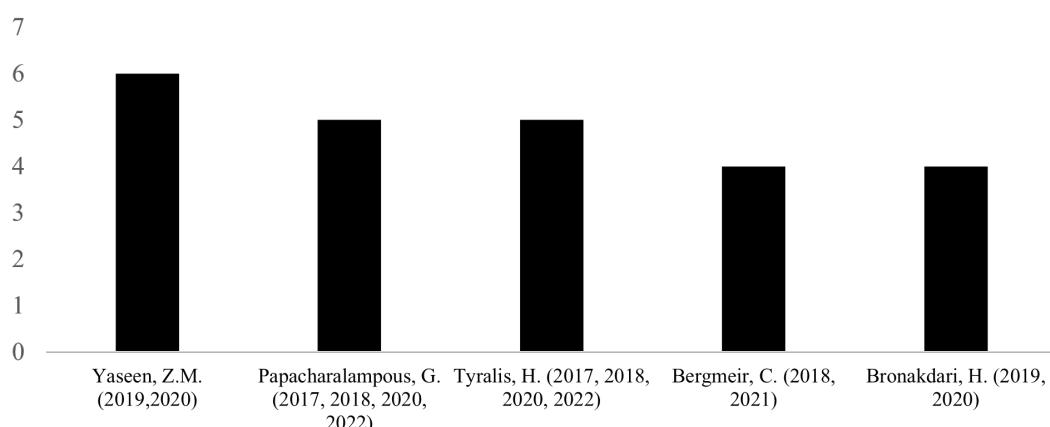
Revista científica	Quantidade de publicação	Qualidade da revista	H-INDEX
Neurocomputing	27	Q1	143
IEEE Access	18	Q1	127
Applied Soft Computing	12	Q1	143
Energies	11	Q2	93
Energy	11	Q1	343

Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

Na Tabela 2, são apresentadas as revistas que mais publicam artigos sobre o tema em questão. É importante destacar que muitas dessas revistas estão localizadas fora do Brasil e têm seus nomes em inglês. No entanto, todas as revistas listadas, incluindo aquelas com um alto fator de impacto, como a categoria **Q1**, apresentam uma correlação significativa com as áreas de **informática, engenharia e matemática**.

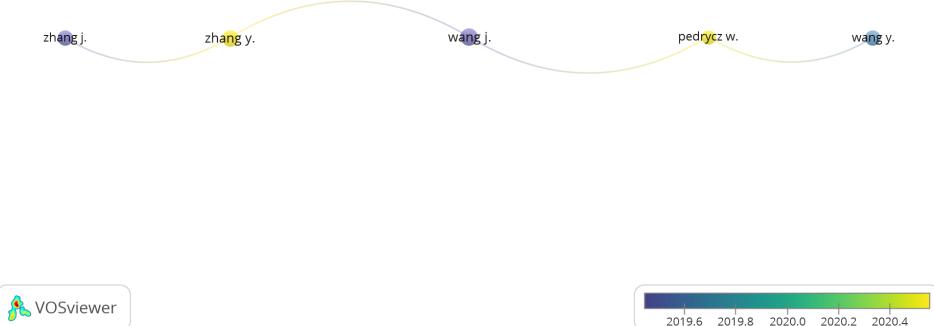
Essa observação ressalta a importância dessas áreas de especialização na pesquisa sobre séries temporais, uma vez que elas estão fortemente representadas nas principais revistas científicas. Essas revistas desempenham um papel fundamental na disseminação do conhecimento e no avanço do campo, garantindo a qualidade e o impacto dos artigos publicados. Portanto, é valioso direcionar a atenção para essas revistas, uma vez que elas são reconhecidas como fontes confiáveis e respeitadas dentro da comunidade científica.

Figura 13: Relação de autores entre artigos publicados



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

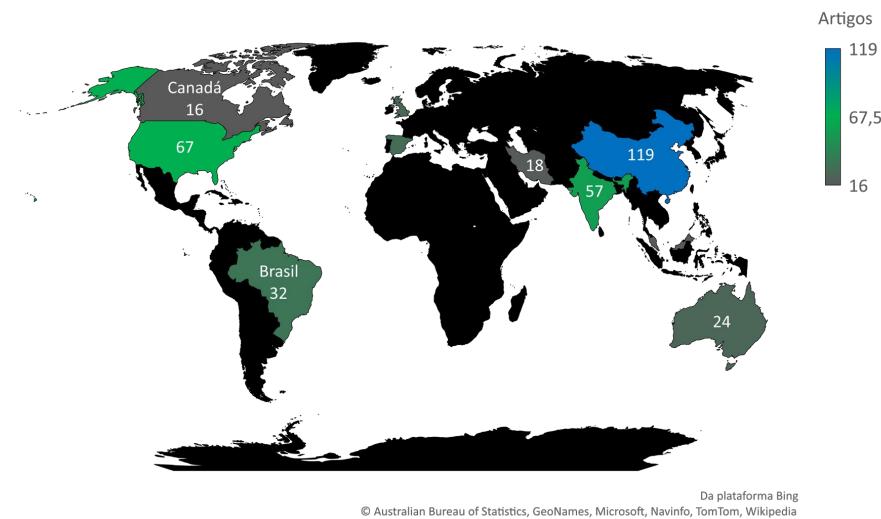
Figura 14: Ligação bibliográfica entre os autores



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

Em resposta à questão colocada anteriormente (**Q 1**), foi utilizada a Figura 13 para visualizar de forma mais clara os autores que mais publicaram sobre o tema em análise. O gráfico apresenta um histograma que destaca os autores cujo número de publicações é maior que 4 durante o período de 2016 a 2022. Essa abordagem visa evitar a inclusão de todos os autores e destacar aqueles que tiveram uma contribuição significativa no campo, considerando o critério estabelecido de pelo menos 4 publicações. Dessa forma, é possível identificar os principais autores que se destacam nesse tópico específico, fornecendo uma visão geral da distribuição da produção científica entre os pesquisadores.

Figura 15: Mapa mundial da publicação de artigos em todo o mundo

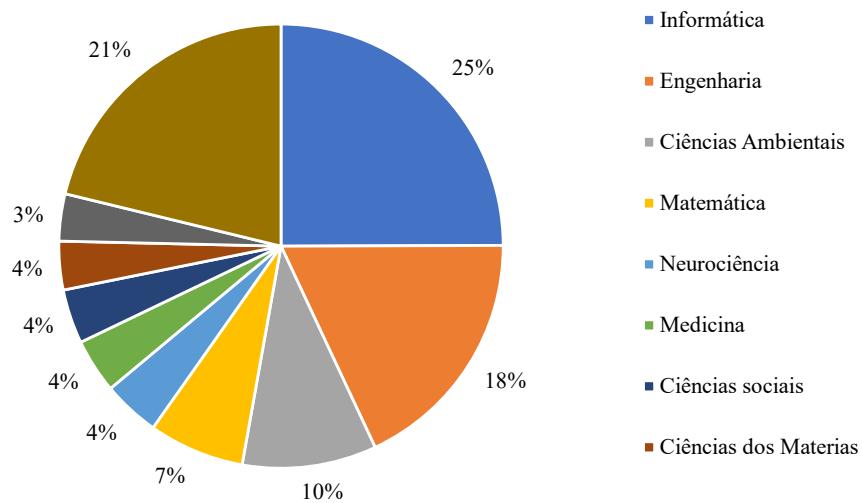


Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

A pergunta de pesquisa **Q 2** foi abordada por meio da análise da Figura 15, que apresenta os países com maior número de publicações sobre o assunto em escala, ordenados de forma decrescente. Os principais países que se destacam nessa análise são os seguintes: China, com 119 publicações; Estados Unidos, com 67 publicações; Índia, com 57 publicações; Brasil, com 32 publicações; Espanha, com 28 publicações; Reino Unido, com 25 publicações; Austrália, com 24 publicações; Irã, com 18 publicações; Malásia, com 17 publicações; e Canadá, com 16 publicações.

É importante ressaltar que o mapa não exibe todos os países e seus respectivos números de publicações, mas destaca aqueles com maior produção nesse contexto específico. Essa análise ajuda a identificar os países com maior contribuição científica nessa área de estudo, fornecendo insights sobre os locais onde a pesquisa sobre séries temporais tem sido mais ativa.

Figura 16: Áreas de aplicação do tema



Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

Para responder à pergunta de pesquisa **Q 3**, foi criado um gráfico circular, apresentado na Figura 16, que ilustra as áreas com maior número de publicações durante o período analisado na revisão. A Tabela 3 complementa o gráfico, fornecendo os valores específicos de cada área e a quantidade de publicações correspondente.

O gráfico circular oferece uma representação visual clara das áreas que se destacam em termos de produção científica no campo das séries temporais. Ao examinar a tabela, é possível identificar as áreas com maior número de publicações, permitindo uma compreensão aprofundada das principais áreas de conhecimento relacionadas ao tema. Essa análise contribui para uma melhor compreensão da distribuição de publicações e áreas de pesquisa ao longo do período estudado.

Tabela 3: Áreas e seus valores respetivos de artigos em cada área.

Informática	240
Engenharia	174
Ciências Ambientais	94
Matemática	67
Neurociência	40
Medicina	38
Ciências sociais	38
Ciências dos Materiais	34
Negócios, Gestão e Contabilidade	33
Outros	204

Fonte: Elaboração própria a partir de dados da Scopus, len e Web of Sicence (2016 a 2022)

Na última pergunta de pesquisa, referente à **Q 4**, foi realizada uma investigação dos artigos mais influentes na revisão. Esses artigos retratam alguns dos métodos utilizados por renomados autores Golyandina (2020), Kumar, Jain e Singh (2021), Xie et al. (2019), Lara-Benitez, Carranza-Garcia e Riquelme (2021), Ahmad et al. (2018), Carvalho Jr. e Costa Jr. (2019), Tan et al. (2021), Liu e Chen (2019), Liu et al. (2021), Rossi (2018), Soyer e Zhang (), Martinović, Hunjet e Turcin (2020), Ursu e Pereau (2016), Wang et al. (2016), Shih, Sun e Lee (2019), Moon et al. (2019), Chou e Tran (2018), Bergmeir, Hyndman e Koo (2018), Boroojeni et al. (2017), Chou e Nguyen (2018), Coelho et al. (2017), Du et al. (2020), Sadaei et al. (2019), Salgotra, Gandomi e Gandomi (2020), Tyralis e Papacharalampous (2017), Vlachas et al. (2020), Yang et al. (2019), Shen et al. (2020), Sezer, Gudelek e Ozbayoglu (2020), Chen et al. (2018), Buyukahin e Ertekin (2019), Li e Bastos (2020), Kulshreshtha e Vijayalakshmi (2020), Samanta et al. (2020), Xu et al. (2019), Graff et al. (2017), Taieb e Atiya (2016).

Esses artigos abordam diferentes métodos usados pelos autores para previsão de séries temporais e análise não-linear dessas previsões. Eles representam contribuições significativas para o avanço do conhecimento e aplicação prática das séries temporais, oferecendo insights valiosos sobre abordagens eficazes nesse campo. Ao incluir esses estudos influentes na análise, obtém-se uma visão abrangente dos métodos e técnicas mais relevantes na previsão de séries temporais.

No estudo conduzido por Xu et al. (2019), um modelo híbrido foi proposto, combinando o modelo linear AR e LR com o modelo não-linear ARIMA e o modelo DBN. Essa abordagem permite capturar tanto os comportamentos lineares quanto os não-lineares de

uma série temporal. Por outro lado, Li e Bastos (2020) comparou o desempenho de previsão da abordagem MAELS com outros modelos de aprendizado de máquina de última geração, como CNN, RNN, LSTM, ARIMA e SVM-VAR. As abordagens CNN, RNN e LSTM são capazes de lidar com dados multivariados de entrada e saída, enquanto o ARIMA utiliza informações passadas para prever o futuro com base em características como autocorrelação e médias móveis.

Dessa forma, por meio dessa revisão sistemática e análise de conteúdo, a pergunta de pesquisa formulada no início do capítulo foi respondida.

Além desses modelos mencionados, também será utilizada a versão atualizada do ARIMA nesta dissertação. Os modelos SARIMA e SARIMAX também serão comparados para determinar qual deles é o mais adequado. Além disso, serão empregados os modelos Light GBM e XGBoost. Quanto às métricas de erro, serão utilizadas MAE, MAPE e RMSE, que são amplamente adotadas na literatura. O coeficiente de determinação (R^2), mencionado na equação (5), não é tão comumente utilizado para comparação de modelos de previsão futura.

2.6 Principais conclusão

A conclusão abrangente da pesquisa de revisão revela que diversas bases de dados foram consultadas, como Scopus, Web of Science e Lens. Cada uma dessas bases proporcionou uma quantidade significativa de artigos relevantes, que foram minuciosamente analisados. Essa abordagem permitiu responder à pergunta de pesquisa formulada no início da revisão.

Apesar da base de dados Lens ser menor em comparação com as demais, também foram encontrados artigos relevantes que contribuíram para enriquecer o processo de dissertação. Além disso, o uso de software especializado foi essencial para lidar com a grande quantidade de artigos e suas inter-relações.

No âmbito específico da revisão sistemática, a análise de séries temporais recebeu uma ênfase particular, com um enfoque aprofundado e atualizado nos últimos seis anos. Os resultados obtidos foram altamente relevantes e significativos. Por meio do cruzamento de palavras-chave e da aplicação de filtros específicos, foram selecionados 308 artigos publicados entre 2016 e 2022.

Com o objetivo de refinar ainda mais a análise, foi realizado um filtro adicional com base em áreas de interesse, como matemática, engenharia e informática. Isso resultou na seleção de 481 artigos relacionados a essas áreas, excluindo aqueles de outras áreas não pertinentes.

A pesquisa de revisão realizada foi minuciosa e abrangente, proporcionando uma

base sólida de artigos relevantes para o desenvolvimento da dissertação. Os resultados obtidos foram fundamentais para orientar as próximas etapas do trabalho e para alcançar uma compreensão aprofundada do tema das séries temporais.

3 Base Teórica

Um dos pilares fundamentais para obter resultados satisfatórios é contar com uma base sólida de conhecimento. Neste capítulo, são abordados diversos aspectos importantes, como métricas de erro e modelos regressivos de previsão. Essas métricas desempenham um papel crucial na avaliação e comparação dos modelos, permitindo uma análise precisa do desempenho de cada um. Além disso, os modelos regressivos de previsão são explorados, fornecendo insights valiosos sobre como essas técnicas podem ser aplicadas para realizar previsões com precisão. Compreender e dominar esses conceitos é essencial para se obter resultados confiáveis e embasar as próximas etapas do trabalho.

3.1 Métricas de Erros

A métrica de Erro Quadrático Médio (MSE) é amplamente utilizada no campo do aprendizado de máquina para avaliar a qualidade dos modelos de previsão. O MSE é calculado pela média da soma dos quadrados das diferenças entre os valores reais e os valores previstos. Sua fórmula é a seguinte:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

Nessa fórmula, n representa o número de amostras, y_i é o valor real correspondente à amostra i e \hat{y}_i é o valor previsto para a mesma amostra. O MSE é calculado como a média das diferenças ao quadrado entre os valores reais e os valores previstos.

A utilização do MSE fornece uma medida quantitativa da precisão do modelo, pois penaliza de forma mais significativa os erros maiores. Ao elevar as diferenças ao quadrado, a métrica enfatiza a importância de minimizar as discrepâncias entre os valores reais e os valores previstos. Dessa forma, quanto menor o valor do MSE, melhor é o desempenho do modelo em termos de previsão.

Portanto, o MSE é uma métrica fundamental para avaliar a qualidade dos modelos de previsão e é amplamente utilizada para comparar diferentes algoritmos e abordagens de aprendizado de máquina.

3.1.1 Erro quadrático médio raiz (RMSE)

O RMSE é uma métrica amplamente empregada na avaliação de modelos de previsão em séries temporais. Ele é calculado tomando a raiz quadrada do MSE, conforme mostrado na seguinte fórmula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

Na equação (7), n representa o número de amostras, y_i é o valor real correspondente à amostra i , e \hat{y}_i é o valor previsto para a mesma amostra. O RMSE fornece uma medida da dispersão média entre os valores reais e os valores previstos pelo modelo.

Uma das vantagens de utilizar o RMSE é que, ao computar a raiz quadrada, o erro passa a ter a mesma escala da variável de interesse. Isso permite uma interpretação mais fácil dos resultados, sendo que um valor baixo de RMSE indica um bom desempenho do modelo, já que o erro se aproxima de zero.

O RMSE possui algumas características positivas. Ele penaliza de forma significativa os valores discrepantes, caso seja necessário para o modelo. Além disso, o erro resultante está nas mesmas unidades da série temporal, facilitando a interpretação. O RMSE pode ser considerado uma combinação das melhores características do MSE e do Erro Absoluto Médio (MAE).

No entanto, o RMSE também apresenta algumas desvantagens. Ele tem uma interpretabilidade reduzida, uma vez que os erros ainda são elevados ao quadrado. Além disso, o RMSE é dependente da escala dos dados, o que impede sua comparação direta com modelos de séries temporais que utilizam unidades diferentes.

Apesar das limitações, o RMSE é uma métrica amplamente utilizada para avaliar modelos de previsão em séries temporais. Ele fornece uma medida da dispersão média entre os valores reais e previstos, auxiliando na compreensão do desempenho do modelo e na comparação com outras abordagens.

3.1.2 Erro Absoluto Médio (MAE)

O Erro Absoluto Médio (MAE) é amplamente utilizado como uma métrica para avaliar o desempenho de modelos de previsão. Em vez de calcular a média das diferenças entre os valores reais e previstos, o MAE calcula a média dos valores absolutos dessas diferenças, garantindo que os erros positivos e negativos não se anulem.

O MAE mede o desvio médio das previsões em relação aos valores reais e é uma métrica intuitiva e fácil de interpretar, representando a magnitude média dos erros em relação à escala dos dados. Por exemplo, um MAE de 2 significa que, em média, as previsões têm um desvio absoluto de 2 unidades em relação aos valores reais.

Uma das vantagens do MAE é a sua insensibilidade a valores extremos, pois trata os erros de forma absoluta. No entanto, como o MAE não considera a magnitude dos

erros individuais, pode não refletir adequadamente a gravidade de desvios significativos em relação aos valores reais.

Para superar essa limitação, uma alternativa é o Erro Médio Absoluto Percentual (MAPE). O MAPE expressa o MAE como uma porcentagem em relação aos valores reais, proporcionando uma medida relativa de erro. Essa métrica é especialmente útil quando se deseja avaliar o desempenho de um modelo em relação à magnitude dos dados.

Em resumo, o MAE é uma métrica simples e fácil de interpretar, que mede o desvio médio das previsões em relação aos valores reais. O MAPE, por sua vez, fornece uma medida relativa de erro, expressa como uma porcentagem dos valores reais. A escolha entre essas métricas depende do contexto do problema e dos requisitos específicos de avaliação.

O cálculo do MAE é realizado utilizando o valor absoluto da diferença entre o valor real e o valor previsto, e em seguida, divide-se pela quantidade n de amostras. Isso resulta no erro médio absoluto. A equação do MAE é dada por:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (8)$$

Sua interpretação é similar ao RMSE, em que o erro é expresso na mesma escala ou ordem de grandeza da variável estudada.

3.1.3 Erro Percentual Absoluto Médio (MAPE)

Com certeza! Aqui está uma versão aprimorada do texto:

O MAPE é uma métrica que expressa o erro de previsão como uma porcentagem relativa ao valor observado. Ele é calculado somando as diferenças entre o valor real e o valor previsto (representando o erro), dividido pelo valor observado.

O MAPE é calculado usando a seguinte fórmula:

$$MAPE = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

No entanto, surge um problema quando o valor observado y_i é igual a zero, pois é matematicamente impossível dividir por zero. O MAPE é uma medida de erro em que valores menores indicam um melhor desempenho de previsão.

Uma alternativa ao MAPE é calcular $1 - \text{MAPE}$, que representa a porcentagem de acerto.

O Erro Médio Percentual Absoluto é a diferença percentual entre o valor real e

o valor previsto. É comumente usado como uma métrica de referência para avaliar o desempenho de modelos de previsão.

Prós:

- Fácil de interpretar
- Independente de escala, permitindo comparações entre diferentes séries temporais

Contras:

- Erro infinito se o valor real estiver próximo ou igual a zero
- Previsões mais baixas estão propensas a ter um erro de 100%, enquanto previsões mais altas podem ter um erro infinito, o que resulta em um viés de subprevisão.

3.2 Modelos de séries temporais univariados

A previsão de séries temporais é um desafio complexo, sem uma resposta fácil. Existem inúmeros modelos estatísticos que afirmam superar uns aos outros, mas nunca está claro qual modelo é o melhor.

Dito isto, os modelos baseados em ARMA são frequentemente uma boa opção para iniciar. Eles podem alcançar pontuações decentes na maioria dos problemas de séries temporais e são adequados como modelos de referência em tais problemas.

Quanto ao modelo ARIMA, ele é dividido em três componentes: AR (Autoregressão), I (Integração) e MA (Média Móvel). O componente AR leva em consideração os valores anteriores da série temporal, o componente I trata das diferenças entre os valores observados para tornar a série estacionária, e o componente MA considera os erros residuais do modelo. Esses componentes combinados ajudam a capturar os padrões e tendências presentes na série temporal.

3.2.1 Componente Autorregressivo

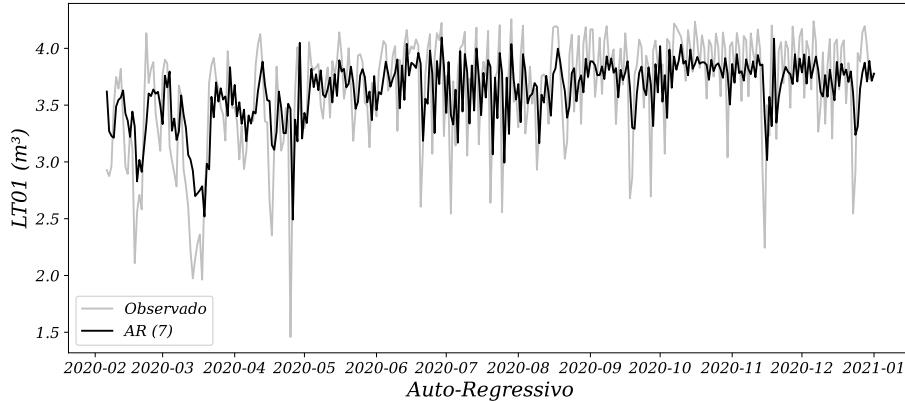
O componente autoregressivo do modelo ARIMA é representado por AR(p), em que o parâmetro p determina o número de séries temporais defasadas utilizadas.

A equação do modelo AR(p) é expressa da seguinte forma:

$$Y_t = c + \sum_{n=1}^p \alpha_n Y_{t-n} + \varepsilon_t \quad (10)$$

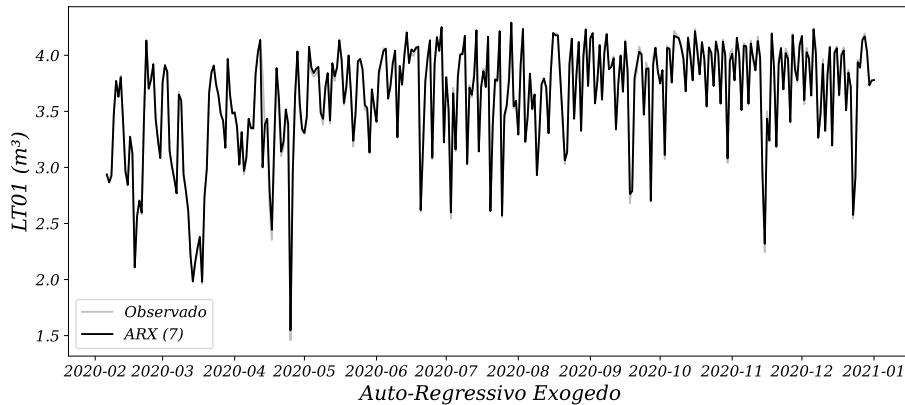
A partir dos dados, é possível obter uma previsão utilizando o modelo AR(7).

Figura 17: Modelo AR(7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 18: ARX (7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na equação (10), o termo ε_t representa o ruído branco. Essa equação pode ser entendida como uma regressão múltipla, em que os valores defasados de y_t são utilizados como preditores. Esse modelo é conhecido como modelo autorregressivo de ordem p , ou AR(p).

A Figura 17 tem como objetivo apresentar uma previsão de um passo à frente (um dia). Nos apêndices C, pode-se observar uma comparação entre os modelos AR, MA e ARX.

O modelo ARX é uma extensão do modelo AR, que incorpora variáveis exógenas nos dados para melhorar as previsões futuras. Esse modelo também é multivariado, como mostrado na subseção 3.3, e foi incluído aqui para fins de comparação com o modelo AR simples, considerando a presença de variáveis exógenas.

Embora o modelo AR possa ser visualmente adequado para a previsão que está

sendo feita, é importante destacar que, por ser um modelo autorregressivo, ele realiza previsões lineares e não captura padrões não lineares presentes nos dados. Para uma análise mais abrangente da série temporal, é necessário considerar exemplos de casos gerais.

3.2.2 AR(0): Ruído branco

Se o parâmetro p for definido como zero (AR(0)), significa que não há termos autorregressivos no modelo. Nesse caso, a série temporal se comporta como um ruído branco. Cada ponto de dados é amostrado de uma distribuição com média zero e variância igual a sigma-quadrado. Isso resulta em uma sequência de números aleatórios que não exibem nenhum padrão ou correlação.

Essa propriedade do ruído branco pode ser útil em análises estatísticas, pois serve como uma hipótese nula. Ao comparar diferentes modelos ou testar a presença de padrões em uma série temporal, podemos usar o ruído branco como referência para avaliar se os resultados observados são estatisticamente significativos ou apenas resultado do acaso. Isso nos ajuda a evitar a detecção de padrões falsos positivos e garante a confiabilidade das análises realizadas.

3.2.3 AR(1): Caminhadas aleatórias e Oscilações

Com o parâmetro p definido como 1, o modelo AR leva em consideração o valor anterior da série temporal multiplicado por um coeficiente α , e, em seguida, adiciona ruído branco. Quando o coeficiente é igual a 0, temos apenas ruído branco, resultando em uma série de tempo completamente aleatória, sem padrões previsíveis.

Quando o coeficiente é igual a 1, temos uma caminhada aleatória, onde cada valor da série é obtido somando-se o valor anterior a um termo de ruído branco. Nesse caso, os valores da série apresentam uma tendência linear, aumentando ou diminuindo ao longo do tempo sem retornar à média.

Se o coeficiente estiver na faixa $0 < \alpha < 1$, temos o fenômeno de reversão média. Isso significa que os valores da série tendem a oscilar em torno de uma média central e a regressar em direção a ela após se afastarem. Esse padrão indica uma tendência de retorno à média ao longo do tempo.

Os diferentes comportamentos da série temporal, determinados pelo coeficiente no modelo AR, têm implicações importantes na análise e previsão de dados. A compreensão desses padrões é fundamental para escolher o modelo adequado e interpretar corretamente os resultados obtidos.

3.2.4 AR(p): Termos de ordem superior

Aumentar ainda mais o parâmetro p no modelo AR significa considerar um número crescente de medições de tempo anteriores, cada uma multiplicada pelo seu próprio coeficiente. Isso permite levar em conta uma memória mais longa da série temporal e capturar padrões de dependência mais complexos ao longo do tempo.

No entanto, é importante ter em mente que aumentar excessivamente o valor de p pode levar a problemas de *overfitting*, onde o modelo se ajusta muito bem aos dados de treinamento, mas tem um desempenho ruim na previsão de novos dados. Portanto, é necessário encontrar um equilíbrio entre a complexidade do modelo e sua capacidade de generalização.

Além disso, é comum combinar o modelo AR com o modelo de média móvel (MA) para formar o modelo ARMA. O modelo MA considera os erros passados, ou seja, as diferenças entre os valores reais e as previsões anteriores, ajustadas por coeficientes. A combinação dos componentes AR e MA permite capturar tanto a dependência autorregressiva quanto a dependência na média móvel, proporcionando uma modelagem mais abrangente da série temporal.

Em suma, aumentar o parâmetro p no modelo AR pode melhorar a capacidade do modelo de capturar padrões complexos da série temporal, mas é necessário ter cuidado para evitar *overfitting*. A combinação com o modelo MA pode fornecer uma modelagem mais completa dos dados. A escolha adequada dos parâmetros depende da análise cuidadosa dos padrões presentes na série temporal e do equilíbrio entre a complexidade do modelo e sua capacidade de generalização.

3.2.5 Média Móvel

No modelo de média móvel (MA), o componente não é uma média móvel simples, mas sim uma combinação de termos de erro de previsão defasados. O parâmetro q no modelo MA representa o número de termos de erro de previsão que são levados em consideração na previsão.

De acordo com Trenberth (1984) este componente não é uma média de rolamento, mas sim os atrasos no ruído branco.

Em um modelo MA(1), por exemplo, a previsão é composta por um termo constante, o produto do termo de erro de previsão anterior por um multiplicador, e o termo de erro de previsão atual. Essa abordagem baseia-se em princípios estatísticos e de probabilidade, ajustando a previsão com base em termos anteriores de erro de previsão.

O modelo MA é uma alternativa ao modelo AR e é usado para capturar padrões de dependência na média móvel, ou seja, a influência de erros passados na previsão atual.

Ao combinar o modelo AR e o modelo MA, como no modelo ARMA, é possível obter uma modelagem mais abrangente que considera tanto a dependência autorregressiva quanto a dependência na média móvel.

Portanto, o modelo MA leva em conta os termos de erro de previsão defasados para ajustar a previsão atual, permitindo considerar a probabilidade e estatística na modelagem da série temporal.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (11)$$

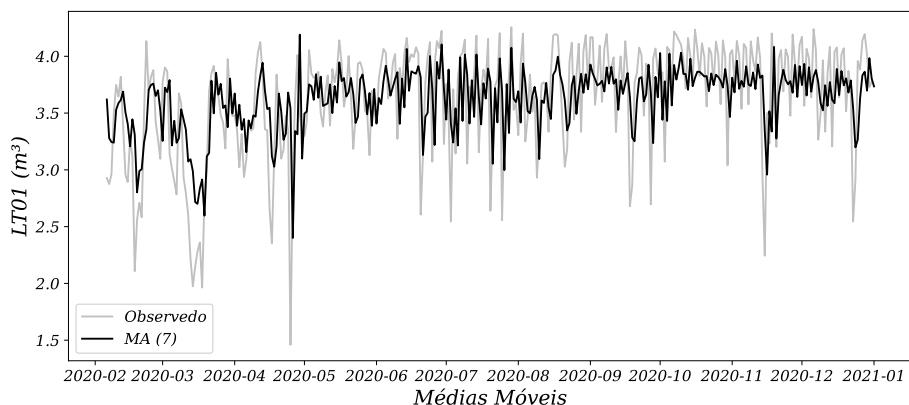
Na equação (11), em que ε_t representa o ruído branco, esse modelo é conhecido como um modelo de média móvel $MA(q)$, em que q é a ordem da média móvel. É importante ressaltar que não observamos diretamente os valores de ε_t , portanto, essa modelagem não se trata de uma regressão no sentido convencional.

Diferentemente de uma regressão comum em que temos variáveis explicativas observadas, no modelo $MA(q)$, estamos usando os termos de ruído branco defasados para estimar e prever os valores da série temporal. O objetivo é capturar a dependência dos termos de erro passados na previsão atual.

Esse modelo é útil para modelar séries temporais em que a média móvel tem um impacto significativo nas observações. Ao ajustar a série temporal com base nos termos de ruído branco defasados, podemos obter uma estimativa mais precisa dos valores futuros.

Embora o modelo $MA(q)$ seja diferente de uma regressão tradicional, ele é uma ferramenta estatística poderosa para modelar e prever séries temporais, levando em consideração a dependência entre os termos de erro passados.

Figura 19: Modelo MA(7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

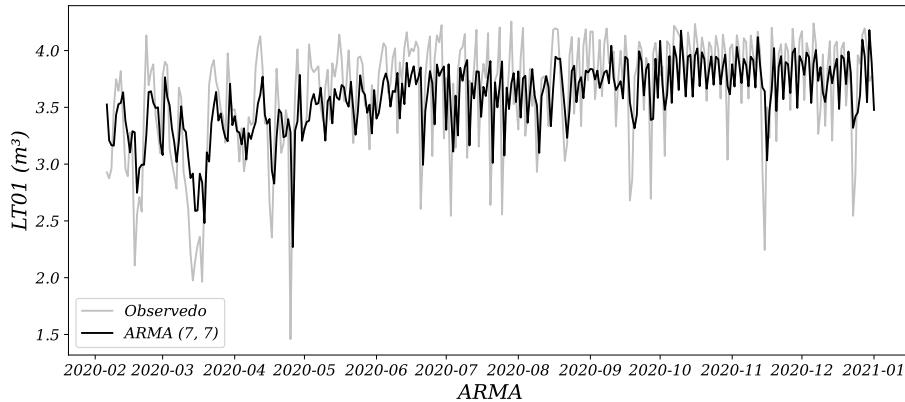
O modelo MA, quando comparado com o modelo AR de mesma ordem, facilita a previsão. Conforme ilustrado na Figura 19, a previsão gráfica se assemelha ao modelo apresentado na Figura 17, embora não seja comparável ao modelo exibido na Figura 18. É importante notar que esse modelo aparenta prever com precisão o período de tempo que foi considerado.

3.2.6 Modelos ARMA e ARIMA

A arquitetura ARMA é uma combinação dos modelos AR e MA, onde o modelo AR é adicionado ao modelo MA.

No modelo ARMA, é adicionada uma constante à soma dos termos autorregressivos multiplicados pelos seus coeficientes, juntamente com a soma dos termos de média móvel multiplicados pelos seus coeficientes, além do ruído branco. Essa estrutura é amplamente utilizada em diversos modelos de previsão em diferentes áreas.

Figura 20: ARMA (7,7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

A Figura 20 ilustra a combinação dos modelos AR e MA em um modelo ARMA. Essa abordagem pode levar a uma redução significativa no erro de previsão, como observado nos apêndices A e B, onde são apresentadas comparações com um maior número de passos de previsão.

3.2.7 ARIMA

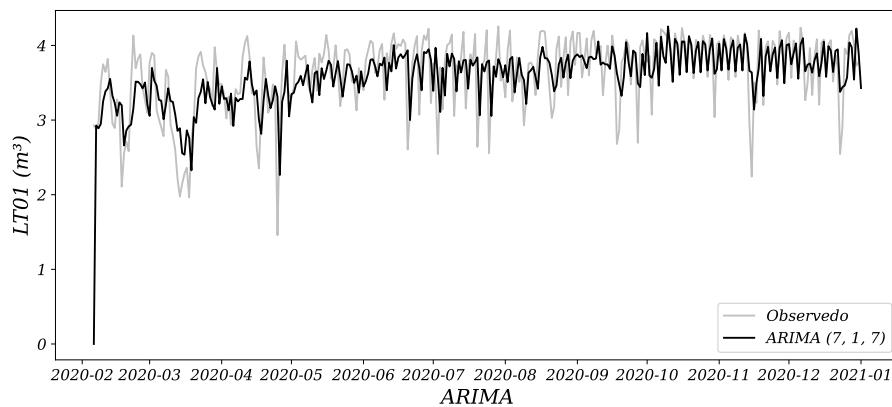
$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \dots + \omega_q \varepsilon_{t-q} + \varepsilon_t \quad (12)$$

Na equação (12), a variável Y_t representa a série temporal que foi diferenciada

(possivelmente mais de uma vez). Os "preditores" no lado direito da equação incluem os valores defasados de Y_t e os erros defasados. Esse tipo de modelo é conhecido como ARIMA (p, d, q) .

O modelo ARIMA é uma extensão do modelo ARMA que incorpora uma etapa adicional de pré-processamento chamada de diferenciação. Essa etapa é representada pela notação $I(d)$, em que d denota a ordem de diferenciação, ou seja, o número de transformações necessárias para tornar a série temporal estacionária. Portanto, um modelo ARIMA é simplesmente um modelo ARMA aplicado à série temporal diferenciada. Isso permite lidar com séries temporais que possuem tendências ou padrões não estacionários.

Figura 21: ARIMA $(7,1,7)$



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Ao analisar a Figura 21, não se nota uma diferença visual significativa em relação aos outros métodos apresentados anteriormente. O método ARX ainda parece ser superior aos demais com base na análise visual.

Embora os modelos ARIMA sejam eficazes, incorporar variáveis sazonais e exógenas ao modelo pode potencializar sua capacidade de previsão. No entanto, é importante destacar que o modelo ARIMA pressupõe que a série temporal seja estacionária. Quando lidamos com séries temporais não estacionárias, é necessário recorrer a outros modelos para a análise e previsão adequadas.

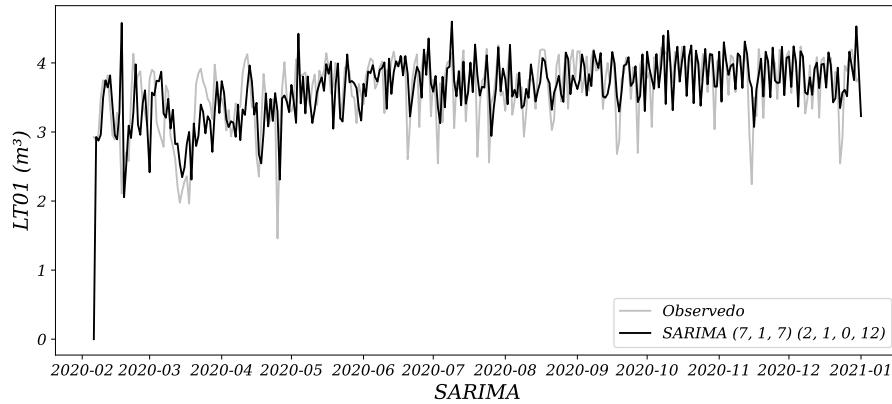
3.2.8 SARIMA

$$Y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^P \phi_n y_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (13)$$

O modelo proposto é uma extensão do modelo ARIMA, com a adição de compo-

nentes autorregressivos e de média móvel sazonal. Esses componentes extras são ajustados levando em consideração os padrões sazonais presentes nos dados, utilizando atrasos correspondentes à frequência sazonal (por exemplo, 12 para dados mensais). Essa abordagem permite capturar e modelar de forma mais precisa as variações sazonais e melhorar a qualidade das previsões em séries temporais com esse comportamento cíclico.

Figura 22: SARIMA $(7, 1, 7)(2, 1, 1)_{12}$



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 22, é possível observar que a previsão em vermelho está mais próxima dos valores observados em preto, mostrando que a inclusão do componente de sazonalidade melhora a qualidade da previsão. Os modelos SARIMA são capazes de lidar com dados que apresentam padrões sazonais, permitindo a diferenciação dos dados em termos de componentes sazonais e não sazonais. Uma abordagem útil para determinar os melhores parâmetros do modelo é utilizar uma estrutura de pesquisa automatizada de parâmetros, como o pmdarima, que auxilia na identificação dos parâmetros ideais para o modelo SARIMA. Isso pode contribuir para uma melhor compreensão e ajuste do modelo aos dados observados.

3.3 Modelos de série temporal multivariada

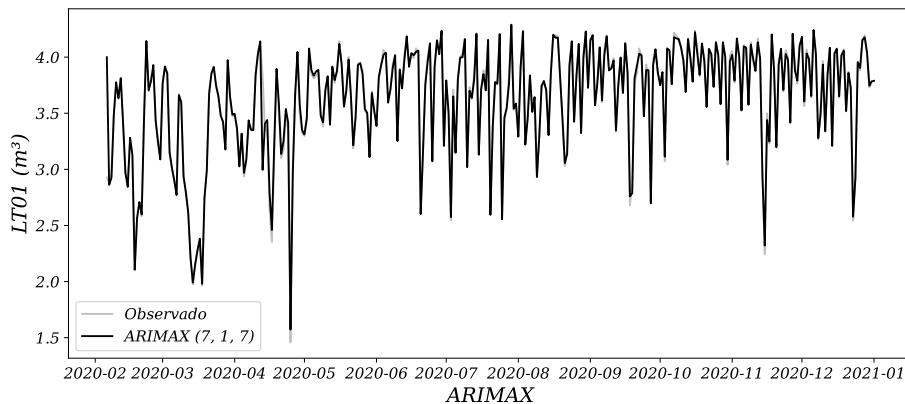
3.3.1 ARIMAX e SARIMAX

$$d_t = c + \sum_{n=1}^p \alpha_n d_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^r \beta_n x_{nt} + \sum_{n=1}^P \phi_n d_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (14)$$

Em (14), o modelo SARIMAX é apresentado. Nesse modelo, são consideradas variáveis exógenas, ou seja, são utilizados dados externos para a realização das previsões. É

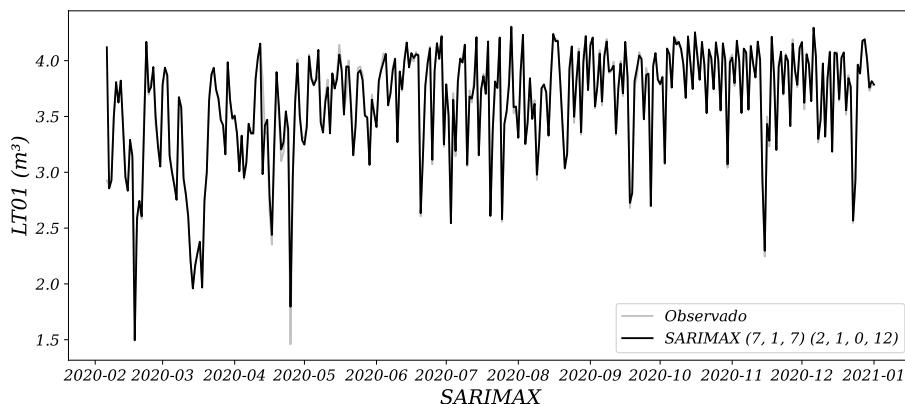
importante ressaltar que mesmo que essas variáveis exógenas sejam indiretamente modeladas no histórico de previsões do modelo, ao incluí-las diretamente, o modelo será capaz de responder de forma mais ágil aos efeitos dessas variáveis. Isso significa que a incorporação de informações externas possibilita uma resposta mais rápida e precisa do modelo em relação aos fatores externos, resultando em previsões mais atualizadas e acuradas.

Figura 23: ARIMAX (7, 1, 7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 24: SARIMAX (7, 1, 7)(2, 1, 1)₁₂



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Entre os modelos com variáveis exógenas, como mostrado nas Figuras 23 e 24, observa-se uma melhora significativa na qualidade das previsões em comparação com os modelos que não incluem variáveis exógenas. A adição dessas variáveis externas permite capturar melhor as influências e os padrões presentes nos dados, resultando em previsões mais completas e precisas. Essa inclusão de informações adicionais contribui para uma compreensão mais abrangente do comportamento da série temporal e possibilita uma melhor adaptação do modelo aos padrões observados.

3.4 Modelos de aprendizado de máquina supervisionados

Os modelos regressivos para séries temporais têm sido amplamente reconhecidos e utilizados na literatura atual, especialmente aqueles baseados em métodos de gradiente. Esses modelos, incluindo a regressão linear simples, têm se destacado como uma escolha popular em competições de séries temporais em todo o mundo.

Esses modelos são valorizados por sua capacidade de capturar relações complexas e não lineares nos dados, permitindo previsões mais precisas e eficientes. Sua popularidade reflete o reconhecimento da eficácia desses modelos em abordar uma ampla gama de problemas de previsão de séries temporais em diferentes áreas de estudo.

A abordagem regressiva, combinada com técnicas de otimização baseadas em gradiente, tem se mostrado particularmente eficaz na obtenção de resultados de alta qualidade. Esses modelos são capazes de aprender a partir dos dados históricos e ajustar seus parâmetros de forma iterativa, otimizando assim o desempenho da previsão.

Com a crescente disponibilidade de dados e avanços na área de aprendizado de máquina, espera-se que os modelos regressivos para séries temporais continuem a evoluir e desempenhar um papel importante na análise e previsão de dados temporais em diversas aplicações.

3.4.1 Regressão Linear (LR)

De acordo com o estudo realizado por Korstanje (2021), nos modelos de aprendizado de máquina supervisionados, é feita uma tentativa de identificar as relações existentes entre diferentes variáveis:

- Variável de destino: a variável que você tenta prever
- Variáveis explicativas: Variáveis que ajudam você a prever o alvo variável

Para realizar previsões, é importante que se compreenda quais tipos de variáveis explicativas podem ser utilizadas. Neste exemplo, a variável **Pressão de Sucção (PT01SU)** será considerada como a variável x , enquanto a variável **Nível do Reservatório (Câmara 1) LT01** será considerada como a variável y , com base na análise de correlação de Pearson ilustrada na Figura 25. O coeficiente de correlação indica a relação entre o eixo x e y , como expresso pela seguinte fórmula.

A fórmula do coeficiente de correlação de Pearson é dada por:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}} \quad (15)$$

Onde x_i e y_i representam os valores das variáveis X e Y , respectivamente. \bar{x} e \bar{y} são as médias dos valores x_i e y_i . O coeficiente de correlação de Pearson mede a força e a direção da relação linear entre as variáveis X e Y . Valores próximos a 1 indicam uma correlação positiva forte, valores próximos a -1 indicam uma correlação negativa forte, e valores próximos a 0 indicam uma ausência de correlação entre as variáveis.

Figura 25: Corelação de Pearson



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

A Figura 25 ilustra a correlação entre as variáveis no conjunto de dados em questão. Essa imagem representa graficamente a relação entre as variáveis e é usada para demonstrar a existência de uma correlação forte entre elas. Com base nessa análise, é possível responder à pergunta de pesquisa **Q 1**, pois a correlação entre as variáveis é significativa.

3.4.2 Definição do modelo

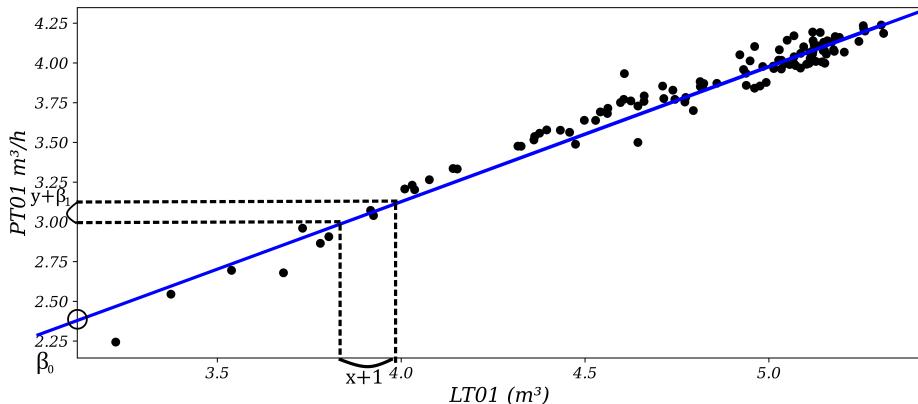
A regressão linear é definida da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (16)$$

Da Equação (16), temos as seguintes variáveis:

- Há p variáveis explicativas, denotadas por x .
- Existe uma variável alvo, denotada por y .
- O valor de y é calculado como uma constante β_0 , somada aos valores das variáveis x multiplicados por seus coeficientes β_1 a β_p .

Figura 26: Regressão linear LT01 vs PT01 correlação 98%



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

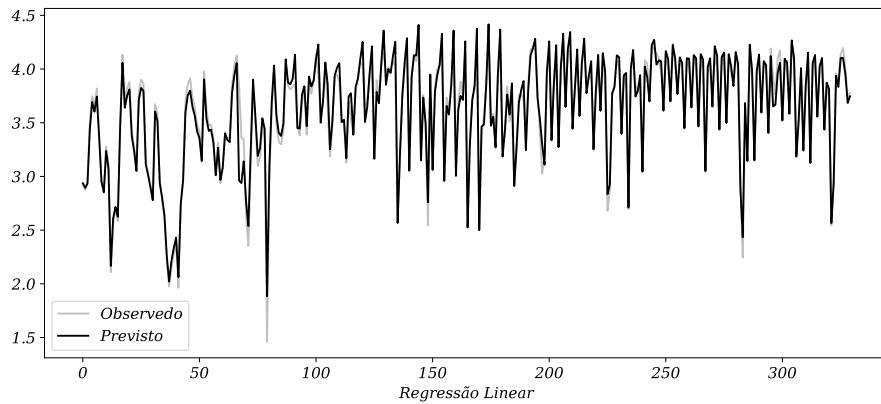
A Figura 26 fornece uma representação visual da interpretação dos coeficientes β_0 e β_1 . Ela ilustra que um aumento de 1 na variável x está associado a um aumento proporcional de β_1 na variável y . O valor de β_0 representa o valor de y quando x é igual a 0.

Para utilizar a regressão linear, é necessário estimar os coeficientes (betas) com base em um conjunto de dados de treinamento. Esses coeficientes podem ser estimados por meio da seguinte fórmula, expressa em notação matricial:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (17)$$

A fórmula mencionada, conhecida como **OLS** (método dos mínimos quadrados ordinários), é amplamente utilizada na regressão linear Korstanje (2021). Esse método é conhecido por ser rápido de ajustar, pois requer apenas cálculos matriciais para estimar os coeficientes β . No entanto, ele é mais adequado para processos lineares e pode ser menos adequado para modelos mais complexos que envolvam relações não-lineares. Portanto, é importante considerar suas limitações ao aplicar a regressão linear em contextos mais complexos.

Figura 27: Regressão linear (LR) um passo a frente

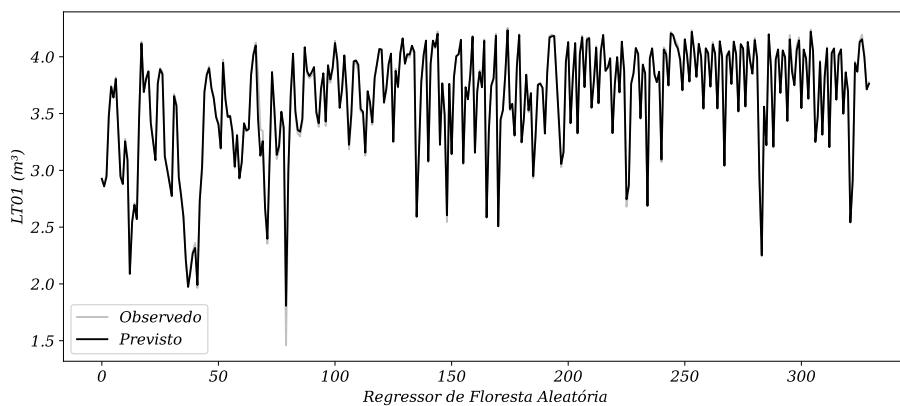


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

3.4.3 Floresta Aleatória

Pode-se observar que ter exatamente a mesma árvore de decisão repetidas vezes não adiciona valor significativo em comparação a usar essa mesma árvore de decisão apenas uma vez. Em modelos de conjunto, cada modelo individual deve ser ligeiramente diferente dos demais. Existem dois métodos amplamente reconhecidos para criar conjuntos: o ensacamento (*bagging*) e o reforço (*boosting*). A floresta aleatória utiliza o ensacamento para criar um conjunto de árvores de decisão, onde cada árvore é construída com uma amostra aleatória do conjunto de dados original. Isso garante que as árvores sejam distintas e diversificadas, contribuindo para a robustez e eficácia do modelo.

Figura 28: Regressão da Floresta Aleatória (RFR)

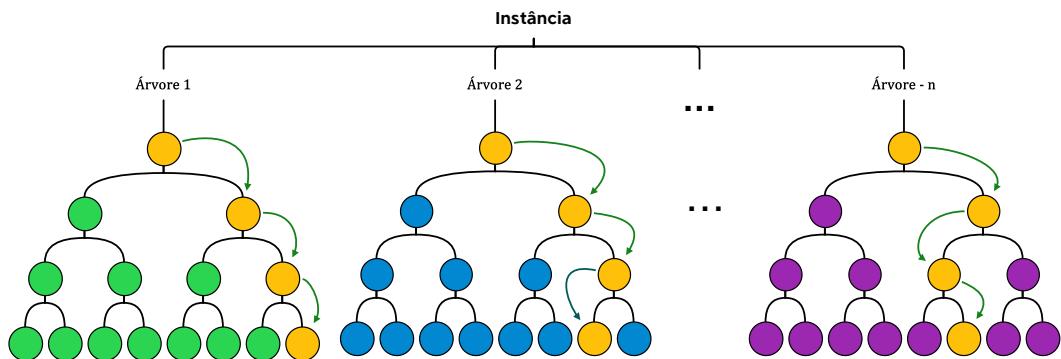


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Segundo Pelletier et al. (2016), cada árvore em um modelo de Floresta Aleatória de Regressão (RFR) é construída por meio de um algoritmo de aprendizado individual que

divide o conjunto de variáveis de entrada em subconjuntos, com base em um teste de valor de atributo, como o coeficiente de Gini. Ao contrário das árvores de decisão clássicas, as árvores de RFR são construídas sem poda e selecionam aleatoriamente um subconjunto de variáveis de entrada em cada nó. Atualmente, o número de variáveis utilizadas para dividir um nó em uma RFR (denotado por m) corresponde à raiz quadrada do número total de variáveis de entrada. Essa abordagem ajuda a aumentar a diversidade das árvores e aprimorar o desempenho do modelo.

Figura 29: Esquema da Floresta Aleatória



Fonte: Elaboração própria

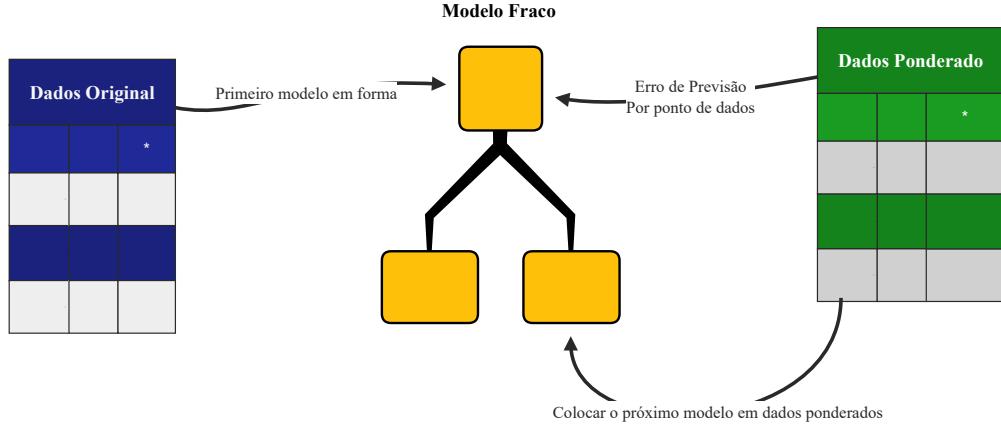
3.4.4 LightGBM e XGboost

O aumento de gradiente (do inglês *gradient boosting*) é um método que combina vários modelos de árvore de decisão para realizar previsões. Cada uma dessas árvores de decisão é única, pois a diversidade é um elemento importante nesse processo. A diversidade é alcançada através de um processo chamado boosting, que é uma abordagem iterativa. O boosting adiciona modelos fracos ao conjunto de forma inteligente, dando mais peso aos pontos de dados que ainda não foram bem previstos.

O processo de boosting melhora o conjunto ao focar nas partes dos dados que ainda não são compreendidas. A Figura 31 apresenta uma visão esquemática desse processo. À medida que novos modelos fracos são adicionados, todos os modelos fracos intermediários são mantidos. O modelo final é uma combinação de todos esses modelos fracos, resultando em um ensemble que oferece uma melhor capacidade de previsão do que um único modelo.

O boosting é apenas um dos métodos de ensemble utilizados em conjunto com o bagging. O bagging também é um método que utiliza múltiplos modelos de árvore de decisão, porém, em vez de adicionar os modelos de forma iterativa, cada modelo é treinado independentemente em subconjuntos aleatórios dos dados de treinamento. Ambos os métodos, boosting e bagging, têm como objetivo melhorar o desempenho do modelo combinando as previsões de múltiplos modelos individuais.

Figura 30: Impulsionando gradiente com XGBoost e LightGBM



Fonte: Adaptação de Korstanje (2021)

3.4.5 O Gradiente em Gradiente de Boosting (Reforço)

O processo iterativo utilizado no aumento de gradiente, como descrito por Korstanje (2021), recebe esse nome por um motivo. O termo “gradiente” refere-se a um campo vetorial de derivadas parciais que apontam na direção da inclinação mais acentuada. De forma simplificada, podemos pensar nos gradientes como as inclinações das estradas: quanto maior a inclinação, mais íngreme a colina. Para calcular os gradientes, são realizadas derivadas ou derivadas parciais de uma função.

No aumento de gradiente, ao adicionar árvores adicionais ao modelo, o objetivo é incorporar uma árvore que explique melhor a variação que ainda não foi explicada pelas árvores anteriores. Dessa forma, a nova árvore tem como objetivo ajustar-se aos erros ou resíduos deixados pelas árvores anteriores.

$$y - \hat{y} \quad (18)$$

A equação (18) pode ser reescrita como a derivada parcial negativa da função de perda em relação às previsões \hat{y} :

$$y - \hat{y} = -\frac{\partial L}{\partial \hat{y}} \quad (19)$$

Isso é definido como o objetivo da nova árvore a ser adicionada no modelo de aumento de gradiente, garantindo que ela explique a máxima quantidade de variação adicional no modelo geral. Essa é a razão pela qual o modelo é chamado de "aumento de gradiente" ("gradient boosting", em inglês). O processo utiliza o gradiente da função de perda para guiar a adição de novas árvores, buscando minimizar o erro e melhorar a capacidade do modelo em explicar a variação nos dados.

3.4.6 Algoritmos de boosting de gradiente

Existem muitos algoritmos que executam versões ligeiramente diferentes de aumento de gradiente. Quando o método de aumento de gradiente foi inventado, o algoritmo não tinha um desempenho tão bom, mas isso mudou com o advento do algoritmo AdaBoost: o primeiro algoritmo capaz de se adaptar a modelos fracos.

O algoritmo de aumento de gradiente é uma das ferramentas de aprendizado de máquina com melhor desempenho no mercado. Após o AdaBoost, uma longa lista de algoritmos de aumento levemente diferentes foi adicionada à literatura, incluindo XGBoost, LightGBM, LPBoost, BrownBoost, MadaBoost, LogitBoost e TotalBoost. Ainda há muitas contribuições para melhorar a teoria do aumento de gradiente. Nesta subseção, dois algoritmos são apresentados: XGBoost e LightGBM.

O **XGBoost** é um dos algoritmos de aprendizado de máquina mais utilizados. É uma forma rápida de obter bom desempenho. Devido à sua facilidade de uso e alto desempenho, é frequentemente o primeiro algoritmo escolhido por muitos profissionais de aprendizado de máquina.

O **LightGBM** é outro algoritmo de aumento de gradiente que é importante conhecer. Atualmente, é um pouco menos difundido que o XGBoost, mas está ganhando popularidade rapidamente. A vantagem esperada do LightGBM em relação ao XGBoost é um ganho de velocidade e uma utilização mais eficiente de memória.

Nesta subseção, você encontrará as implementações de ambos os algoritmos de aumento de gradiente.

3.4.7 A diferença entre XGBoost e LightGBM

Se alguém planeja utilizar os dois algoritmos de aumento de gradiente, é importante que essa pessoa compreenda suas diferenças, o que também proporciona uma visão das várias divergências que existem entre os modelos disponíveis no mercado.

Uma diferença fundamental reside na maneira como esses algoritmos identificam as melhores divisões entre os nós das árvores de decisão individuais. É crucial lembrar que uma divisão em uma árvore de decisão ocorre quando a árvore precisa encontrar a separação que mais melhora o desempenho do modelo.

A abordagem intuitiva e simples para encontrar a melhor divisão é iterar por todas as possibilidades e selecionar a melhor. No entanto, essa abordagem é computacionalmente custosa, e algoritmos mais recentes apresentam alternativas mais eficientes.

Uma alternativa proposta pelo XGBoost é a segmentação baseada em histograma. Nesse caso, em vez de iterar por todas as partições possíveis, o modelo constrói um histograma para cada variável e utiliza-os para encontrar a melhor divisão geral entre as

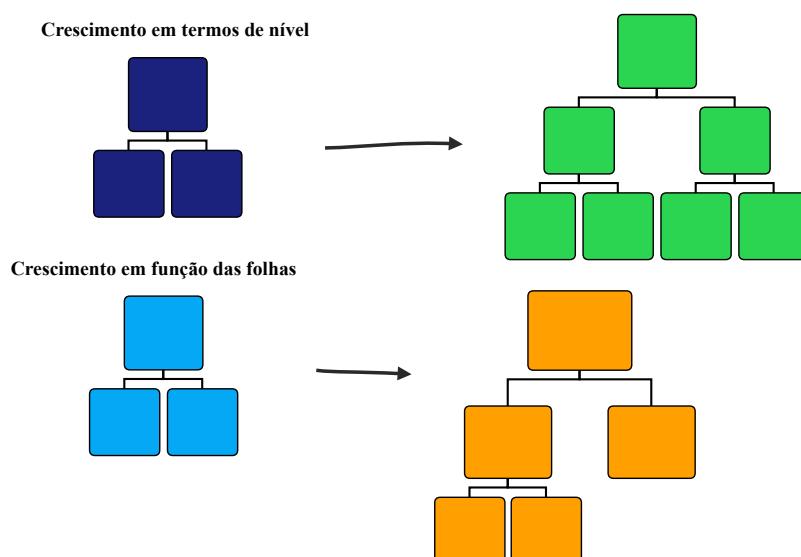
variáveis.

O LightGBM, desenvolvido pela Microsoft, adota uma abordagem mais eficiente para a definição das divisões. Essa abordagem é conhecida como amostragem unilateral baseada em gradiente (GOSS). O GOSS calcula o gradiente para cada ponto de dados e utiliza-o para filtrar os pontos de dados com gradientes baixos. Afinal, os pontos de dados com gradientes baixos já são bem compreendidos, enquanto aqueles com gradientes altos precisam ser melhor aprendidos.

O LightGBM também utiliza uma abordagem chamada Exclusive Feature Bundling (EFB), que acelera a seleção de muitas variáveis correlacionadas. Outra diferença é que o modelo LightGBM é adequado para o crescimento de folhas (leaf-wise growth), enquanto o XGBoost cultiva as árvores em níveis (level-wise growth). Essa diferença pode ser visualizada na Figura 31.

Essa diferença teoricamente favorece o LightGBM em termos de precisão, mas também apresenta um maior risco de overfitting (sobreajuste) quando há poucos dados disponíveis. Portanto, é importante que a pessoa considere essas distinções ao escolher entre os dois algoritmos de aumento de gradiente.

Figura 31: Crescimento em folha versus crescimento em nível



Fonte: Adaptação de Korstanje (2021)

Na Figura 31, é possível visualizar como cada modelo é ajustado durante o processo de crescimento de árvore em folhas e em níveis. Essa representação gráfica oferece uma compreensão visual das diferenças entre os dois métodos.

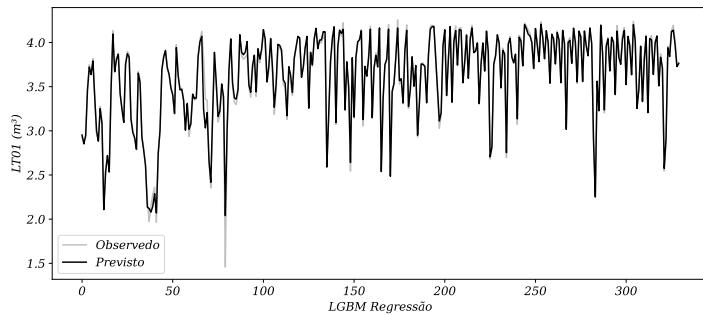
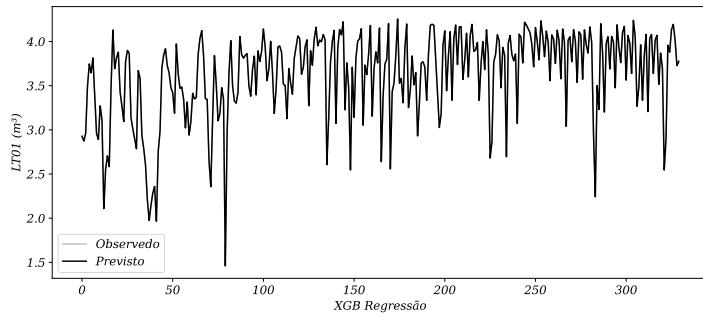
No crescimento de árvore em folhas, como no LightGBM, novas folhas são adicionadas à árvore de forma iterativa, visando maximizar a redução do erro de treinamento. Isso significa que as árvores são expandidas adicionando folhas, uma a uma, até que o

critério de parada seja alcançado.

Por outro lado, no crescimento em níveis, como no XGBoost, as árvores são expandidas em profundidade de forma simultânea em todos os níveis. Ou seja, em cada nível, todas as folhas são expandidas ao mesmo tempo, resultando em um crescimento mais uniforme da árvore.

Essa distinção no modo de crescimento das árvores pode afetar o comportamento e o desempenho do modelo. Portanto, compreender essa diferença é importante ao escolher entre esses algoritmos de aumento de gradiente.

Figura 32: XGBoost e LighGBM regressão



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 32 é um modelo baseado nos dados coletados da SANEPAR.

3.5 Estudo de Caso

Estudo de caso: Previsão de demanda de água usando análise de série temporal

Introdução: A previsão da demanda de água é uma preocupação fundamental para muitas organizações e autoridades responsáveis pelo abastecimento de água. A análise de séries temporais é uma abordagem comumente usada para prever padrões futuros com base em dados históricos. Neste estudo de caso, será explorado como a análise de séries temporais pode ser aplicada para prever a demanda de água ao longo do tempo.

3.5.1 Definição do problema

O primeiro passo é definir claramente o problema que queremos abordar. Por exemplo, vamos considerar a seguinte pergunta de pesquisa: "Como prever a demanda diária de água em uma determinada cidade para os próximos seis meses?"

Na subseção 1.2.1 estão as perguntas de pesquisa que serão abordadas no estudo de caso, da pergunta **Q 1 à Q 5**, com as ramificações da **Q 5**.

3.5.2 Coleta de dados

Na subseção 1.3, são apresentadas as variáveis contidas no conjunto de dados coletado no período de 2018 a 2020, durante uma grave falta de água que afetou a cidade. Devido a essa situação, foi implementado um rodízio de abastecimento de água para os residentes. Os dados foram coletados em intervalos de uma hora, levando em consideração cada variável, com ênfase na variável-alvo, denominada LT01, que representa o nível do reservatório.

O conjunto de dados possui um total de 26.306 linhas e 9 colunas. Durante a coleta dos dados, verificou-se que eles apresentam padrões sazonais, indicando variações recorrentes ao longo do tempo. Além disso, constatou-se que o consumo diário foi significativamente afetado no ano de 2020, diferindo dos anos anteriores, nos quais as mudanças não foram tão significativas.

3.5.3 Análise exploratória dos dados

Ao longo do trabalho realizado, pôde-se observar na subseção 2.1 que foi realizada uma análise gráfica do problema antes da aplicação de qualquer método. A detecção de anomalias mostrou-se desafiadora, porém não impossível de ser realizada. Essa detecção permitiu a análise da presença de sazonalidade nos dados. A decomposição STL foi utilizada para essa finalidade, conforme descrito na etapa **Etapa 3** e detalhado na subseção 4.1.3, onde são apresentadas as decomposições realizadas.

É fundamental lembrar que, durante a análise exploratória, os dados sofreram algumas alterações. Por exemplo, a média diária foi calculada em vez de ser considerada a nível horário, resultando em uma redução do conjunto de dados de 26.306 linhas para 1.096 linhas. A decomposição STL foi aplicada nos formatos aditivo e multiplicativo, e ambas as abordagens estão ilustradas nas Figuras 33 e 34, respectivamente.

Adicionalmente, na subseção 4.1.3, foi realizada a verificação da estacionariedade da série. O teste de Dickey-Fuller (DF) foi empregado para auxiliar na tomada de decisões, e os resultados demonstraram que a série em análise é estacionária, conforme evidenciado pelo teste DF.

3.5.4 Escolha do modelo

Como os dados apresentam sazonalidade, foram selecionados modelos simples de ARIMA, como AR, MA, ARMA, ARIMA e SARIMA. Esses modelos são univariados. Já os modelos com variável exógena, como ARX, ARIMAX e SARIMAX, são considerados multivariados. No contexto dos dados analisados, qualquer variável que possa interferir na variável preditora é considerada exógena. Para este caso específico, todas as outras variáveis foram incluídas como exógenas para melhorar a previsão.

Outros modelos utilizados são os modelos de aprendizado de máquina supervisados, como LR, RFR, LightGBM e XGBoost. Esses modelos são regressores baseados em árvores de decisão ou gradientes, especialmente os modelos XGBoost e LightGBM, que são amplamente reconhecidos como eficazes na previsão e tomada de decisões, conforme mencionado por Chen e Guestrin (2016) em seu estudo de benchmarking de frameworks de deep learning para tarefas de manutenção preditiva. Sánchez, Díaz e López (2020), em seu estudo comparativo de XGBoost, AdaBoost e CatBoost em algoritmos de aprendizado de máquina, também destacam o desempenho superior do XGBoost em várias métricas de avaliação.

3.5.5 Divisão dos dados

Para obter a divisão mais adequada dos dados, verificam-se a média e o desvio padrão de cada um desses conjuntos. O conjunto de dados é dividido em três partes: treinamento, validação e teste. Nessa divisão, utiliza-se inicialmente 70% dos dados para treinamento e validação, e os 30% restantes para teste. Em seguida, a porção de treinamento e validação é subdividida em 80% para treinamento e 20% para validação.

3.5.6 Ajuste do modelo

Nesta etapa, você aplicará o modelo selecionado aos dados de treinamento. Ajuste os parâmetros do modelo com o objetivo de minimizar os erros de previsão. Dependendo do modelo escolhido, você pode usar técnicas de otimização para encontrar os melhores parâmetros.

Ao ajustar o modelo para a base de dados, foi feita uma alteração na ordem do modelo sugerido pelo pmddarima. A escolha foi trocar o modelo SARIMAX(1,1,1)(2,1,0,12) para SARIMAX(7,1,7)(2,1,0,12). Essa decisão foi tomada com base na observação de um ajuste mais preciso aos dados, evidenciado pela redução nos resíduos e uma melhor captura das características da série temporal. Além disso, considerando o conhecimento do problema e as características específicas dos dados, foi identificado que padrões mais complexos requeriam ordens mais altas para serem adequadamente capturados. Dessa

forma, foi realizado um processo iterativo de experimentação e avaliação para determinar o modelo SARIMAX(7,1,7)(2,1,0,12) como o mais adequado para a base de dados em questão. É importante ressaltar que o desempenho do novo modelo será avaliado por meio de diagnósticos adicionais e análise dos resultados obtidos.

Os modelos XGBRegressor e LGBMRegressor foram ajustados usando as técnicas de GridSearchCV e BayesSearchCV. Essas abordagens permitiram encontrar as melhores combinações de hiperparâmetros para esses modelos, buscando maximizar o desempenho e a precisão das previsões. Por outro lado, os modelos LR (Regressão Linear) e RFR (Random Forest Regressor) não passaram por ajustes, pois não apresentaram melhorias significativas nos resultados após as etapas de GridSearchCV, BayesSearchCV e RandomizedSearchCV. Portanto, esses modelos mantiveram as configurações padrão, uma vez que as tentativas de otimização dos hiperparâmetros não resultaram em melhorias substanciais para eles.

- **GridSearchCV:** O GridSearchCV é uma técnica de busca exaustiva que é usada para ajustar os hiperparâmetros de um modelo de aprendizado de máquina. Ele realiza uma busca sistemática por todas as combinações possíveis de valores especificados para cada hiperparâmetro e avalia o desempenho do modelo para cada combinação. Essa abordagem avalia todas as opções disponíveis, mas pode ser computacionalmente intensiva. Ao final, fornece os melhores hiperparâmetros encontrados que otimizam a métrica de avaliação escolhida.
- **BayesSearchCV:** O BayesSearchCV é uma técnica de otimização de hiperparâmetros baseada em Bayesian optimization. Ele usa um processo de amostragem sequencial para encontrar a melhor combinação de hiperparâmetros de forma mais eficiente do que o GridSearchCV. O BayesSearchCV usa uma função de perda estimada e um modelo probabilístico para determinar quais configurações de hiperparâmetros são mais promissoras e, em seguida, realiza novas amostragens para refinar a busca. Essa abordagem permite uma exploração mais inteligente do espaço de hiperparâmetros e a descoberta de melhores configurações com menos iterações.
- **RandomizedSearchCV:** O RandomizedSearchCV é uma técnica de busca aleatória de hiperparâmetros. Ao contrário do GridSearchCV, que testa todas as combinações possíveis, o RandomizedSearchCV seleciona aleatoriamente um subconjunto do espaço de hiperparâmetros e avalia o modelo para cada combinação escolhida. Essa abordagem é útil quando o espaço de hiperparâmetros é grande e não é possível testar todas as combinações em tempo razoável. O RandomizedSearchCV permite uma exploração mais ampla do espaço de hiperparâmetros, embora com menor garantia de encontrar a melhor combinação.

3.5.7 Avaliação do modelo

Após ajustar o modelo, é hora de avaliar sua precisão. Use os dados de teste para comparar as previsões do modelo com os valores reais de demanda de água. Métricas como erro médio absoluto (MAE), erro quadrático médio (RMSE) e coeficiente de determinação (R^2) podem ser usadas para avaliar a qualidade das previsões.

3.5.8 Previsões futuras

Uma vez que você esteja satisfeito com a precisão do modelo, você pode usá-lo para fazer pre-

visões futuras. Aplique o modelo aos dados futuros para estimar a demanda de água nos próximos seis meses, por exemplo.

3.5.9 Monitoramento e ajuste contínuo

É importante lembrar que a demanda de água pode ser afetada por fatores externos imprevisíveis, como mudanças climáticas, eventos especiais ou mudanças no comportamento dos consumidores. Portanto, é necessário monitorar continuamente os resultados das previsões e ajustar o modelo conforme necessário.

3.5.10 Principais Conclusão

Este estudo de caso descreveu uma abordagem geral para prever a demanda de água usando análise de série temporal. No entanto, é importante adaptar essas etapas aos dados e às características específicas do seu caso. A análise de séries temporais pode ser uma ferramenta valiosa para previsão de demanda de água e pode ajudar a tomar decisões informadas para o gerenciamento do abastecimento de água.

4 Resultados

Neste capítulo, são fornecidos uma síntese e uma visão geral dos resultados obtidos até o momento. É apresentado um resumo sucinto das principais realizações e descobertas que foram alcançadas até agora.

4.1 Planejamento do Problema

Assim como apresentado na seção 1.4.1, os passos da dissertação delinearam o processo pelo qual cada modelo foi construído e os métodos utilizados para responder às questões de pesquisa abordadas na seção 1.2.1. Esses passos proporcionaram uma cronologia lógica das etapas realizadas ao longo do tempo com os dados da SANEPAR, ilustrando o progresso e os resultados alcançados até o momento.

4.1.1 Análise Exploratória dos dados (EDA)

A partir do passo **Etapa 1**, foi realizado o EDA (Exploratory Data Analysis) para processar os dados obtidos até o momento. O EDA permite responder às questões de pesquisa levantadas. Conforme mencionado por Yu (2016), na era dos grandes dados, é desafiador descobrir as regras, modelos analíticos e hipóteses por trás dos volumes massivos de dados caóticos, não estruturados e multimídia coletados por meio de vários canais. A análise exploratória de dados foi promovida por John Tukey como uma abordagem para explorar os dados, resumir suas principais características e formular hipóteses que possam direcionar a coleta adicional de dados e experimentos. No contexto de grandes análises de dados, várias técnicas de EDA têm sido adotadas.

Ao analisar a pergunta **Q 1**, que relaciona a demanda com a variável prevista e a pressão para a variável PT01, pode-se observar na Figura 25 que ambas as variáveis apresentam uma correlação quase perfeita, com um coeficiente de correlação de Pearson (r) igual a 1. Portanto, para responder a essa pergunta, basta observar a correlação de Pearson na Figura 25.

Para responder à pergunta **Q 2**, é criada uma tabela para fornecer uma resposta mais completa.

Tabela 4: Descrição estatística dos dados com o filtro aplicado das 18h às 21h

18 a 21h	B1	B2	B3	LT01	FT01	FT02	FT03	PT01	PT02
Contagem	366	366	366	366	366	366	366	366	366
Média	43,87	22,26	8,70	3,34	164,83	133,08	102,01	4,23	17,29
STD	23,22	18,47	17,81	0,69	114,60	67,99	47,55	0,81	8,59
Min	0	0	0	0,99	0,07	0	0	1,88	0
25%	37,93	0	0	2,87	64,31	131,06	107,92	3,69	16,77
50%	57,99	30,92	0	3,41	201,37	146,17	121,40	4,22	22,46
75%	57,99	37,25	0	3,86	268,61	158,71	127,07	4,85	22,52
Max	59,99	57,33	53,74	4,40	379,20	285,56	170,56	5,66	24,23

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Tabela 4, o desvio padrão é representado pela sigla STD, que corresponde à expressão em inglês "standard deviation". Além disso, em resposta à pergunta **Q 2**, é importante mencionar que, assim como em qualquer empresa de tratamento de água, é utilizado um mecanismo de acionamento automático chamado "trava de segurança" para evitar que o nível do tanque chegue a zero e haja falta de água nos locais abastecidos por esse tanque. O nível mínimo que o tanque pode alcançar é de $1.459m^3$ (equivalente a 1459 litros). As bombas são ativadas em sua potência máxima para evitar que sejam acionadas quando o nível do tanque estiver fora da faixa de $[3.843, 4.256] m^3$. No entanto, a bomba 1 ainda estaria operando para completar o nível do tanque caso ele esteja dentro dessa faixa.

Em situações de demanda de pico, uma abordagem ideal, embora não necessariamente a mais econômica, seria ter um tanque de reserva adicional e instalar uma tubulação que os conecte. Durante o dia, ambos os tanques seriam abastecidos e, à noite, por meio da ação da gravidade, eles manteriam o mesmo nível até que o consumo atinja um ponto em que as bombas sejam acionadas. Essa estratégia permite um abastecimento contínuo e eficiente de água.

Na pergunta **Q 3**, observa-se que o tanque tem uma capacidade máxima de $4,256m^3$, o que equivale a 4.256 litros. Para atender a essa demanda e manter o tanque quase cheio ou sempre cheio, é necessário que o fluxo de entrada esteja na faixa de $[238, 302] m^3/h$, o fluxo de gravidade esteja entre $[126, 182] m^3/h$, o fluxo de retorno esteja entre $[110, 144] m^3/h$, a pressão de sucção esteja entre $[1.92, 4.24] mca$ e a pressão de retorno esteja entre $[21, 24] mca$.

Para responder à pergunta **Q 4**, o ponto de equilíbrio, onde as bombas não precisam ser acionadas, ocorre quando o fluxo de FT01 é de $211 m^3/h$, FT02 é de $114 m^3/h$, FT03

é de $100\ m^3/h$ e o nível do tanque está em $3.545\ m^3$. No que diz respeito à pergunta **Q 5a.**, o nível do tanque deve ser de $4,00\ m^3$ para evitar o funcionamento das bombas durante as horas de pico.

4.1.2 Múltiplas entradas e saída única (MISO)

Na etapa **Etapa 2**, foram explorados modelos MISO (do inglês *Multiple Inputs, Single Output*) na dissertação. Os modelos ARIMA e suas derivações foram amplamente estudados, juntamente com modelos regressivos que envolvem múltiplas variáveis de entrada e uma variável de saída, neste caso, a LT01. As demais variáveis foram utilizadas como suporte para melhorar os modelos do tipo ARIMAX ou modelos com variáveis exógenas. Quando aplicados sem o uso de variáveis exógenas, os modelos ARIMA apresentam apenas uma entrada, semelhante ao modelo de regressão linear (LR). No entanto, ao incluir variáveis exógenas, os modelos se tornam MISO, permitindo uma modelagem mais abrangente e considerando a interação de várias variáveis para prever a variável de interesse.

4.1.3 Decomposição STL

A decomposição sazonal e de tendência utilizando o procedimento de Loess (STL) é uma técnica amplamente utilizada para decompor séries temporais em seus componentes sazonais, de tendência e restantes. De acordo com Theodosiou (2011), o método STL realiza a decomposição aditiva dos dados por meio de uma sequência de aplicações do Loess mais suave, onde regressões polinomiais ponderadas localmente são aplicadas em cada ponto do conjunto de dados, tendo como variáveis explicativas os valores mais próximos do ponto cuja resposta está sendo estimada.

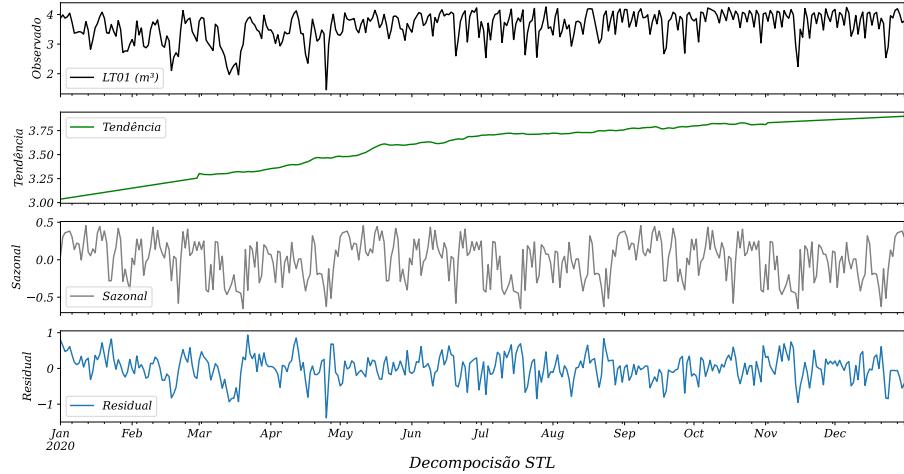
A decomposição STL é especialmente útil para identificar e isolar padrões sazonais e de tendência presentes nas séries temporais. Ela permite a separação dos componentes sazonais, que ocorrem em intervalos regulares ao longo do tempo, da componente de tendência, que indica a direção geral dos dados ao longo do tempo. A decomposição também resulta em uma componente restante, que representa a variação não explicada pelos componentes sazonais e de tendência.

Ao aplicar a decomposição STL, a série temporal pode ser expressa como a soma dos componentes sazonais, de tendência e restantes. Essa técnica é útil para análise e modelagem de séries temporais, pois proporciona uma compreensão mais clara dos padrões de variação presentes nos dados.

A decomposição STL é formalmente definida como:

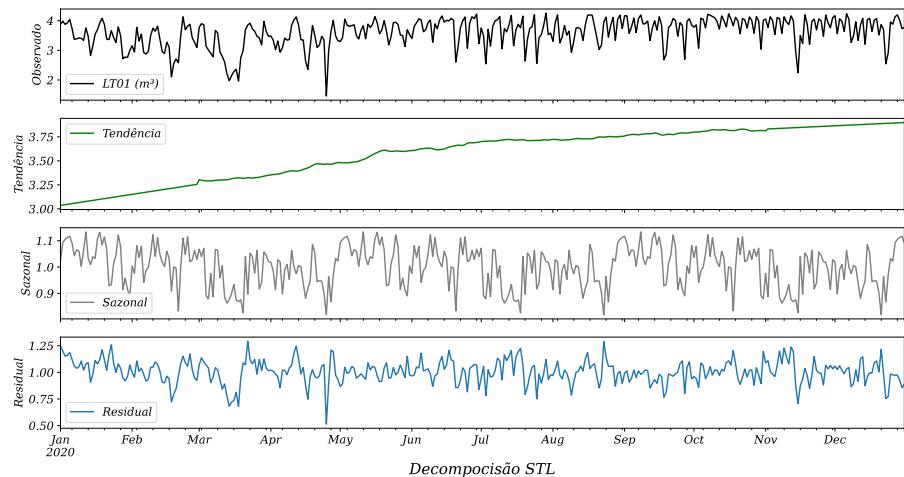
$$y_t = f(S_t, T_t, R_t) = \begin{cases} y_t = S_t + T_t + R_t & \text{aditivo} \\ y_t = S_t T_t R_t & \text{multiplicativo} \end{cases} \quad (20)$$

Figura 33: Decomposição STL aditiva dos dados coletados



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 34: Decomposição STL multiplicativa dos dados coletados



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na resposta à pergunta **Q 5b.**, as Figuras 33 e 34 fornecem informações sobre a presença de tendência, sazonalidade e resíduos na série temporal.

Através da decomposição, é possível analisar se a série apresenta tendência, sazonalidade e resíduos. Ao observar as Figuras 33 e 34, é evidente que os dados exibem

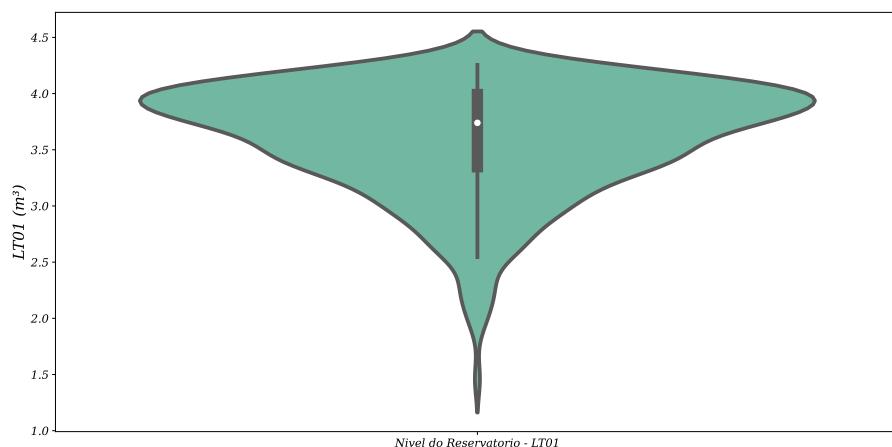
ambos os padrões. Isso indica que a série é estacionária, como confirmado pelo seguinte teste.

Teste de Dickey-Fuller (DF) Aumentado:

- Estatística de teste ADF: -4.248
- Valor de p: 0.001
- Atrasos utilizados: 21.000
- Observações: 1074.000
- Valor crítico (1%): -3.436
- Valor crítico (5%): -2.864
- Valor crítico (10%): -2.568

Com base na forte evidência contra a hipótese nula, podemos rejeitar a hipótese nula. Isso indica que os dados não possuem raiz unitária e são estacionários em Q 5c.. Identificar as horas de pico entre 18h e 21h não é uma tarefa fácil. No entanto, ao observar a Figura 35, podemos notar um aumento na demanda durante essas horas durante o ano de 2020.

Figura 35: Violino no nível do reservatório

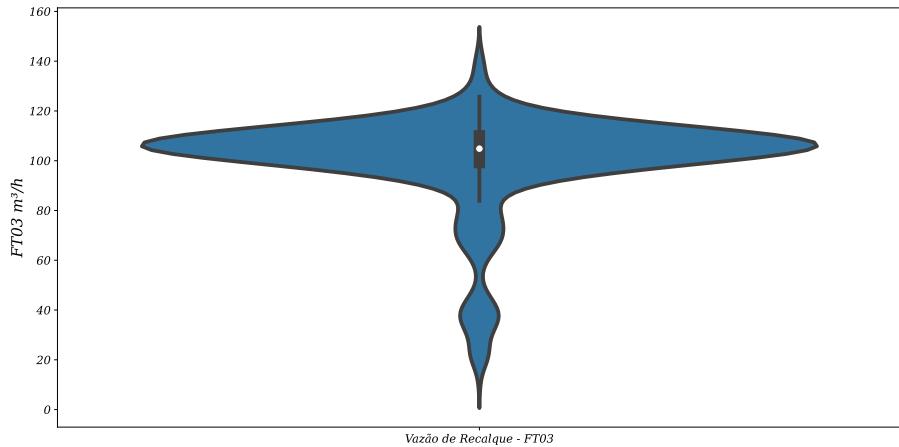


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Conforme mencionado na subseção 1.1.1, as anomalias climáticas ocorridas em 2020, especialmente a falta de chuvas, tiveram um impacto significativo nos resultados. Isso contribuiu para as mudanças observadas na demanda de água ao longo desse período.

Com relação à pergunta **Q 5d.**, durante as horas de pico, é necessário que o nível do tanque esteja dentro da faixa de $[3.545, 4.256] m^3$ para evitar o acionamento das bombas. Manter o nível do tanque dentro dessa faixa permitirá que o sistema opere de forma eficiente, atendendo à demanda sem a necessidade de acionar as bombas.

Figura 36: Violino da vazão de recalque



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Para responder à pergunta **Q 5e.**, a Figura 36 ilustra como a vazão pode ser afetada pelo nível do tanque. É interessante observar que a vazão de recalque tem um impacto mais significativo no nível do tanque em comparação com as outras vazões. Isso ocorre porque a vazão de recalque está associada à injeção de água diretamente no tanque por meio da bomba localizada próxima à base do tanque. Por outro lado, as demais vazões apresentam alguns valores ausentes, o que limita sua influência na análise geral.

De acordo com o Reisen et al. (2017), o teste DF tem as seguintes equações

$$z_t = y_t + \theta \beta_t, \quad t = 1, \dots, T, \quad (21)$$

$$\hat{\rho}_{\text{DF}} - 1 = \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (22)$$

De (22) onde $\Delta z_t = z_t - z_{t-1}$. Sob a hipótese nula (H_0) : “ $\rho = 1$ ”, as estatísticas do teste DF e suas distribuições limitantes são dadas da seguinte forma:

$$T(\hat{\rho}_{\text{DF}} - 1) = T \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (23)$$

e

$$\hat{\tau}_{\text{DF}} = \frac{\hat{\rho}_{\text{DF}} - 1}{\hat{\sigma}_{\text{DF}} \left(\sum_{t=1}^T z_{t-1}^2 \right)^{-1/2}} \quad (24)$$

De (24) onde $\hat{\sigma}_{\text{DF}}^2 = T^{-1} \sum_{t=1}^T (\Delta z_t - (\hat{\rho}_{\text{DF}} - 1) z_{t-1})^2$.

Suponha que $(z_t)_{1 \leq t \leq T}$ são dadas por (21), então quando $\rho = 1$,

$$T(\hat{\rho}_{\text{DF}} - 1) \xrightarrow{d} \frac{W(1)^2 - 1}{2 \int_0^1 W(r)^2 dr} - \left(\frac{\theta}{\sigma} \right)^2 \frac{\pi}{\int_0^1 W(r)^2 dr}, \text{ como } T \rightarrow \infty \quad (25)$$

$$\hat{\tau}_{\text{DF}} \xrightarrow{d} [1 + 2(\theta/\sigma)^2 \pi]^{-1/2} \left\{ \frac{W(1)^2 - 1}{2 \left(\int_0^1 W(r)^2 dr \right)^{1/2}} - \frac{(\theta/\sigma)^2 \pi}{\left(\int_0^1 W(r)^2 dr \right)^{1/2}} \right\} \quad (26)$$

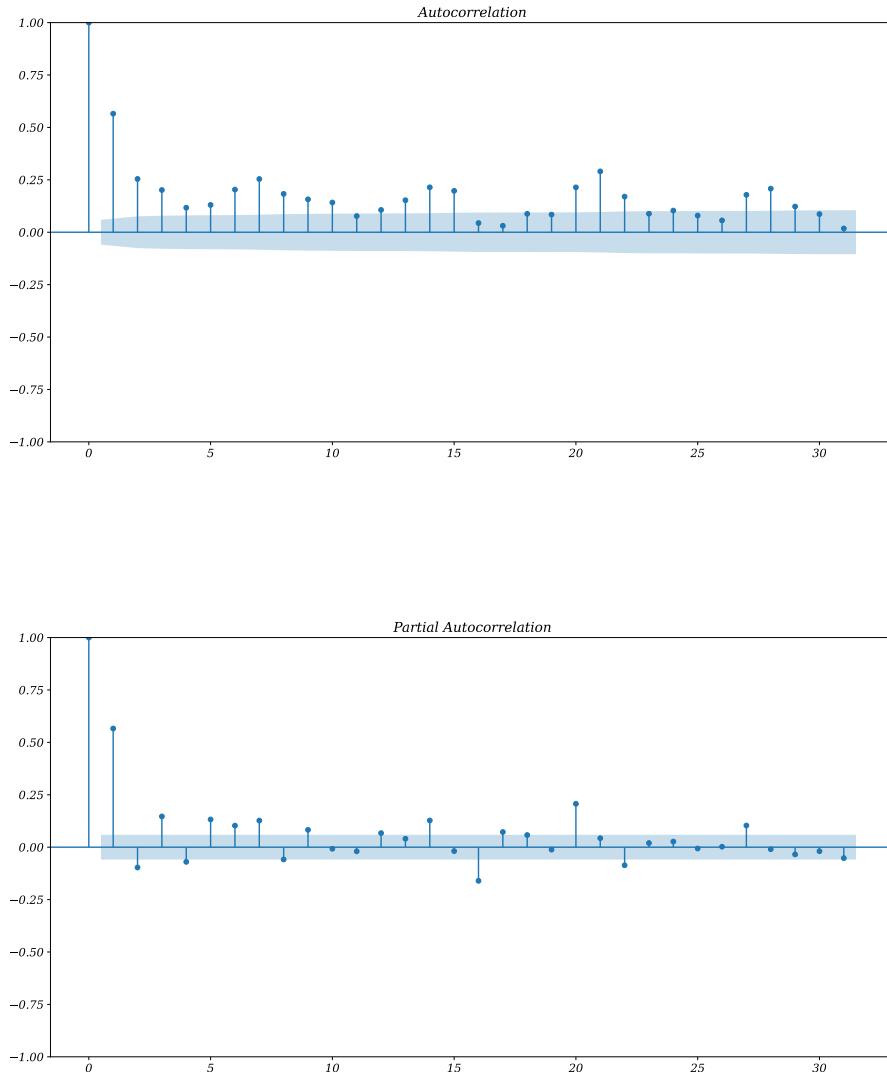
$$\text{como } T \rightarrow \infty \quad (27)$$

A partir de (27), onde \xrightarrow{d} denota convergência na distribuição e onde $\{W(r), r \in [0, 1]\}$ denota o movimento Browniano padrão.

O ACF (do inglês *Auto-Correlation Function*) é uma medida estatística utilizada para identificar a presença de correlação serial em uma série temporal. Ele calcula a autocorrelação entre os valores da série em diferentes defasagens, ou seja, a correlação entre os valores atuais e os valores passados da série.

O ACF é útil para analisar a dependência temporal dos dados e identificar padrões de sazonalidade, tendência ou outros efeitos temporais. Através do ACF, é possível avaliar se a série exibe autocorrelação significativa em defasagens específicas, o que pode indicar a presença de não estacionariedade ou estrutura temporal que precisa ser considerada na análise ou modelagem da série temporal.

Figura 37: Autocorrelação e Autocorrelação parcial



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

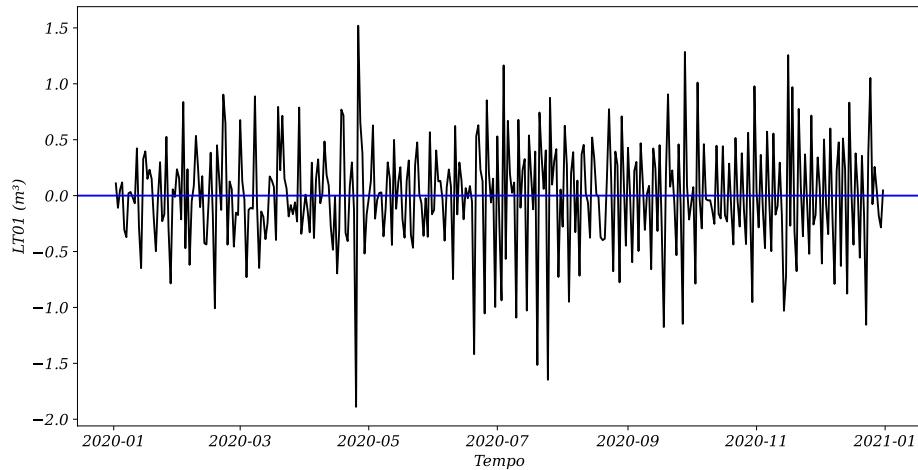
Na Figura 37, é possível observar a diferença entre a autocorrelação e a autocorrelação parcial (PACF). A autocorrelação mede a correlação entre os valores da série temporal em diferentes defasagens, levando em consideração tanto a correlação direta quanto a correlação indireta. Por outro lado, a autocorrelação parcial mede apenas a correlação direta entre os valores, eliminando a influência das defasagens intermediárias.

O intervalo de confiança padrão de 95% é representado pela marca azul na Figura. As observações que estão fora desse intervalo são consideradas estatisticamente correlacionadas, indicando a presença de padrões ou estrutura na série temporal.

A correlação visualizada na Figura 37 é fundamental para a interpretação do teste DF. Em uma série de ruído branco, os valores são completamente aleatórios e não apresentam correlação significativa. Portanto, quando há correlação presente na série, isso

indica a existência de padrões ou dependências entre os valores, o que pode ser explorado para a modelagem e previsão da série temporal.

Figura 38: Ruído branco



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 38, é possível observar uma série temporal que pode ser caracterizada como ruído branco. Uma série temporal é considerada ruído branco se suas variáveis forem independentes e distribuídas de forma idêntica, com média zero. Isso implica que todas as variáveis possuem a mesma variância (σ^2) e que cada valor não possui correlação com os demais valores da série.

Além disso, é importante destacar o comprimento dos zeros na variável prevista, o que conclui a etapa **Etapa 3**.

4.1.4 Separação dos dados

Na etapa **Etapa 4**, os dados foram divididos em conjuntos de treinamento, teste e validação. Essa prática é comum entre profissionais de aprendizado de máquina, pois permite avaliar o desempenho do modelo em conjuntos de dados diferentes.

Em relação ao processamento de modelos de aprendizado profundo, é importante mencionar as inovações trazidas pela empresa Nvidia ao longo dos anos, especialmente no campo do processamento de imagens. O lançamento da placa de vídeo GeForce RTX 4090 tem sido bastante aguardado tanto por gamers quanto por profissionais que lidam com aprendizado de máquina.

No contexto do estudo, foram utilizados dois computadores para realizar os cálculos dos modelos. Um deles é equipado com um processador Intel Core i5-3330 e o outro é um notebook com um processador Intel Core i7-5500. Ambos os processadores possuem

4 threads, sendo que o notebook possui 2 núcleos físicos e o i5 possui 4 núcleos físicos. Cada processador tem suas especificações e desempenho adequados a diferentes necessidades. Vale ressaltar que não é obrigatório utilizar as últimas gerações de processadores para realizar esses processamentos, e sim compreender e aplicar corretamente os recursos disponíveis.

Quanto à divisão dos dados, foi adotada uma estratégia básica em que 70% dos dados foram destinados ao conjunto de treinamento e os 30% restantes foram reservados para o conjunto de teste. Dentro dos 70% de treinamento, foi realizada uma subdivisão em que 80% desses dados foram usados novamente para treinamento e os 20% restantes foram utilizados para validação. Essa abordagem foi implementada em linguagem de programação para facilitar o processo e evitar a necessidade de recalculá-la a cada modificação do modelo.

4.1.5 Estratégia de Previsão

Na Etapa **Etapa 5**, discute-se a previsão dos dados em uma janela de horizonte de previsão estendida, abrangendo diferentes períodos de tempo, como um dia, uma semana, duas semanas e um mês. Essa estratégia de previsão recorrente permite a comparação entre modelos de regressão e modelos ARIMA em diferentes horizontes temporais.

Essa abordagem é vantajosa, pois cada modelo possui suas próprias características e desempenho ao lidar com previsões de curto prazo, como um dia, e previsões de prazo mais longo, como um mês. Ao utilizar uma janela de previsão mais ampla, é possível observar e avaliar melhor as diferenças entre os modelos e analisar seu desempenho em horizontes de tempo variados.

4.1.6 Horizonte

Na etapa **Etapa 6**, o horizonte de previsão foi personalizado com base no método recursivo de previsão de série temporal e na previsão do nível do tanque LT01. Foram selecionados os seguintes passos para a previsão à frente: um dia, uma semana, duas semanas e um mês. Essa escolha do horizonte de previsão foi feita levando em consideração a estratégia recursiva e os objetivos específicos do estudo. Foi identificado que essa janela de tempo proporcionaria uma análise mais adequada e comparável entre os modelos utilizados.

4.1.7 Modelos de previsão e métricas de desempenho

A partir da etapa **Etapa 7**, foram utilizadas três métricas amplamente empregadas na literatura para a previsão e comparação de modelos ARIMA e modelos de regressão.

Essas métricas foram detalhadas na seção 3.1.

Ao analisar os modelos desenvolvidos, observou-se que o modelo de regressão linear (LR) obteve o melhor desempenho tanto na previsão de curto prazo, considerando uma modelagem de 24 horas, quanto nas horas de pico entre 18h e 21h. Os modelos MA, AR, SARIMA, ARIMA, SARIMAX, ARIMAX, ARX, LGBMRegressor, XGBRegressor e RFR também apresentaram um desempenho satisfatório, seguindo uma ordem de melhor para pior.

Para previsões de longo prazo, como no caso dos 30 dias, foram avaliados os modelos ARMA, AR, MA, ARIMA, ARIMAX, ARX, SARIMA, SARIMAX, XGBRegressor, RFR, LGBMRegressor e LR, novamente seguindo a ordem de melhor desempenho. No entanto, ao analisar os resultados graficamente nos apêndices, percebe-se que os modelos que incorporam variáveis exógenas parecem ter uma capacidade de previsão superior em relação aos demais modelos. Essa tendência pode ser observada nas Figuras de 41 a 52 e nas Tabelas de 6 a 21.

4.1.8 Teste de Significância

Na etapa **Etapa 8**, foi utilizado o teste de Friedman e Nemenyi para comparar as classificações médias entre os classificadores. O teste de Nemenyi é um teste de comparação múltipla utilizado após a aplicação de testes não paramétricos com três ou mais fatores.

Tabela 5: Teste Nemenyi

Nemenyi	0	1	2	3	4	5	6	7	8
0	1,000	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
1	0,001	1,000	0,001	0,001	0,001	0,001	0,001	0,001	0,157
2	0,001	0,001	1,000	0,847	0,001	0,001	0,001	0,001	0,001
3	0,001	0,001	0,847	1,000	0,001	0,001	0,001	0,001	0,001
4	0,001	0,001	0,001	0,001	1,000	0,001	0,001	0,001	0,001
5	0,001	0,001	0,001	0,001	0,001	1,000	0,001	0,001	0,001
6	0,001	0,001	0,001	0,001	0,001	0,001	1,000	0,001	0,001
7	0,001	0,001	0,001	0,001	0,001	0,001	0,001	1,000	0,001
8	0,001	0,157	0,001	0,001	0,001	0,001	0,001	0,001	1,000

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Para calcular a estatística de teste F_r de Friedman, inicialmente cria-se uma tabela com os dados, onde cada linha representa uma amostra e cada coluna representa uma condição de teste. Em seguida, as amostras são ordenadas ao longo das condições, da

melhor situação para a pior. Se não houver empates, a estatística de teste F_r é calculada utilizando a seguinte fórmula:

$$F_r = \left(\frac{12}{nk(k+1)} \sum_{i=1}^k R_i^2 \right) - 3n(k+1) \quad (28)$$

Nessa fórmula, n é o número de linhas (ou amostras), k é o número de colunas (ou condições) e R_i é a soma das fileiras da coluna (ou condição) i .

Além disso, o valor crítico CD (Critical Difference) é utilizado para determinar se dois classificadores são significativamente diferentes um do outro. O CD é calculado usando a fórmula que mencionei anteriormente:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (29)$$

Na fórmula do CD, q_α é o valor crítico obtido da tabela de teste de Nemenyi, k é o número de classificadores e N é o número total de amostras.

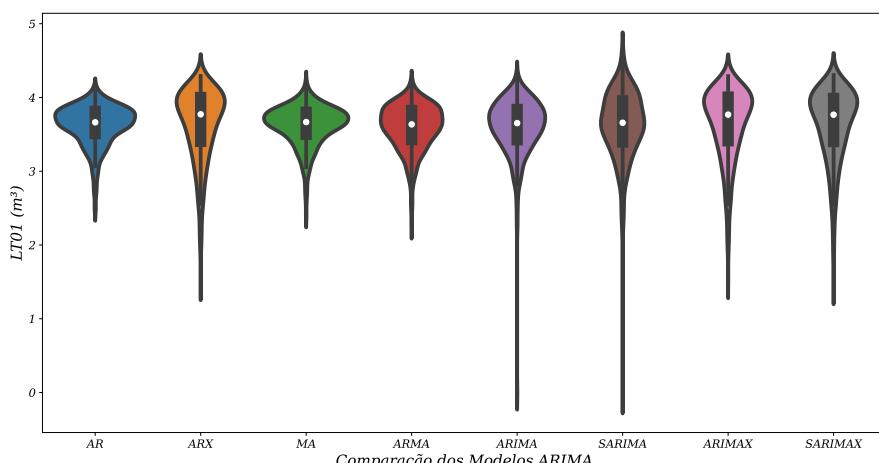
De acordo com essa equação, os resultados da pesquisa foram os seguintes:

statistic = 8015.611, *p-value* = 0.0 com um total de 26.306 linhas por 9 colunas.

4.1.9 Comparação dos modelos

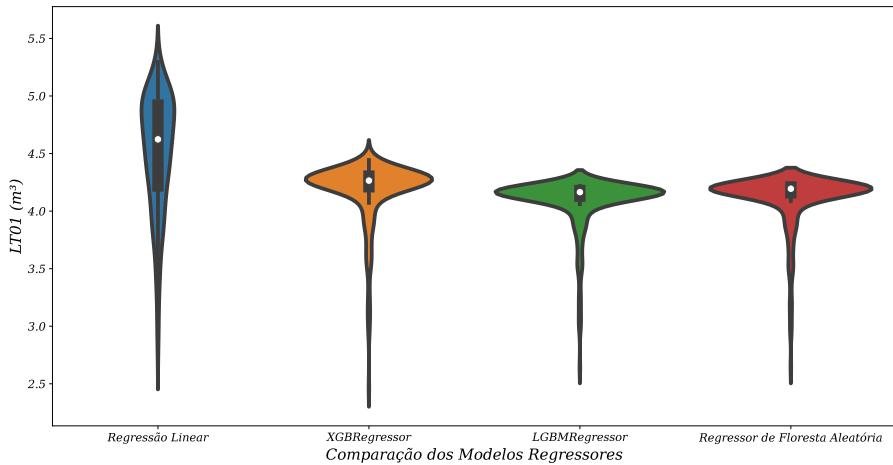
Com o objetivo de obter uma análise mais aprofundada do desempenho de cada modelo, foi realizada uma comparação por meio de um gráfico de violino. Dessa forma, pôde-se observar qual dos modelos apresentava o melhor desempenho.

Figura 39: Comparação dos modelos ARIMAS



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 40: Comparação de modelos de regressão



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Ao comparar os modelos apresentados nas Figuras 39 e 40, é possível observar quais são os modelos que se destacam, levando em consideração a modelagem dos dados. Os modelos ARIMA que mostram melhor desempenho são o AR, ARX, MA, ARMA, ARIMAX e SARIMAX, devido à sua capacidade de lidar com *outliers* e limites inferiores em alguns modelos. No caso dos modelos baseados em gradientes e regressão, é perceptível que eles exibem resultados semelhantes, graças às técnicas de otimização matemática conhecidas como Grid Search e Randomized Search, que permitem aprimorar os métodos utilizados.

Quando se trata de um horizonte de previsão curto, o modelo de LR apresenta melhor desempenho em comparação com os demais. No entanto, em horizontes de previsão mais longos, os modelos XGBoost e Light GBM demonstram maior precisão. A Random Forest também é capaz de realizar previsões precisas, ficando ligeiramente atrás do XGBoost em previsões de longo prazo.

Para avaliar a eficiência dos modelos ARIMA em previsões de longo prazo, utiliza-se o método conhecido como Ljung-Box. Os modelos que mostram melhor desempenho nesse contexto são o ARX, ARIMAX e SARIMAX, os quais incorporam variáveis exógenas. Esses modelos não lineares têm capacidade de previsão mais robusta em horizontes de tempo mais distantes, em comparação com os outros modelos ARIMA.

4.2 Estudo de caso

5 Conclusões

Nesta dissertação, o objetivo foi analisar a escassez de água em Curitiba e propor uma abordagem baseada nos 12 passos descritos por ALMEIDA (2013). Essa abordagem busca compreender o ambiente sem interferências e, quando necessário, considera as variáveis exógenas nos modelos ARX, ARIMAX e SARIMAX. Embora os modelos regressivos sejam adequados para lidar com interferências, sua inclusão não foi realizada nesta etapa.

Para identificar anomalias nos dados, sugere-se consultar os registros de 2020, período em que ocorreu uma grande anomalia na SANEPAR. Os resultados detalhados dessas anomalias podem ser encontrados no capítulo 4.

5.1 Limitações da pesquisa e propostas futuras

As limitações deste trabalho estão relacionadas ao tempo disponível e à abordagem dos modelos de aprendizado de máquina. Durante esta dissertação, foram explorados alguns modelos que podem ser aplicados em conjunto com séries temporais, como os modelos de rede neural LSTM, CNN, RNN, entre outros. No entanto, devido à complexidade desses modelos e à necessidade de um maior período de tempo para sua aplicação adequada, eles não foram incluídos neste momento. Os modelos que foram utilizados inicialmente foram escolhidos de forma a atender à pergunta de pesquisa levantada.

Em trabalhos futuros, seria interessante aprofundar a investigação desses modelos de previsão, uma vez que existem muitos autores na literatura que trabalham com eles. Também seria válido realizar comparações entre os modelos mais conhecidos, como o Light GBM e o XGBoost, para previsões de curto prazo. Para previsões de longo prazo, cada modelo possui sua relevância, sendo que a regressão linear (LR) é eficiente e ágil quando se trata de dados com poucas variáveis.

No próximo trabalho, que complementará esta dissertação, é recomendado abordar toda a literatura disponível, não apenas os últimos 6 anos, e considerar outras fontes, como dissertações, teses e capítulos de livros. Apesar de terem sido considerados apenas alguns artigos relevantes, existem muitos outros disponíveis sobre o assunto.

No contexto da otimização matemática, alguns modelos, como a floresta aleatória, XGBoost e Light GBM, poderiam se beneficiar do uso de técnicas para aumentar o gradiente e melhorar a precisão dos resultados. Métodos de otimização como Grid Search, Randomized Search e Bayesian Optimization (Otimização Bayesiana) foram aplicados para melhorar os modelos. Em teoria, esses métodos deveriam reduzir os erros, conforme discutido na seção de métricas (Seção 3.1), no entanto, observou-se um aumento nos erros ao longo do tempo. Um exemplo disso pode ser visto no apêndice C, onde os modelos

apresentaram um aumento dos erros de 6% para 30%. Para obter previsões mais precisas, é necessário minimizar esses erros e torná-los próximos de zero.

Portanto, seria relevante realizar uma pesquisa mais aprofundada sobre os hiperparâmetros e explorar estratégias de otimização para uma melhor utilização dos modelos baseados em árvores e gradientes.

Referências

- AHMAD, T. et al. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. **ENERGY AND BUILDINGS**, v. 165, p. 301–320, 2018. ISSN 0378-7788.
- ALMEIDA, A. T. D. **Processo de Decisão nas Organizações-Construindo Modelos de Decisão Multicritério. Atlas.** [S.l.]: São Paulo, 2013.
- BERGMEIR, C.; HYNDMAN, R.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. **Computational Statistics and Data Analysis**, v. 120, p. 70–83, 2018.
- BOROOJENI, K. et al. A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. **Electric Power Systems Research**, v. 142, p. 58–73, 2017.
- BRANDÃO, G. A. **Séries Temporais: Parte 1.** DEV Community, 2020. Disponível em: <<https://dev.to/giselyalves13/series-temporais-parte-1-13l8>>.
- BROWNLEE, J. **Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end.** [S.l.]: Machine Learning Mastery, 2016.
- BUYUKSAHIN, U.; ERTEKIN. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. **Neurocomputing**, v. 361, p. 151–163, 2019.
- Carvalho Jr., J. G.; Costa Jr., C. T. Non-iterative procedure incorporated into the fuzzy identification on a hybrid method of functional randomization for time series forecasting models. **Applied Soft Computing Journal**, Elsevier Ltd, Postgraduate Program in Electrical Engineering, Federal University of Pará, Brazil, v. 80, p. 226–242, 2019. ISSN 15684946 (ISSN). Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064441622&doi=10.1016%2Ffj.asoc.2019.03.059&partnerID=40&md5=84d0bd291cc451de280dc9ed77524736>>.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.** [S.l.: s.n.], 2016. p. 785–794.
- CHEN, Y. Y. et al. Applications of Recurrent Neural Networks in Environmental Factor Forecasting: A Review. **NEURAL COMPUTATION**, v. 30, n. 11, p. 2855–2881, 2018. ISSN 0899-7667.
- CHOU, J.-S.; NGUYEN, T.-K. Forward Forecast of Stock Price Using Sliding-Window Metaheuristic-Optimized Machine-Learning Regression. **IEEE Transactions on Industrial Informatics**, v. 14, n. 7, p. 3132–3142, 2018.
- CHOU, J.-S.; TRAN, D.-S. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. **Energy**, v. 165, p. 709–726, 2018.

- COELHO, I. et al. A GPU deep learning metaheuristic based model for time series forecasting. **Applied Energy**, v. 201, p. 412–418, 2017.
- DU, S. et al. Multivariate time series forecasting via attention-based encoder–decoder framework. **Neurocomputing**, v. 388, p. 269–279, 2020.
- GOLYANDINA, N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. **WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS**, v. 12, n. 4, 2020. ISSN 1939-0068.
- GRAFF, M. et al. Time series forecasting with genetic programming. **Natural Computing**, v. 16, n. 1, p. 165–174, 2017.
- KORSTANJE, J. **Advanced Forecasting with Python**. [S.l.]: Springer, 2021.
- KULSHRESHTHA, S.; VIJAYALAKSHMI, A. An ARIMA-LSTM hybrid model for stock market prediction using live data. **Journal of Engineering Science and Technology Review**, v. 13, n. 4, p. 117–123, 2020.
- KUMAR, G.; JAIN, S.; SINGH, U. P. Stock Market Forecasting Using Computational Intelligence: A Survey. **ARCHIVES OF COMPUTATIONAL METHODS IN ENGINEERING**, v. 28, n. 3, p. 1069–1101, 2021. ISSN 1134-3060.
- LARA-BENITEZ, P.; CARRANZA-GARCIA, M.; RIQUELME, J. C. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. **INTERNATIONAL JOURNAL OF NEURAL SYSTEMS**, v. 31, n. 3, 2021. ISSN 0129-0657.
- LI, A. W.; BASTOS, G. S. Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review. **IEEE ACCESS**, v. 8, p. 185232–185242, 2020. ISSN 2169-3536.
- LIU, H.; CHEN, C. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. **APPLIED ENERGY**, v. 249, p. 392–408, 2019. ISSN 0306-2619.
- LIU, Z. Y. et al. Forecast Methods for Time Series Data: A Survey. **IEEE ACCESS**, v. 9, p. 91896–91912, 2021. ISSN 2169-3536 J9 - IEEE ACCESS JI - IEEE Access.
- MARTINOVIĆ, M.; HUNJET, A.; TURCIN, I. Time series forecasting of the austrian traded index (Atx) using artificial neural network model. **Tehnicki Vjesnik**, v. 27, n. 6, p. 2053–2061, 2020.
- MARTINS, L. E. G.; GORSCHEK, T. Requirements engineering for safety-critical systems: A systematic literature review. **Information and Software Technology**, v. 75, p. 71–89, 2016. ISSN 0950-5849. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0950584916300568>>.
- MOON, J. et al. Temporal data classification and forecasting using a memristor-based reservoir computing system. **Nature Electronics**, v. 2, n. 10, p. 480–487, 2019.

PELLETIER, C. et al. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. **Remote Sensing of Environment**, v. 187, p. 156–168, 2016. Cited By 296. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84992151859&doi=10.1016%2fj.rse.2016.10.010&partnerID=40&md5=09efc79bab8e893b97fd21cb4844b98d>>.

PINHEIRO, N. M. **Introdução a Series Temporais — Parte 1**. Data Hackers, 2022. Disponível em: <<https://medium.com/data-hackers/series-temporais-parte-1-a0e75a512e72>>.

QUININO, R. C.; REIS, E. A.; BESSEGATO, L. F. O coeficiente de determinação r² como instrumento didático para avaliar a utilidade de um modelo de regressão linear múltipla. **Belo Horizonte: UFMG**, 1991.

REISEN, V. et al. Robust dickey–fuller tests based on ranks for time series with additive outliers. **Metrika**, v. 80, n. 1, p. 115–131, 2017. Cited By 1. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84986317325&doi=10.1007%2fs00184-016-0594-8&partnerID=40&md5=c83f82d0c372e22d5970aff448f05411>>.

RIBEIRO, M. H. D. M. et al. Time series forecasting based on ensemble learning methods applied to agribusiness, epidemiology, energy demand, and renewable energy. Pontifícia Universidade Católica do Paraná, 2021.

ROSSI, R. Relational time series forecasting. **Knowledge Engineering Review**, v. 33, 2018.

SADAEI, H. et al. Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series. **Energy**, v. 175, p. 365–377, 2019.

SALGOTRA, R.; GANDOMI, M.; GANDOMI, A. Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. **Chaos, Solitons and Fractals**, v. 138, 2020.

SAMANTA, S. et al. Learning elastic memory online for fast time series forecasting. **Neurocomputing**, v. 390, p. 315–326, 2020.

SÁNCHEZ, A. M.; DÍAZ, A. A.; LÓPEZ, A. O. A comparative study of xgboost, adaboost, and catboost in machine learning algorithms. In: SPRINGER. **International Conference on Learning and Optimization Algorithms: Theory and Applications (LOPAL)**. [S.l.], 2020. p. 292–303.

SEZER, O. B.; GUDELEK, M. U.; OZBAYOGLU, A. M. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. **APPLIED SOFT COMPUTING**, v. 90, 2020. ISSN 1568-4946.

SHEN, Z. et al. A novel time series forecasting model with deep learning. **Neurocomputing**, v. 396, p. 302–313, 2020.

SHIH, S.-Y.; SUN, F.-K.; LEE, H.-Y. Temporal pattern attention for multivariate time series forecasting. **Machine Learning**, v. 108, n. 8-9, p. 1421–1441, 2019.

SOYER, R.; ZHANG, D. Bayesian modeling of multivariate time series of counts. **WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS**. ISSN 1939-0068.

TAIEB, S. B.; ATIYA, A. F. A Bias and Variance Analysis for Multistep-Ahead Time Series Forecasting. **IEEE Transactions on Neural Networks and Learning Systems**, Institute of Electrical and Electronics Engineers Inc., Department of Computer Science, Université Libre de Bruxelles, Brussels, 1050, Belgium, v. 27, n. 1, p. 62–76, 2016. ISSN 2162237X (ISSN). Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84925431469&doi=10.1109%2FTNNLS.2015.2411629&partnerID=40&md5=e1c7f3c7a1136a0e0e4d2aff817b4008>>.

TAN, Y. F. et al. Exploring Time-Series Forecasting Models for Dynamic Pricing in Digital Signage Advertising. **FUTURE INTERNET**, v. 13, n. 10, 2021. ISSN 1999-5903.

THEODOSIOU, M. Forecasting monthly and quarterly time series using stl decomposition. **International Journal of Forecasting**, v. 27, n. 4, p. 1178–1195, 2011. Cited By 86. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052160927&doi=10.1016%2fj.ijforecast.2010.11.002&partnerID=40&md5=e8242471ba1ec14ada46ab567f3a364d>>.

TRENBERTH, K. E. Signal versus noise in the southern oscillation. **Monthly Weather Review**, v. 112, n. 2, p. 326–332, 1984.

TYRALIS, H.; PAPACHARALAMPOUS, G. Variable selection in time series forecasting using random forests. **Algorithms**, v. 10, n. 4, 2017.

URSU, E.; PEREAU, J. C. Application of periodic autoregressive process to the modeling of the Garonne river flows. **STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT**, v. 30, n. 7, p. 1785–1795, 2016. ISSN 1436-3240.

VASCONCELOS, F. **Falta d'água em curitiba e região metropolitana não É culpa só da estiagem**. 2020. Disponível em: <<https://www.brasildefato.com.br/2020/11/03/falta-d-agua-em-curitiba-e-regiao-metropolitana-nao-e-culpa-so-da-estiagem>>.

VLACHAS, P. et al. Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics. **Neural Networks**, v. 126, p. 191–217, 2020.

WANG, Y. et al. Recycling combustion ash for sustainable cement production: A critical review with data-mining and time-series predictive models. **CONSTRUCTION AND BUILDING MATERIALS**, v. 123, p. 673–689, 2016. ISSN 0950-0618.

XIE, T. et al. Hybrid forecasting model for non-stationary daily runoff series: A case study in the Han River Basin, China. **JOURNAL OF HYDROLOGY**, v. 577, 2019. ISSN 0022-1694.

XU, W. et al. Deep belief network-based AR model for nonlinear time series forecasting. **Applied Soft Computing Journal**, v. 77, p. 605–621, 2019.

YANG, W. et al. Hybrid wind energy forecasting and analysis system based on divide and conquer scheme: A case study in China. **Journal of Cleaner Production**, v. 222, p. 942–959, 2019.

YU, C. Research of time series air quality data based on exploratory data analysis and representation. In: . Institute of Electrical and Electronics Engineers Inc., 2016. ISBN 9781509023509. Cited By 5; Conference of 5th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2016 ; Conference Date: 18 July 2016 Through 20 July 2016; Conference Code:124077. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994079422&doi=10.1109%2fAgro-Geoinformatics.2016.7577697&partnerID=40&md5=fef861624a35632bf2d84acf63986bbe>>.

A Apêndice - Comparação dos modelos de previsão de series temporais média de 24h

$(p = 7, d = 1, q = 7)(P = 2, D = 1, Q = 1)_{M=12}$ Média 24h

Tabela 6: Comparação dos modelos com 1 dia de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,096	0,306	0,419
ARX	0,118	0,377	0,513
MA	0,093	0,296	0,403
ARMA	0,102	0,325	0,435
ARIMA	0,095	0,302	0,405
SARIMA	0,105	0,342	0,450
ARIMAX	0,119	0,378	0,511
SARIMAX	0,118	0,377	0,512
LR	0,015	0,069	0,077
RFR	0,190	0,624	0,672
XGBRegressor	0,207	0,683	0,720
LGBMRegressor	0,184	0,599	0,655

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 7: Comparação dos modelos com 1 dia de antecedência 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,084	0,285	0,366
ARX	0,103	0,354	0,459
MA	0,082	0,278	0,361
ARMA	0,086	0,295	0,372
ARIMA	0,082	0,280	0,351
SARIMA	0,097	0,333	0,421
ARIMAX	0,102	0,353	0,458
SARIMAX	0,104	0,358	0,463
LR	0,014	0,066	0,073
RFR	0,172	0,587	0,633
XGBRegressor	0,192	0,658	0,692
LGBMRegressor	0,166	0,564	0,616

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 8: Comparação dos modelos com 1 dia de antecedência 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,100	0,329	0,424
ARX	0,137	0,462	0,586
MA	0,102	0,336	0,431
ARMA	0,102	0,340	0,433
ARIMA	0,103	0,346	0,440
SARIMA	0,118	0,398	0,501
ARIMAX	0,137	0,461	0,587
SARIMAX	0,138	0,464	0,590
LR	0,018	0,087	0,098
RFR	0,153	0,494	0,587
XGBRegressor	0,170	0,560	0,643
LGBMRegressor	0,145	0,465	0,568

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 9: Comparação dos modelos com 1 dia de antecedência 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,093	0,302	0,165
ARX	0,123	0,402	0,283
MA	0,107	0,344	0,460
ARMA	0,097	0,316	0,424
ARIMA	0,094	0,303	0,406
SARIMA	0,106	0,350	0,448
ARIMAX	0,120	0,394	0,521
SARIMAX	0,122	0,401	0,530
LR	0,016	0,074	0,084
RFR	0,176	0,579	0,642
XGBRegressor	0,194	0,643	0,694
LGBMRegressor	0,170	0,554	0,624

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 10: Comparação dos modelos com 7 dias de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,093	0,296	0,399
ARX	0,118	0,377	0,524
MA	0,104	0,329	0,444
ARMA	0,103	0,330	0,439
ARIMA	0,108	0,342	0,463
SARIMA	0,111	0,360	0,487
ARIMAX	0,118	0,379	0,525
SARIMAX	0,118	0,379	0,525
LR	1,197	5,230	5,230
RFR	0,224	0,705	0,821
XGBRegressor	0,260	0,823	0,934
LGBMRegressor	0,215	0,673	0,793

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 11: Comparação dos modelos com 7 dias de antecedência 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,073	0,245	0,319
ARX	0,093	0,319	0,423
MA	0,080	0,269	0,353
ARMA	0,081	0,274	0,347
ARIMA	0,087	0,292	0,384
SARIMA	0,095	0,324	0,438
ARIMAX	0,093	0,318	0,422
SARIMAX	0,094	0,320	0,424
LR	1,174	5,224	5,224
RFR	0,188	0,630	0,712
XGBRegressor	0,223	0,756	0,828
LGBMRegressor	0,179	0,598	0,684

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 12: Comparação dos modelos com 7 dias de antecedência 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,118	0,383	0,499
ARX	0,147	0,479	0,632
MA	0,125	0,403	0,530
ARMA	0,117	0,384	0,494
ARIMA	0,120	0,393	0,505
SARIMA	0,131	0,437	0,544
ARIMAX	0,148	0,480	0,632
SARIMAX	0,148	0,481	0,636
LR	1,161	5,212	5,213
RFR	0,187	0,578	0,755
XGBRegressor	0,222	0,693	0,870
LGBMRegressor	0,177	0,543	0,727

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 13: Comparação dos modelos com 7 dias de antecedência 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,098	0,316	0,177
ARX	0,125	0,406	0,305
MA	0,105	0,337	0,450
ARMA	0,097	0,312	0,418
ARIMA	0,097	0,314	0,420
SARIMA	0,118	0,386	0,506
ARIMAX	0,124	0,402	0,546
SARIMAX	0,125	0,405	0,551
LR	1,183	5,224	5,224
RFR	0,208	0,656	0,787
XGBRegressor	0,243	0,775	0,901
LGBMRegressor	0,199	0,623	0,759

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 14: Comparação dos modelos com 14 dias de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,105	0,334	0,445
ARX	0,126	0,399	0,548
MA	0,106	0,336	0,447
ARMA	0,110	0,350	0,463
ARIMA	0,111	0,353	0,477
SARIMA	0,114	0,367	0,489
ARIMAX	0,126	0,401	0,547
SARIMAX	0,126	0,401	0,547
LR	2,606	11,394	11,394
RFR	0,221	0,696	0,812
XGBRegressor	0,269	0,859	0,962
LGBMRegressor	0,215	0,673	0,792

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 15: Comparação dos modelos com 14 dias de antecedência 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,078	0,264	0,346
ARX	0,090	0,309	0,430
MA	0,079	0,265	0,349
ARMA	0,093	0,317	0,403
ARIMA	0,088	0,295	0,389
SARIMA	0,092	0,315	0,402
ARIMAX	0,090	0,308	0,429
SARIMAX	0,090	0,308	0,429
LR	2,558	11,388	11,388
RFR	0,185	0,619	0,702
XGBRegressor	0,233	0,790	0,859
LGBMRegressor	0,179	0,598	0,683

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 16: Comparação dos modelos com 14 dias de antecedência 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,118	0,378	0,504
ARX	0,120	0,384	0,560
MA	0,120	0,385	0,509
ARMA	0,107	0,344	0,464
ARIMA	0,105	0,338	0,462
SARIMA	0,113	0,364	0,496
ARIMAX	0,120	0,384	0,560
SARIMAX	0,119	0,383	0,558
LR	2,531	11,376	11,377
RFR	0,186	0,572	0,748
XGBRegressor	0,227	0,710	0,889
LGBMRegressor	0,177	0,542	0,725

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 17: Comparação dos modelos com 14 dias de antecedência 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,104	0,335	0,204
ARX	0,122	0,393	0,297
MA	0,106	0,340	0,452
ARMA	0,097	0,311	0,423
ARIMA	0,099	0,318	0,431
SARIMA	0,113	0,365	0,492
ARIMAX	0,121	0,389	0,539
SARIMAX	0,122	0,393	0,543
LR	2,577	11,388	11,388
RFR	0,206	0,648	0,779
XGBRegressor	0,251	0,804	0,927
LGBMRegressor	0,198	0,623	0,758

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 18: Comparação dos modelos com 30 dias de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,121	0,383	0,514
ARX	0,135	0,432	0,592
MA	0,120	0,379	0,510
ARMA	0,120	0,383	0,508
ARIMA	0,124	0,395	0,527
SARIMA	0,126	0,405	0,538
ARIMAX	0,136	0,434	0,594
SARIMAX	0,136	0,435	0,596
LR	5,827	25,483	25,484
RFR	0,224	0,705	0,821
XGBRegressor	0,282	0,902	0,998
LGBMRegressor	0,211	0,659	0,780

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 19: Comparação dos modelos com 30 dias de antecedência 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,091	0,311	0,390
ARX	0,086	0,302	0,434
MA	0,090	0,306	0,383
ARMA	0,089	0,304	0,384
ARIMA	0,100	0,343	0,426
SARIMA	0,098	0,337	0,412
ARIMAX	0,086	0,301	0,433
SARIMAX	0,086	0,302	0,434
LR	5,721	25,478	25,478
RFR	0,187	0,628	0,710
XGBRegressor	0,245	0,831	0,896
LGBMRegressor	0,174	0,580	0,666

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 20: Comparação dos modelos com 30 dias de antecedência 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,117	0,375	0,495
ARX	0,141	0,462	0,628
MA	0,120	0,384	0,504
ARMA	0,118	0,384	0,496
ARIMA	0,120	0,390	0,509
SARIMA	0,132	0,431	0,570
ARIMAX	0,140	0,459	0,627
SARIMAX	0,142	0,463	0,627
LR	5,663	25,466	25,466
RFR	0,189	0,583	0,759
XGBRegressor	0,239	0,754	0,918
LGBMRegressor	0,174	0,532	0,716

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 21: Comparação dos modelos com 30 dias de antecedência 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	0,114	0,367	0,237
ARX	0,137	0,447	0,360
MA	0,113	0,361	0,477
ARMA	0,120	0,385	0,508
ARIMA	0,117	0,375	0,497
SARIMA	0,124	0,404	0,531
ARIMAX	0,136	0,443	0,596
SARIMAX	0,137	0,446	0,601
LR	5,763	25,477	25,477
RFR	0,208	0,657	0,788
XGBRegressor	0,264	0,847	0,961
LGBMRegressor	0,195	0,610	0,746

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

B Apêndice - Comparação dos modelos de previsão com o método Ljung Box

Modelos ARIMAS para previsão de longo prazo usando a defasagem de 10.

Tabela 22: Comparação dos modelos Ljung Box **Treinamento**

Ljung Box	Estatística de Teste	Valor De p
ARX	6,30	0,79
AR	7,13	0,07
MA	34,34	0,00
ARMA	11,60	0,31
ARIMA	13,01	0,22
SARIMA	10,17	0,43
ARIMAX	30,36	0,00
SARIMAX	11,63	0,31

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 23: Comparação dos modelos Ljung Box **Validação**

Ljung Box	Estatística de Teste	Valor De p
ARX	7,47	0,68
AR	2,43	0,99
MA	1,39	1,00
ARMA	5,42	0,86
ARIMA	4,04	0,95
SARIMA	4,45	0,93
ARIMAX	0,02	1,00
SARIMAX	0,04	1,00

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 24: Comparação dos modelos Ljung Box **Teste**

Ljung Box	Estatística de Teste	Valor De p
ARX	0,86	1,00
AR	7,80	0,65
MA	7,89	0,64
ARMA	19,34	0,04
ARIMA	9,50	0,49
SARIMA	3,57	0,97
ARIMAX	0,60	1,00
SARIMAX	3,72	0,96

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

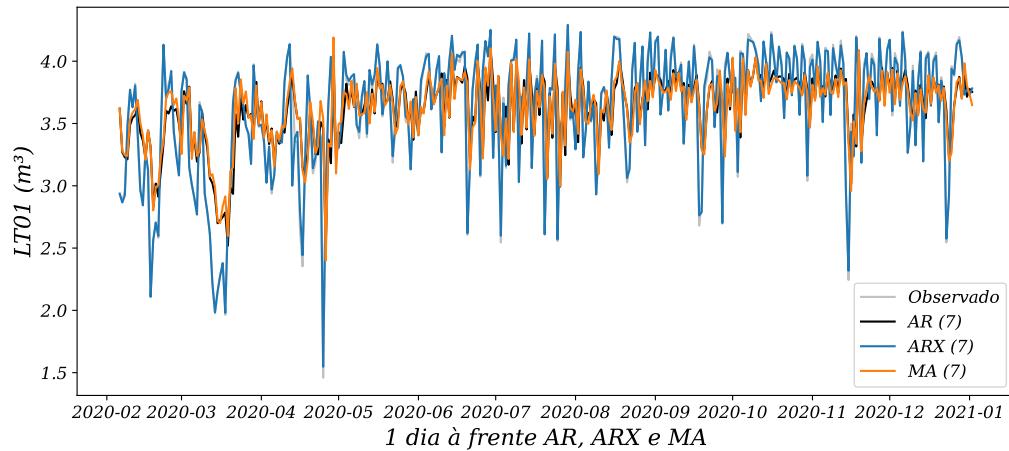
Tabela 25: Comparação dos modelos Ljung Box **Completo**

Ljung Box	Estatística de Teste	Valor De p
ARX	4,70	0,91
AR	4,26	0,16
MA	49,16	0,00
ARMA	40,49	0,00
ARIMA	40,49	0,00
SARIMA	40,49	0,00
ARIMAX	60,91	0,00
SARIMAX	5,83	0,83

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

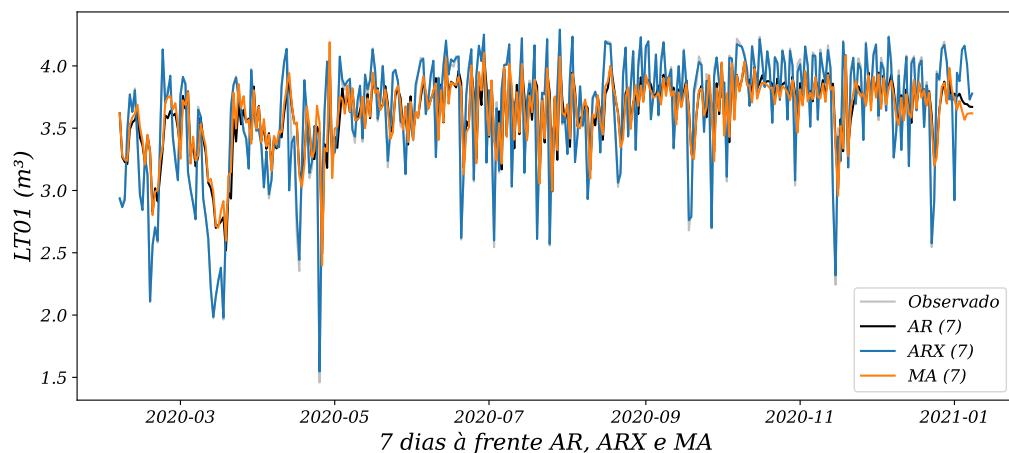
C Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h

Figura 41: Comparação dos modelos AR, ARX e MA, 1 dia à frente



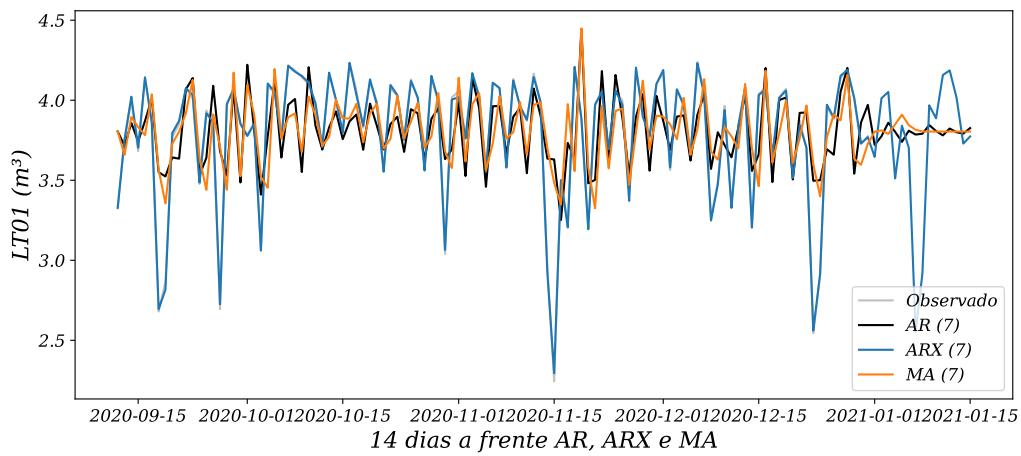
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 42: Comparação dos modelos AR, ARX e MA, 7 dias à frente



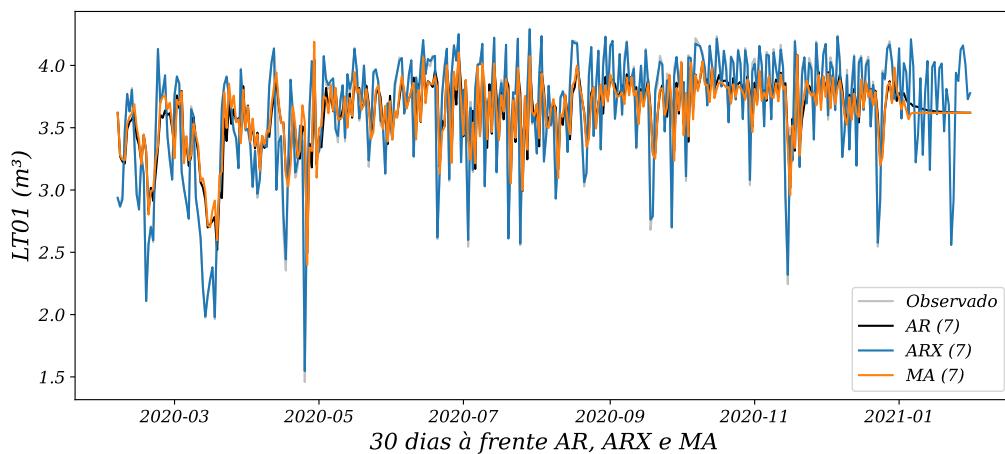
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 43: Comparação dos modelos AR, ARX e MA, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

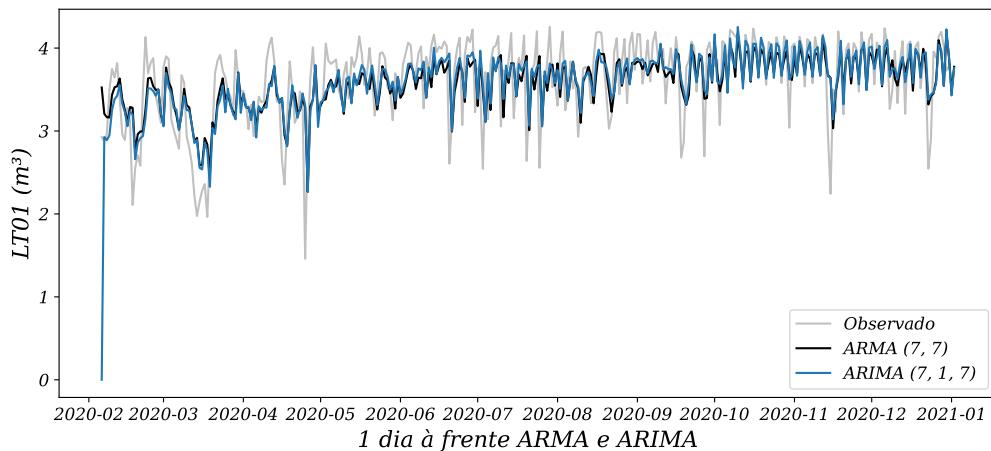
Figura 44: Comparação dos modelos AR, ARX e MA, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

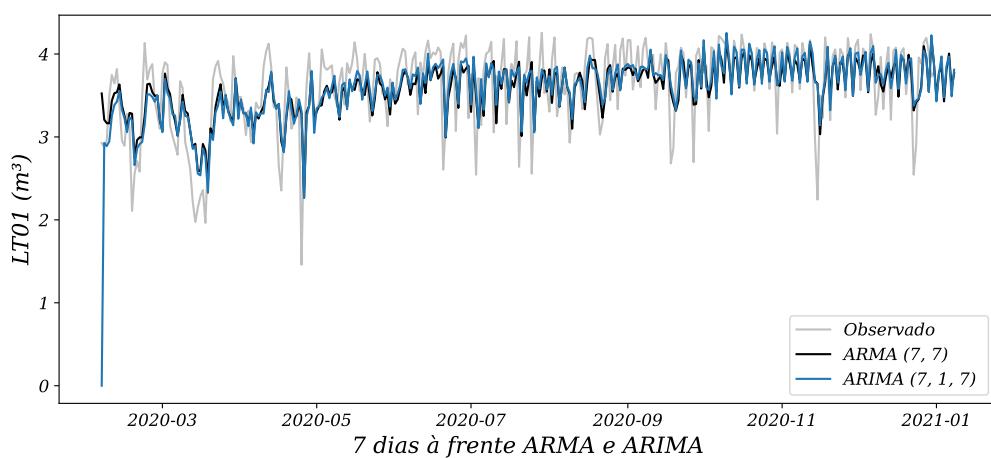
D Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h

Figura 45: Comparação dos modelos ARMA e ARIMA, 1 dia à frente



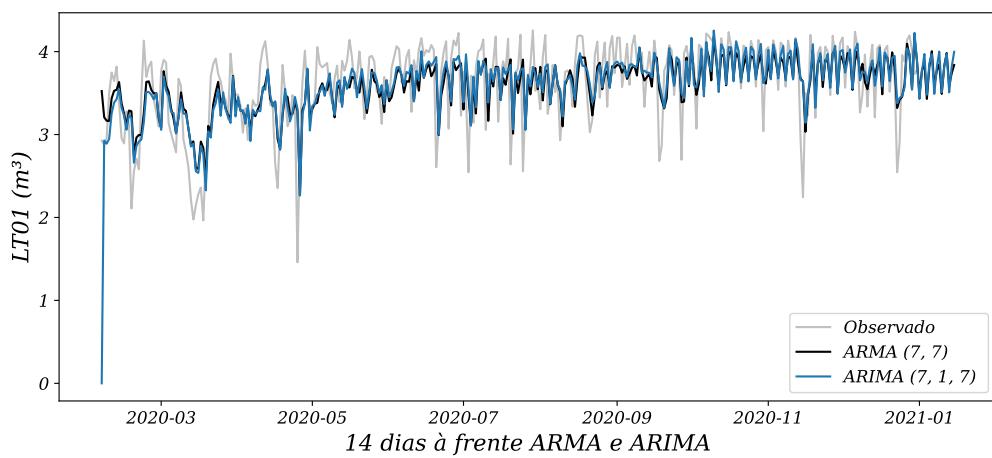
Fonte: Autoria própria.

Figura 46: Comparação dos modelos ARMA e ARIMA, 7 dias à frente



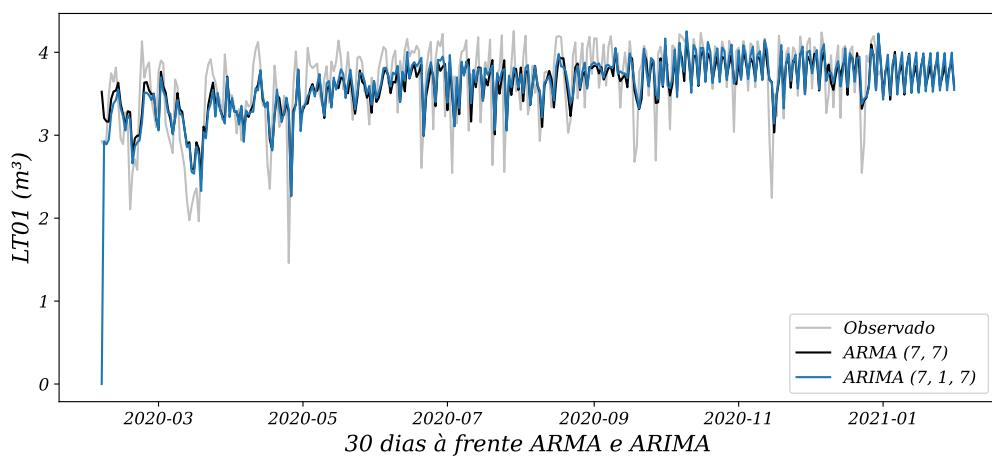
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 47: Comparação dos modelos ARMA e ARIMA, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

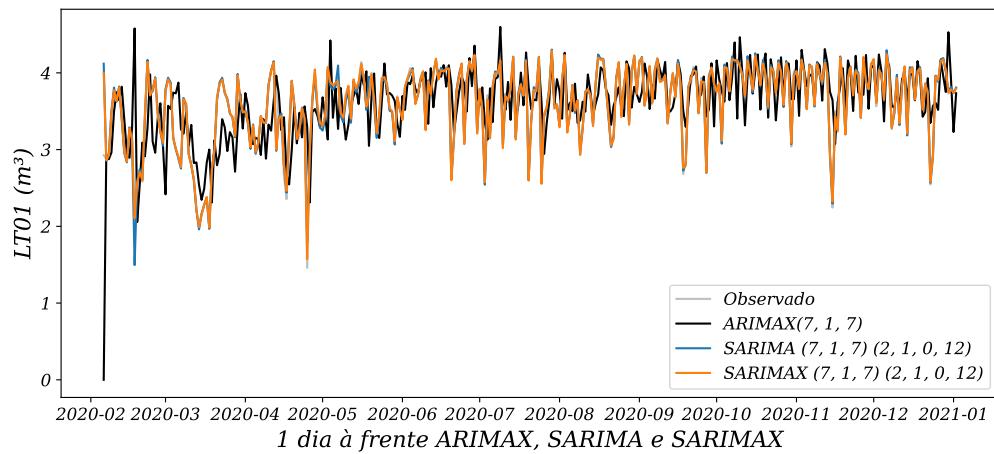
Figura 48: Comparação dos modelos ARMA e ARIMA, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

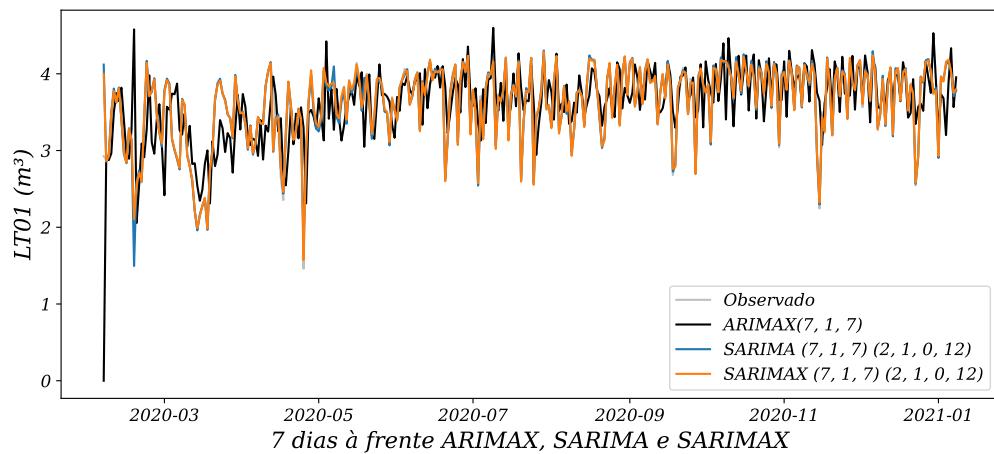
E Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e SARIMAX (7,1,7) (2,1,0,12) 24h

Figura 49: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente



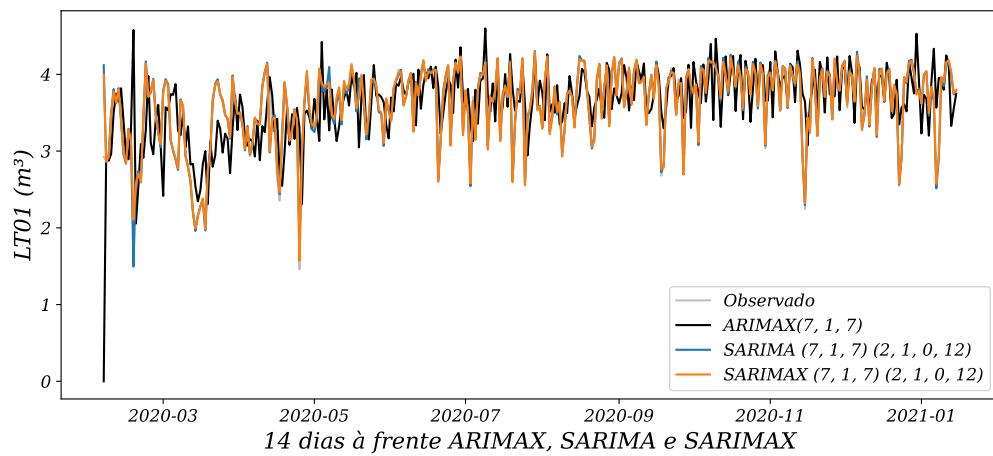
Fonte: Autoria própria.

Figura 50: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente



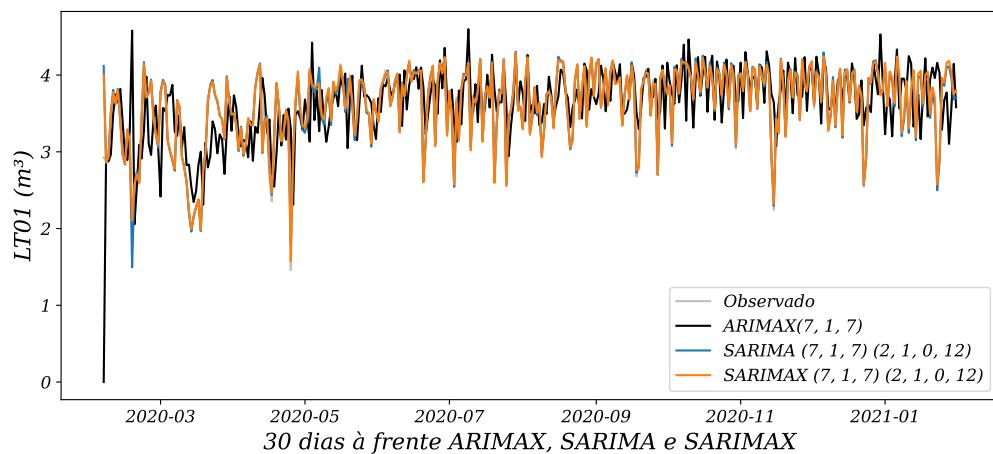
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 51: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 52: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)