



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO, DE CIÊNCIAS EXATAS E EDUCAÇÃO
DEPARTAMENTO DE ENG. DE CONTROLE, AUTOMAÇÃO E COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Cláudio Lourenço Moreira

**Previsão do Nível do Lago Guaíba a partir de Dados Meteorológicos: Aplicação
de Técnicas de Aprendizado de Máquina com Regressão Ridge**

Blumenau
2025

Cláudio Lourenço Moreira

**Previsão do Nível do Lago Guaíba a partir de Dados Meteorológicos: Aplicação
de Técnicas de Aprendizado de Máquina com Regressão Ridge**

Trabalho de Conclusão de Curso de Graduação em Engenharia de Controle e Automação do Centro Tecnológico, de Ciências Exatas e Educação da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenheiro de Controle e Automação.

Orientador: Prof. Dr. Maiquel de Brito

Blumenau
2025

Ficha de identificação da obra

A ficha de identificação é elaborada pelo próprio autor.

Orientações em:

<http://portalbu.ufsc.br/ficha>

Cláudio Lourenço Moreira

**Previsão do Nível do Lago Guaíba a partir de Dados Meteorológicos: Aplicação
de Técnicas de Aprendizado de Máquina com Regressão Ridge**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de “Engenheiro de Controle e Automação” e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Controle e Automação.

Blumenau, dia de julho de 2025.

Banca Examinadora:

Prof. Dr. Maiquel de Brito
Instituição xxxx

Prof. Segundo, Dr.
Instituição xxxx

Prof. Terceiro, Dr.
Instituição xxxx

Dedico este trabalho aos meus pais, amigos, professores, e
a todos que passaram e deixaram uma marca, seja ela
qual for, ao longo desse capítulo da graduação.

AGRADECIMENTOS

Ao longo dessa jornada árdua, desafiadora, porém imensamente prazerosa de se viver, diversas pessoas foram importantes para a minha formação como engenheiro, pessoa e profissional. Não poderia começar diferente meu agradecimento, senão aos meus pais, que são minha base, minha referência e minha motivação para sempre buscar ser melhor em todos os sentidos. Dar orgulho a eles é o que me move, e sem eles eu não teria sequer começado essa jornada.

Na mesma prateleira de troféus e presentes dessa vida, estão meus amigos, que tenho o prazer e a alegria de dizer que ultrapassaram a barreira da universidade e se tornaram meus amigos pra vida, ou melhor, uma segunda família. A turma 19.1 foi especial desde o primeiro dia, e ao longo dos semestres, pandemia e demais percausos, essa turma se manteve unida e compartilhando dúvidas, risadas, histórias e momentos que guardo com imenso carinho. Meu obrigado a todos da turma 19.1, em especial ao time que me acompanhou um passo a mais de perto, e caminhou comigo dentro e fora da sala de aula. Vítor, Augusto, Samuel, Felipe (Borto) e Lauro, os Business Boys, a melhor equipe de robótica, a melhor turma de visão, de controle, de sistemas computacionais, ou de qualquer disciplina do curso, modéstia a parte. Vocês me mostraram o que é ser um estudante e um profissional de extrema qualidade, e eu sou muito grato pela convivência e pelos ensinamentos diários que tive perto de vocês.

Não seria justo deixar de dedicar um parágrafo separado ao professor que caminhou comigo como aluno, amigo e atualmente, família, João Victor Zanoni. Desde o primeiro dia de aula, você esteve ao meu lado como uma das maiores referências que tive na graduação. A tua determinação, dedicação, qualidade, profissionalismo e "n" outras virtudes fizeram do curso uma formação ímpar, sem sombra de dúvidas a melhor e mais especial que eu poderia ter. A UFSC, embora seja uma universidade de excelência, jamais teria me proporcionado uma formação tão completa e especial se não fosse por você. Cada trabalho, seja ele em dupla, trio, quarteto ou qualquer tamanho que fosse, eu sabia que seria no mínimo, eu e você. Foi uma honra vivenciar essa etapa da vida e todos esses obstáculos contigo, parceiro. Obrigado por tudo, professor Zanoni.

Aos meus amigos que tive a felicidade de conhecer fora da turma ao qual entrei, principalmente através da Integre Jr., meu muito obrigado. Passar pelo MEJ com vocês me trouxe um crescimento pessoal e profissional que nenhuma disciplina poderia agregar, e tudo só se tornou mais especial porque essa empresa estava formada por pessoas como vocês. Agradeço por cada evento, cada reunião, desafios, projetos, imersões, viagens e momentos que cultivamos juntos.

Por fim, meu agradecimento a todos os professores que passaram pela minha vida acadêmica, e que contribuíram de alguma forma para a minha formação. o Campus de Blumenau pode ser pequeno em tamanho, mas em compensação tem uma equipe de

professores extremamente qualificados e com grande gabarito, que tornaram a transmissão de conhecimento algo rico e determinante na minha formação. Em especial, agradeço ao meu professor orientador, Maiquel de Brito, que me acompanhou nesse trabalho e me deu todo o suporte necessário nessa reta final, além de estar presente desde o meu primeiro ano no curso.

É difícil agradecer a todos que contribuíram de alguma forma nessa metade de década da minha vida, acredito que cada pessoa que passar por nós e deixa uma marca, não importa o tamanho, é importante para chegarmos aonde estamos agora, e por isso, agradeço de coração a todos que fizeram parte de capítulo tão transformador.

"A única maneira de se definir o limite do possível é ir além dele, para o impossível."
(CLARKE, 1962)

RESUMO

Este trabalho descreve o desenvolvimento de um modelo de previsão do nível do Lago Guaíba, no Rio Grande do Sul, utilizando dados meteorológicos integrados a técnicas de aprendizado de máquina, especificamente a Regressão Ridge. O estudo envolveu a coleta e preparação de dados meteorológicos do INMET e níveis hidrométricos dos rios da bacia do Guaíba, abrangendo etapas de limpeza, normalização, redução de dimensionalidade e sincronização temporal das bases. O modelo foi implementado utilizando diferentes valores do parâmetro de regularização alpha e múltiplas proporções de divisão dos dados em conjuntos de treinamento e teste (80:20, 70:30, 60:40). As métricas MSE, apropriadas indicam bom desempenho, com o coeficiente de determinação entre 87% e 89% e erro absoluto médio de 15 a 16,81 centímetros do nível do rio, demonstrando robustez e baixo risco de sobreajuste. Os resultados confirmam que variáveis meteorológicas têm forte influência no comportamento hidrológico do rio.

Palavras-chave: Previsão hidrológica; Aprendizado de máquina; Regressão Ridge; Lago Guaíba; Meteorologia.

ABSTRACT

This study aimed to develop a forecasting model for the Guaíba River level in Rio Grande do Sul, using meteorological data integrated with machine learning techniques, specifically Ridge Regression. The study involved the collection and preparation of meteorological data from INMET and water level data from the Guaíba basin rivers, encompassing steps such as data cleaning, normalization, dimensionality reduction, and temporal synchronization of the datasets. The model was implemented using different values of the regularization parameter alpha and multiple train-test split ratios (80:20, 70:30, 60:40). The MSE, RMSE, MAE, and R² metrics indicated good performance, with a coefficient of determination between 87% and 89% and a mean absolute error of 15 to 16.81 centimeters in river level, demonstrating robustness and low risk of overfitting. The results confirm that meteorological variables have a strong influence on the river's hydrological behavior.

Keywords: Hydrological forecasting; Machine learning; Ridge Regression; Guaíba River; Meteorology.

LISTA DE FIGURAS

Figura 1 – Rios que desemboram e influenciam no nível do Lago Guaíba.	19
Figura 2 – Rio Jacuí e seu desemboque no Lago Guaíba.	19
Figura 3 – Relação entre o índice de felicidade e expectativa de vida.	21
Figura 4 – Diferentes correlações entre variáveis.	22
Figura 5 – Interpretação de uma regressão linear	24
Figura 6 – Situações de inadequação da RLS	25
Figura 7 – Passo 1 da regressão linear pelo método MQO.	28
Figura 8 – Passo 2 da regressão linear pelo método MQO.	29
Figura 9 – Passo 3 da regressão linear pelo método MQO.	30
Figura 10 – Iterações da aplicação do método MQO	31
Figura 11 – Passo 10 da regressão linear pelo método MQO.	32
Figura 12 – Dados meteorológicos.	36
Figura 13 – Dados do nível do rio.	37
Figura 14 – Limpeza dos dados coletados.	38
Figura 15 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior (°C) não tratado.	38
Figura 16 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior (°C) tratado. .	39
Figura 17 – Gráfico de Umidade Relativa Máxima na Hora Anterior (%) sem tratamento.	40
Figura 18 – Gráfico de Umidade Relativa Máxima na Hora Anterior (%) com tratamento.	40
Figura 19 – Comparativo de gráficos de temperatura em diferentes tipos de medição	42
Figura 20 – Gráfico de Temperatura do Ar - Bulbo Seco	43
Figura 21 – Gráfico de Pressão Atmosférica ao Nível da Estação.	43
Figura 22 – Gráfico de Vento - Velocidade Horária.	44
Figura 23 – Gráfico de Umidade Relativa do Ar	44
Figura 24 – Gráfico de Radiação Global	45
Figura 25 – Gráfico de Precipitação Total	45
Figura 26 – Gráfico comparativo de temperatura e nível quanto a sazonalidade . .	46
Figura 27 – Gráfico comparativo de pressão atmosférica e nível quanto a sazonalidade	46
Figura 28 – Gráfico comparativo de radiação global e nível quanto a sazonalidade .	47
Figura 29 – Nivelamento superior e inferior dos dados dos rios.	48
Figura 30 – Exemplo de código Python para configuração da Regressão Ridge . .	51
Figura 31 – Gráfico modelo de previsão de nível do Lago Guaíba com $\alpha = 1$ e split de treinamento e teste 60:40	52
Figura 32 – Gráfico modelo de previsão de nível do Lago Guaíba com $\alpha = 1$ e split de treinamento e teste 70:30	52

Figura 33 – Gráfico modelo de previsão de nível do Lago Guaíba com <code>alpha</code> = 1 e split de treinamento e teste 80:20	53
Figura 34 – Diferenças entre os valores reais e previstos do nível do Lago Guaíba . .	56

LISTA DE TABELAS

Tabela 1 – Tabela de Valores Ordenados: Variável Independente vs. Variável Dependente	28
Tabela 2 – Tabela de tipos de dados da base de informações meteorológicas.	37
Tabela 3 – Tabela de dados meteorológicos reduzidos - 1 ^a Filtragem	41
Tabela 4 – Datas mínima e máxima disponíveis para cada rio analisado	48
Tabela 5 – Tabela de avaliação de desempenho do modelo com alpha 0.01	53
Tabela 6 – Tabela de avaliação de desempenho do modelo com alpha 0.1	53
Tabela 7 – Tabela de avaliação de desempenho do modelo com alpha 1	53
Tabela 8 – Tabela de avaliação de desempenho do modelo com alpha 10	54
Tabela 9 – Tabela de avaliação de desempenho do modelo com alpha 100	54
Tabela 10 – Tabela de avaliação de desempenho do modelo com alpha 1000000.	55
Tabela 11 – Diferença entre os valores reais e previstos do nível do Lago Guaíba	56

LISTA DE ABREVIATURAS E SIGLAS

CSV	Comma Separated Values (Valores Separados por Vírgula)
INMET	Instituto Nacional de Meteorologia
IQR	Interquartile Range (Intervalo Interquartil)
LGBMR	Light Gradient Boosting Machine Regression (Regressão por Máquina de Reforço de Gradiente Leve)
ML	Machine Learning (Aprendizado de Máquina)
MLPR	Multilayer Perceptron Regression (Regressão por Perceptron Multicamadas)
MQO	Mínimos Quadrados Ordinários
NSE	Nash-Sutcliffe Efficiency (Eficiência de Nash-Sutcliffe)
RFR	Random Forest Regression (Regressão por Floresta Aleatória)
RLM	Regressão Linear Múltipla
RLS	Regressão Linear Simples
RSS	Residual Sum of Squares (Soma dos Quadrados dos Resíduos)
SEMA-RS	Secretaria do Meio Ambiente e Infraestrutura do Rio Grande do Sul
SVR	Support Vector Regression (Regressão por Vetores de Suporte)

SUMÁRIO

1	INTRODUÇÃO	15
1.1	OBJETIVOS	16
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	AS MUDANÇAS CLIMÁTICAS E AS CATÁSTROFES NATURAIS .	17
2.1.1	Cenário de enchentes no sul do Brasil	17
2.1.2	Dinâmica do Lago Guaíba	18
2.2	APRENDIZADO DE MÁQUINA	20
2.2.1	Categorias de aprendizado de máquina	20
2.3	REGRESSÃO	21
2.4	REGRESSÃO LINEAR	22
2.5	MÉTODO DOS QUADRADOS ORDINÁRIOS	25
2.6	MODELO RIDGE	32
3	DESENVOLVIMENTO	35
3.1	COLETA DE DADOS	35
3.2	PRÉ PROCESSAMENTO DOS DADOS	36
3.3	LIMPEZA DOS DADOS	37
3.4	REDUÇÃO DOS DADOS	41
3.5	TRANSFORMAÇÃO DOS DADOS	48
3.6	DIVISÃO DOS DADOS EM CONJUNTOS DE TREINAMENTO E TESTE	49
3.7	APLICAÇÃO DO MODELO DE PREVISÃO	50
4	ANÁLISE DOS RESULTADOS	52
5	CONCLUSÃO	59
	REFERÊNCIAS	61

1 INTRODUÇÃO

As mudanças climáticas têm intensificado a frequência e a severidade de eventos climáticos extremos, como chuvas intensas, secas prolongadas e inundações, impactando diretamente grandes centros urbanos e áreas rurais (VEJA, 2024). No Brasil, esses fenômenos têm se tornado mais frequentes, afetando a região Sul com inundações causadas por chuvas intensas que representam um desafio recorrente. As enchentes de 2024, que devastaram diversas cidades gaúchas, evidenciaram a necessidade de ferramentas eficazes para prever e mitigar os impactos de desastres naturais.

O problema central abordado neste trabalho é a previsão do nível do Lago Guaíba, um dos principais mananciais de abastecimento de água de Porto Alegre e região metropolitana. A capacidade de prever o nível do rio é crucial para antecipar eventos de inundaçāo, que podem causar perdas materiais significativas, deslocamento de populações e até mesmo mortes. A previsão do nível do rio, utilizando dados meteorológicos como variáveis preditoras, permite a elaboração de planos de evacuação, a alocação eficiente de recursos para mitigação de desastres e a proteção de infraestruturas críticas, contribuindo para a segurança e o bem-estar da população (ANDRADE, M. M. *et al.*, 2017).

A importância de resolver esse problema reside na possibilidade de reduzir os impactos socioeconômicos e ambientais causados pelas cheias. As inundações no Rio Grande do Sul, como as observadas em 1941 e 2024, demonstram a vulnerabilidade da região a eventos climáticos extremos (VEJA, 2024). A elaboração de um modelo de previsão do nível do rio pode embasar decisões de políticas públicas e defesa civil, além de otimizar a gestão de recursos hídricos e outras aplicações que tangem a preservação e uso consciente do lago.

Apesar dos avanços em modelos hidrológicos e de previsão, as soluções atuais apresentam limitações que justificam a busca por novas abordagens. Algumas delas são computacionalmente custosas ou requerem dados extensivos que nem sempre estão disponíveis, dificultando sua implementação em tempo real (ANDRADE, M. M. *et al.*, 2017).

No contexto do Lago Guaíba, a problemática é particularmente relevante devido à sua sensibilidade a chuvas intensas, tanto em seu leito principal quanto em seus afluentes, como os rios Jacuí, Caí, Gravataí e Sinos. As cheias de 2024 destacaram como a variabilidade climática e as contribuições dos afluentes amplificam as flutuações no nível do Guaíba, tornando essencial o desenvolvimento de modelos preditivos robustos (VEJA, 2024). Este trabalho propõe o uso de dados meteorológicos, como precipitação, temperatura, umidade e pressão atmosférica, combinados com técnicas de aprendizado de máquina, para prever o nível do rio, oferecendo uma ferramenta para a gestão de riscos de inundações.

1.1 OBJETIVOS

Posto o contexto da realização do trabalho, é possível definir os objetivos que guiarão o desenvolvimento do modelo preditivo, com a estipulação de objetivos e métricas de interesse para a avaliação do modelo.

O objetivo geral deste trabalho é desenvolver um modelo de previsão do nível do Lago Guaíba utilizando dados meteorológicos por meio de técnicas de aprendizado de máquina, especificamente a regressão Ridge.

Os objetivos específicos são:

- Coletar e tratar dados meteorológicos e de níveis de rios relevantes para o modelo.
- Aplicar técnicas de pré-processamento de dados, incluindo limpeza e redução de dimensionalidade, para garantir a qualidade das informações.
- Implementar e treinar um modelo de regressão Ridge para prever o nível do Lago Guaíba com base nos dados meteorológicos.
- Avaliar o desempenho do modelo utilizando métricas de erro como MSE, RMSE, MAE e R^2 .
- Analisar os resultados obtidos e propor ajustes para otimizar a previsão.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos teóricos que sustentam o desenvolvimento do modelo preditivo para o nível do Lago Guaíba, com base em dados meteorológicos e na técnica de Regressão *Ridge*. Inicialmente, aborda-se o impacto das mudanças climáticas e sua relação com eventos extremos, como as enchentes, destacando a relevância do monitoramento de rios em regiões vulneráveis, como o Rio Grande do Sul. Em seguida, são discutidos os princípios de aprendizado de máquina, com ênfase em abordagens supervisionadas, que formam a base para a modelagem proposta. Posteriormente, explora-se a teoria da regressão, incluindo a regressão linear simples e múltipla, e o método dos Mínimos Quadrados Ordinários, que são fundamentais para compreender a Regressão *Ridge*. Por fim, detalha-se a técnica da regressão proposta para implementação, destacando sua capacidade de lidar com multicolinearidade e melhorar a generalização do modelo, essencial para a previsão precisa do nível do rio em cenários de alta variabilidade climática.

2.1 AS MUDANÇAS CLIMÁTICAS E AS CATÁSTROFES NATURAIS

As grandes cidades brasileiras enfrentam desafios mais frequentes relacionados às mudanças climáticas, que agravam problemas como enchentes, inundações e deslizamentos. Projeções indicam que, até 2030, a mancha urbana de São Paulo pode aumentar em até 38%, ampliando o risco para mais de 20% das áreas de expansão urbana, que se tornarão suscetíveis a acidentes naturais (NOBRE *et al.*, 2011). O estudo também destaca que o aumento na frequência de eventos de chuvas intensas pode dobrar o número de dias com precipitação acima de 10 milímetros, agravando a vulnerabilidade da população, especialmente nas áreas periféricas e de menor infraestrutura.

2.1.1 Cenário de enchentes no sul do Brasil

Com base no histórico das enchentes no Rio Grande do Sul, observa-se que os desastres relacionados ao excesso de chuvas não são um fenômeno recente. Desde 1941, o estado lida com eventos catastróficos, como a enchente que devastou Porto Alegre naquele ano, considerada uma das mais graves da história da cidade. Ao longo das décadas, esses episódios continuaram a ocorrer, expondo a vulnerabilidade da região diante de chuvas intensas e repentinhas. A combinação de fatores naturais, como a geografia da região e os ciclos climáticos, aliado as ações humanas nocivas ao meio ambiente, contribui para a repetição e intensificação dessas tragédias (VEJA, 2024).

Em Santa Catarina, estado adjacente ao Rio Grande do Sul, as enchentes também são fenômenos recorrentes que, ao longo dos anos, têm causado impactos sociais, econômicos e ambientais. Um dos eventos mais recentes foi registrado em maio de 2024, quando o estado registrou vários dias com altos índices pluviométricos, levando ao transbordamento

de rios, deslizamentos de terra e bloqueios em diversas rodovias (G1, 2024b).

2.1.2 Dinâmica do Lago Guaíba

O Lago Guaíba, principal manancial de abastecimento de água para a capital do Rio Grande do Sul e região, é alvo de estudo sobre diversos temas, incluindo sua hidrodinâmica e nível ao longo do ano. O Lago Guaíba apresenta flutuações significativas no volume de descarga, variando de 407 m³/s a 14.270 m³/s (ANDRADE, M. M. *et al.*, 2017). Grande parte desta variação sofre influência dos rios que desemborciam no lago, como Rio Jacuí, que contribui com cerca de 84,6% da água que aflui ao lago, além dos rios Sinos, Caí e Gravataí que contribuem com 7,5%, 5,2% e 2,7%, respectivamente (ANDRADE, L. C. d. *et al.*, 2019).

Outro fator relevante em relação ao risco de enchentes do lago está no tempo de retenção das águas que chegam dos rios. Ao investigar os níveis de poluição do lago, devido a capital carecer de um tratamento de água 100% efetivo, notou-se que grande parte da água do Guaíba fica retida por longos períodos, gerando baixa circulação e menor diluição de poluentes (ANDRADE, L. C. d. *et al.*, 2019). Além do agravante da qualidade da água, esse comportamento faz com que o fluxo dos rios que chegam ao lago, em caso de um aumento atípico do volume, desencadeie enchentes que demoram para escoar, prejudicando ainda mais a população das cidades banhadas pelo lago.

Na Figura 1, é mostrado alguns rios que fazem parte da bacia hidrográfica do Guaíba, e que influenciam diretamente no seu nível. À esquerda, percebe-se que um rio não é identificado, sendo este o Rio Jacuí, que contribui majoritariamente com o fluxo de água que aflui ao lago. Na Figura 2, fica evidente o motivo de seu maior impacto, dado sua extensão e largura maior que os demais rios.

Figura 1 – Rios que desemborciam e influenciam no nível do Lago Guaíba.



Fonte: (G1, 2024a).

Figura 2 – Rio Jacuí e seu desemboque no Lago Guaíba.



Fonte: (PORTO IMAGEM, 2009).

2.2 APRENDIZADO DE MÁQUINA

Desde que os computadores foram inventados, criou-se o questionamento da possibilidade de fazê-los pensar de modo semelhante ao ser humano. Por meio desse avanço, diversas áreas sofreriam grandes transformações, uma vez que a capacidade da máquina aprender e aprimorar o seu conhecimento sobre determinado assunto traria melhorias e uma maior performance na atividade desejada (CARBONELL; MICHALSKI; MITCHELL, 1983).

Embora os computadores ainda não alcancem o mesmo nível de aprendizado geral do ser humano, nos últimos anos, o ML (Machine Learning (Aprendizado de Máquina)) se tornou realidade, com aplicações em diversos setores relacionados ou não à tecnologia, agregando valor e conhecimento por meio de dados e informações antes tratados apenas por profissionais da área.

Esse conceito envolve a criação de sistemas que são capazes de aprender a partir de dados, identificando padrões e realizando previsões sem a necessidade de programação explícita. De acordo com (CARBONELL; MICHALSKI; MITCHELL, 1983), o principal objetivo do ML é construir algoritmos que permitam que os computadores adquiram conhecimento e melhorem sua performance de forma autônoma, baseando-se em experiências passadas.

2.2.1 Categorias de aprendizado de máquina

Os quatro principais tipos de ML são: supervisionado, não supervisionado, semi-supervisionado e reforço (SARAVANAN; SUJATHA, 2018). Estes tipos de ML são descritos a seguir:

- Supervisionado: envolve a utilização de dados rotulados, no qual o modelo é treinado com entradas e saídas conhecidas para fazer previsões sobre novos dados;
- Não supervisionado: lida com dados não rotulados, onde o sistema busca encontrar padrões ou agrupamentos nos dados;
- Semissupervisionado: combina elementos de ambos os métodos, utilizando uma pequena quantidade de dados rotulados e uma grande quantidade de dados não rotulados, sendo útil em cenários onde a rotulação de dados é cara ou complexa;
- Aprendizado por reforço: se baseia em um sistema de recompensas e punições, onde o sistema interage com o ambiente e aprende a otimizar suas ações para alcançar um objetivo a partir de *feedbacks* recebidos.

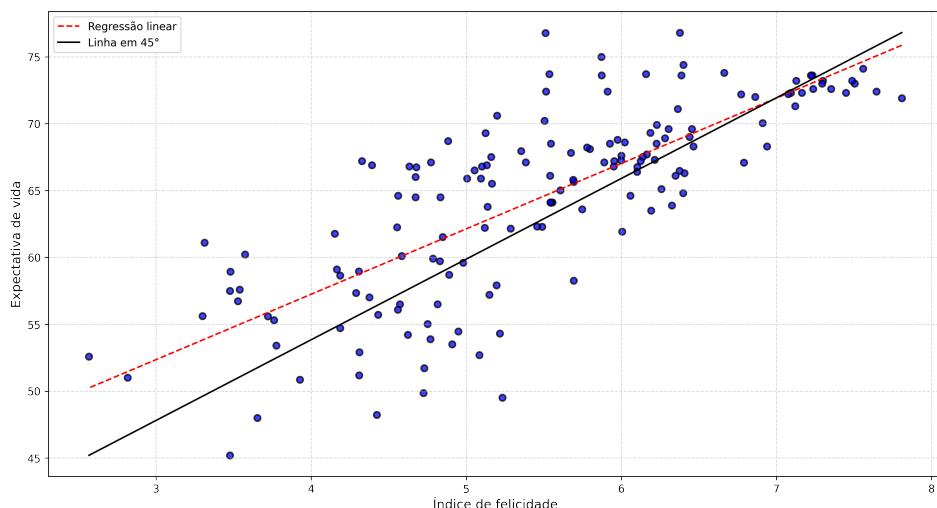
2.3 REGRESSÃO

A partir da necessidade de realizar previsões, visando compreender e estimar a dinâmica dos fenômenos estudados, a regressão se apresenta como uma ferramenta que busca modelar relações entre variáveis dependentes e independentes através de métodos estatísticos (SOTO, 2013).

Em uma equação linear, uma variável independente, comumente representada pela letra x , caracteriza uma grandeza que está sendo manipulada durante um experimento. Dado esse comportamento, a variável x não sofre influência de outras variáveis. A variável dependente, comumente representada pela letra y , caracteriza valores que estão diretamente associados à variável independente. Assim, de forma direta ou indireta, x exerce influência sobre y .

A Figura 3 ilustra um exemplo de regressão, mostrando a relação entre o índice de felicidade e a expectativa de vida em diversos países, conforme dados de (HELLIWELL *et al.*, 2020). Nesse contexto, o índice de felicidade é considerado a variável independente, enquanto a expectativa de vida é a variável dependente. Observando o gráfico, é possível perceber uma tendência de que países com maior índice de felicidade apresentam também uma expectativa de vida mais elevada. Assim, a regressão busca ajustar uma reta que, de forma aproximada, modela essa relação entre as variáveis, permitindo estimar valores de expectativa de vida a partir de valores do índice de felicidade. Tal reta pode ser representada pela equação linear mencionada anteriormente, e, caso disponível, a equação gerada pelos dados pode ser informada para descrever matematicamente a tendência observada na figura.

Figura 3 – Relação entre o índice de felicidade e expectativa de vida.



Fonte: (HELLIWELL *et al.*, 2020)

Embora uma inferência inicial permita constatar uma correlação entre as variáveis

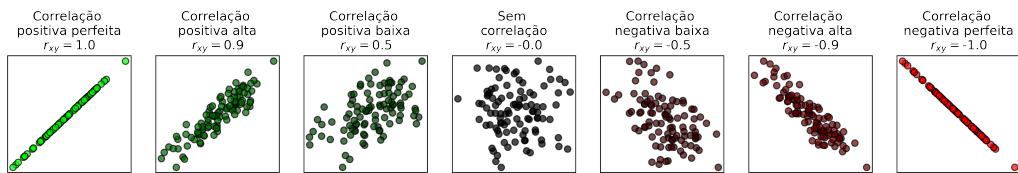
da equação, a criação de um modelo de previsão necessita de métodos que comprovem a correlação pressuposta. Para determinar as relações entre as variáveis dependentes e independentes de um sistema, coeficientes de correlação são calculados, gerando valores que medem e comprovam estatisticamente o grau de correspondência dos fatores estudados. Uma das métricas de correlação mais utilizadas é o coeficiente de Pearson, que mede a associação linear entre duas variáveis (KIRCH, 2008).

Esse coeficiente de correlação pode ser definido pela Equação (1), onde n é o total de amostras, \bar{x} e \bar{y} são as médias aritméticas de ambas as variáveis. Os valores do coeficiente de Pearson variam entre -1 e 1, de tal forma que quanto mais próximos desses extremos, melhor correlacionado estão as variáveis.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

A Figura 4 mostra alguns exemplos com gráficos de dispersão de variáveis com diferentes correlações.

Figura 4 – Diferentes correlações entre variáveis.



Fonte: (HELLIWELL *et al.*, 2020)

Quando o coeficiente de correlação indica uma alta correlação entre variáveis independentes (por exemplo, dados meteorológicos) e a variável dependente (nível do Lago Guaíba), métodos mais simples de regressão podem ser utilizados para estimar valores não presentes no conjunto de dados, aproveitando a relação estatística identificada. No entanto, em casos onde o coeficiente indica baixa correlação entre as variáveis ou alta multicolinearidade entre as preditoras, métodos como a regressão Ridge se mostram vantajosos, pois incorporam uma penalidade (regularização L2) que reduz o impacto de variáveis menos relevantes ou correlacionadas, permitindo previsões mais robustas sem depender exclusivamente da força da correlação linear.

2.4 REGRESSÃO LINEAR

A regressão linear é um tipo específico de regressão que modela a relação entre uma variável independente e uma variável dependente, conforme citado em 2.3. Amplamente utilizada em áreas como engenharia, ciências físicas, economia e ciências sociais, essa técnica assume que a relação entre as variáveis é linear, permitindo prever valores da

variável dependente com base em uma ou mais variáveis independentes (MONTGOMERY; PECK; VINING, 2012).

A aplicação da técnica é relevante devido à sua simplicidade e capacidade de fornecer previsões baseadas em uma fórmula matemática interpretável. Além disso, o método é base para implementações de algoritmos na área de ciência de dados como aprendizado de máquina, otimizando o processamento de dados complexos e viabilizando a criação de modelos de previsão (SERVICES, 2024).

O método de regressão linear é dividido em dois grupos, sendo eles: RLS (Regressão Linear Simples) e RLM (Regressão Linear Múltipla) (MONTGOMERY; PECK; VINING, 2012). A RLS tem como objetivo estabelecer uma relação entre duas variáveis através de uma função, cuja definição é dada por:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2)$$

Onde y é a variável dependente, x a variável independente, enquanto β_0 e β_1 são coeficientes calculados pela regressão, que representam o valor de y quando $x = 0$ e o grau de inclinação da reta, respectivamente.

A RLM, embora seja semelhante à RLS, possui múltiplas variáveis preditoras, sendo definida por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (3)$$

Na equação (3), y é a variável alvo, x_1 a x_k as variáveis regressoras, e β_0 permanece sendo o coeficiente de intercepto do eixo Y enquanto β_1 a β_n representam os coeficientes associados à n -ésima variável (SASSI *et al.*, 2012).

Em (2) e (3), nota-se a presença do erro estatístico representado por ε , que é a diferença entre o valor observado e o valor previsto pela equação de regressão. Esse erro é considerado aleatório e contabiliza a falha do modelo ao tentar se aproximar do comportamento denotado pelos dados amostrados (MONTGOMERY; PECK; VINING, 2012).

Para compreender o modelo de regressão linear sob suas suposições fundamentais, considera-se que a variável independente x (por exemplo, precipitação meteorológica) é conhecida e usada para prever a variável dependente y (como o nível do Lago Guaíba). Sob essas condições, todos os termos do lado direito da equação $y = \beta_0 + \beta_1 x + \varepsilon$ são conhecidos, exceto o erro ε , que determina as propriedades estatísticas de y . Assumindo que o erro ε tem média zero e variância constante σ^2 (MONTGOMERY; PECK; VINING, 2012), a resposta média para qualquer valor de x é dada por

$$E(y | x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x \quad (4)$$

e a variância é dada por:

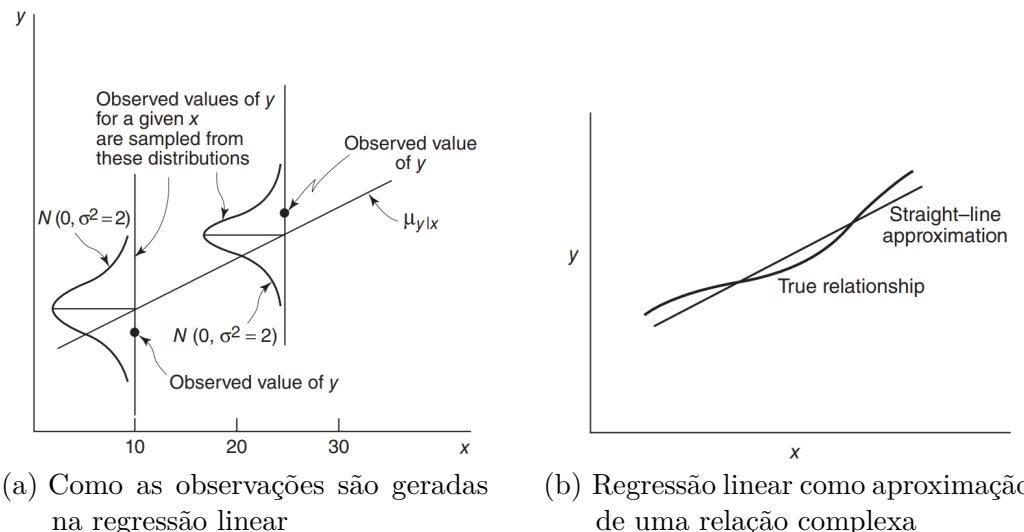
$$Var(y | x) = \sigma_{y|x}^2 = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \quad (5)$$

Desse modo, o modelo de regressão verdadeiro $\mu_{y|x} = \beta_0 + \beta_1 x$ representa uma linha de valores médios, ou seja, a altura da linha de regressão em qualquer valor de x corresponde ao valor esperado de y para aquele x .

Para exemplificar as suposições da regressão linear, considera-se um modelo ilustrado pela Figura 3a, onde a média condicional é $\mu_{y|x} = 3.5 + 2x$ e a variância do erro é $\sigma^2 = 2$. O erro ε segue uma distribuição normal, descrevendo a variação aleatória em torno da média. Como y é a soma de uma componente linear $\beta_0 + \beta_1 x$ (a média) e o erro ε , normalmente distribuído, y também segue uma distribuição normal. Por exemplo, para um valor específico da variável independente $x = 10$, y terá uma distribuição normal com média $\mu_{y|x} = 3.5 + 2(10) = 23.5$ e variância $\sigma^2 = 2$. Quanto menor a variância, mais próximos os pontos estarão da linha de regressão; uma variância maior resulta em maior dispersão em relação à linha de regressão (MONTGOMERY; PECK; VINING, 2012).

A maioria dos fenômenos nos quais se deseja obter a função que descreve o seu comportamento resulta em uma aproximação funcional através das variáveis de interesse. Essas relações funcionais frequentemente baseiam-se em teorias físicas, químicas ou de engenharia e ciências, ou seja, no conhecimento do mecanismo subjacente. Na Figura 5b, é mostrada uma relação entre as variáveis x e y relativamente complexa, mas que pode ser aproximada por uma equação de regressão linear, com um erro relativamente baixo.

Figura 5 – Interpretação de uma regressão linear

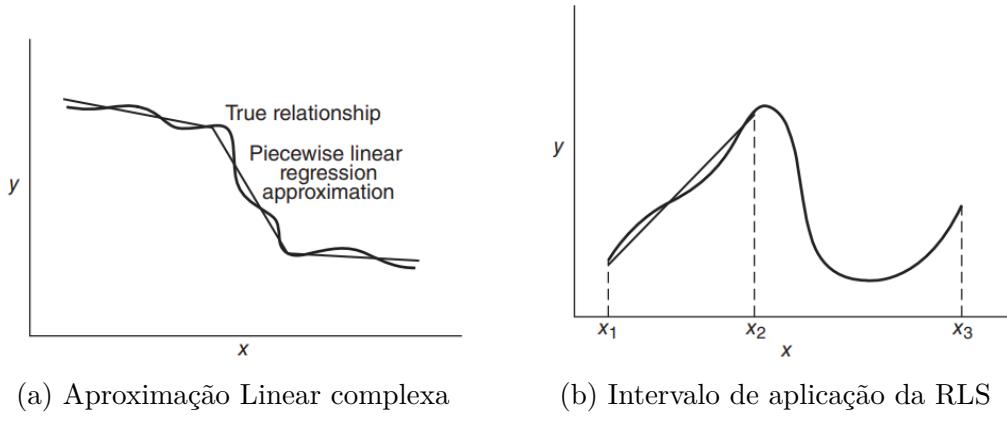


Fonte: (MONTGOMERY; PECK; VINING, 2012)

Contudo, em alguns casos, quando a dinâmica do modelo a ser estimada passa a ter um grau de complexidade maior, como é o caso da Figura 6a, utilizar uma RLS pode implicar em erros que extrapolam a tolerância exigida no estudo. Nesses cenários, utilizar

uma função de regressão linear em intervalos específicos, ou seja, uma RLM, se torna uma alternativa plausível, tendo em vista que, para intervalos menores onde a dinâmica do fenômeno é mais linear, a regressão apresenta um erro menor, como mostra a Figura 10f.

Figura 6 – Situações de inadequação da RLS



Fonte: (MONTGOMERY; PECK; VINING, 2012)

Partindo desses conceitos, para implementação de modelos de regressão linear e múltipla, o método dos Mínimos Quadrados Ordinários se apresenta como uma abordagem para estimar a melhor regressão dos pontos observados, encontrando uma reta com o menor erro entre as amostras e os valores da função estudada.

2.5 MÉTODO DOS QUADRADOS ORDINÁRIOS

O método MQO (Mínimos Quadrados Ordinários) atua como uma ferramenta estatística, visando estimar a relação entre uma variável dependente e uma ou mais variáveis independentes (ALKAMA *et al.*, 2020), permitindo encontrar os coeficientes desejados para o funcionamento do modelo.

Para obter uma regressão que se aproxima da dinâmica analisada, o método visa minimizar a RSS (Residual Sum of Squares (Soma dos Quadrados dos Resíduos)), denotado por:

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (6)$$

para os casos de RLS, ou seja, quando há apenas uma variável independente. Para o caso de RLM, a equação é dada por:

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (7)$$

onde:

- y_i é uma variável aleatória e representa o valor da variável resposta (variável dependente) na i-ésima observação
- x_{ij} representa o valor da variável explicativa (variável independente, variável regressora) na i-ésima observação. Nota-se que podem existir múltiplas variáveis independentes para uma variável independente;
- β_0 e β_1 são os parâmetros do modelo que serão estimados, e que definem a reta de regressão

Para minimizar a SSR em um caso de RLS, por exemplo, são calculadas as derivadas parciais de β_0 e β_1 , igualando ambas a zero.

Derivada em relação à β_0 :

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (8)$$

simplificando:

$$\begin{aligned} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) &= 0 \\ \sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i &= 0 \end{aligned}$$

divindindo por n :

$$\begin{aligned} \frac{\sum_{i=1}^n y_i}{n} - \beta_0 - \beta_1 \frac{\sum_{i=1}^n x_i}{n} &= 0 \\ \bar{y} - \beta_0 - \beta_1 \bar{x} &= 0 \end{aligned}$$

onde \bar{y} e \bar{x} são as médias amostrais de y e x , respectivamente. Assim, a equação pode ser reescrita como:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (9)$$

Derivada em relação à β_1 :

$$\frac{\partial SSR}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (10)$$

substituindo β_0 na equação:

$$\begin{aligned}
 \sum_{i=1}^n x_i(y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i) &= 0 \\
 \sum_{i=1}^n x_i(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) &= 0 \\
 \sum_{i=1}^n x_i(y_i - \bar{y}) + \sum_{i=1}^n x_i(\beta_1 \bar{x} - \beta_1 x_i) &= 0 \\
 \sum_{i=1}^n x_i(y_i - \bar{y}) - \beta_1 \sum_{i=1}^n x_i(x_i - \bar{x}) &= 0
 \end{aligned}$$

sabendo que:

$$\sum_{i=1}^n x_i(x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2$$

e

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

portanto:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 \\
 \beta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{11}
 \end{aligned}$$

Desse modo, a partir de um problema onde uma ou mais entradas geram amostras que resultam em uma saída, torna-se possível estimar uma função que melhor representa seu comportamento, minimizando ao máximo o valor da soma residual dos quadrados entre os pontos amostrais e a curva do modelo.

A Tabela 1 apresenta um exemplo de dados amostrais, onde a variável independente é representada pela primeira coluna e a variável dependente pela segunda coluna. A partir desses dados, é possível aplicar o método dos mínimos quadrados para encontrar os coeficientes que melhor se ajustam à reta de regressão linear.

Ao aplicar o método para resolver um problema, como é o caso da Tabela 1, todos os pontos amostrados são utilizados para encontrar a reta que melhor se ajusta aos dados. Contudo, visando mostrar como a dinâmica de regressão utilizando MQO funciona, nos passos seguintes, as amostras são consideradas de forma cumulativa, alterando a cada iteração os valores de β_0 e β_1 , até que a reta de regressão linear se ajuste aos dados amostrais.

Variável Independente	Variável Dependente
0,38	6,98
0,41	4,05
0,44	5,52
0,59	6,93
0,98	6,57
1,04	6,41
1,22	8,27
1,53	6,93
1,74	8,89
1,84	9,31

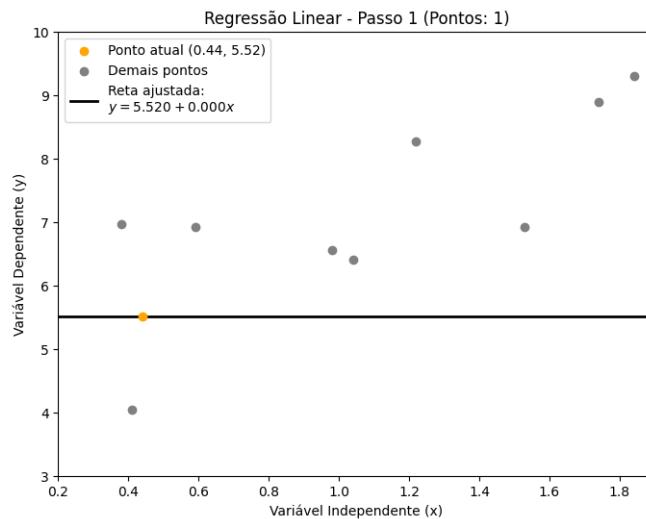
Tabela 1 – Tabela de Valores Ordenados: Variável Independente vs. Variável Dependente

Passo 1: Apenas o ponto (0.44, 5.52)

Com um único ponto, a reta passa exatamente por sobre o mesmo, porém o método MQO exige pelo menos dois pontos para definir uma inclinação. Assim, assume-se uma reta horizontal ao usar o ponto como base inicial.

$$\beta_0 = 5.52, \quad \beta_1 = 0$$

Figura 7 – Passo 1 da regressão linear pelo método MQO.



Fonte: Autor.

Passo 2: Adiciona-se (1.74, 8.89)

n=2

$$\bar{x} = \frac{0.44 + 1.74}{2} = 1.09, \quad \bar{y} = \frac{5.52 + 8.89}{2} = 7.205$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (0.44 - 1.09)(5.52 - 7.205) + (1.74 - 1.09)(8.89 - 7.205) = (-0.65)(-1.685) + (0.65)(1.685) = 1.09525 + 1.09525 = 2.1905$$

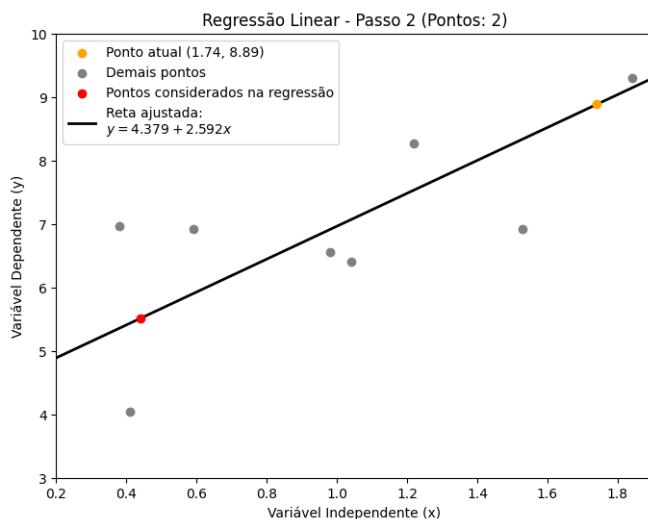
$$\sum(x_i - \bar{x})^2 = (0.44 - 1.09)^2 + (1.74 - 1.09)^2 = 0.4225 + 0.4225 = 0.845$$

$$\beta_0 = \frac{2.1905}{0.845} \approx 2.5923$$

$$\beta_1 = 7.205 - 2.5923 \cdot 1.09 \approx 7.205 - 2.8255 = 4.3795$$

$$\hat{y} = 4.3795 + 2.5923x$$

Figura 8 – Passo 2 da regressão linear pelo método MQO.



Fonte: Autor.

Passo 3: Adiciona $(0.41, 4.05)$

$n = 3$

$$\bar{x} = \frac{0.44 + 1.74 + 0.41}{3} = 0.8633, \quad \bar{y} = \frac{5.52 + 8.89 + 4.05}{3} = 6.1533$$

$$\begin{aligned} \sum(x_i - \bar{x})(y_i - \bar{y}) &= \\ (0.44 - 0.8633)(5.52 - 6.1533) &+ \\ (1.74 - 0.8633)(8.89 - 6.1533) &+ \\ (0.41 - 0.8633)(4.05 - 6.1533) &\approx \end{aligned}$$

$$0.268 + 2.399 + 0.953 = 3.62$$

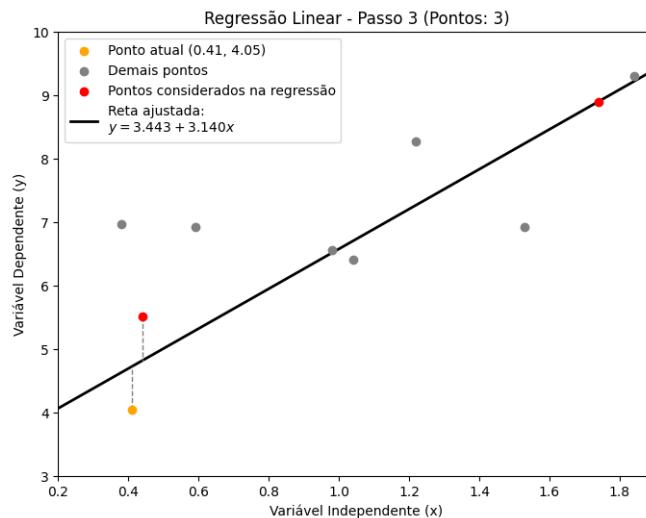
$$\begin{aligned} \sum(x_i - \bar{x})^2 &= (0.44 - 0.8633)^2 + (1.74 - 0.8633)^2 + (0.41 - 0.8633)^2 \\ &\approx 0.179 + 0.769 + 0.205 = 1.153 \end{aligned}$$

$$\beta_0 = \frac{3.62}{1.153} \approx 3.1402$$

$$\beta_1 = 6.1533 - 3.1402 \cdot 0.8633 \approx 6.1533 - 2.711 = 3.4423$$

$$\hat{y} = 3.4423 + 3.1402x$$

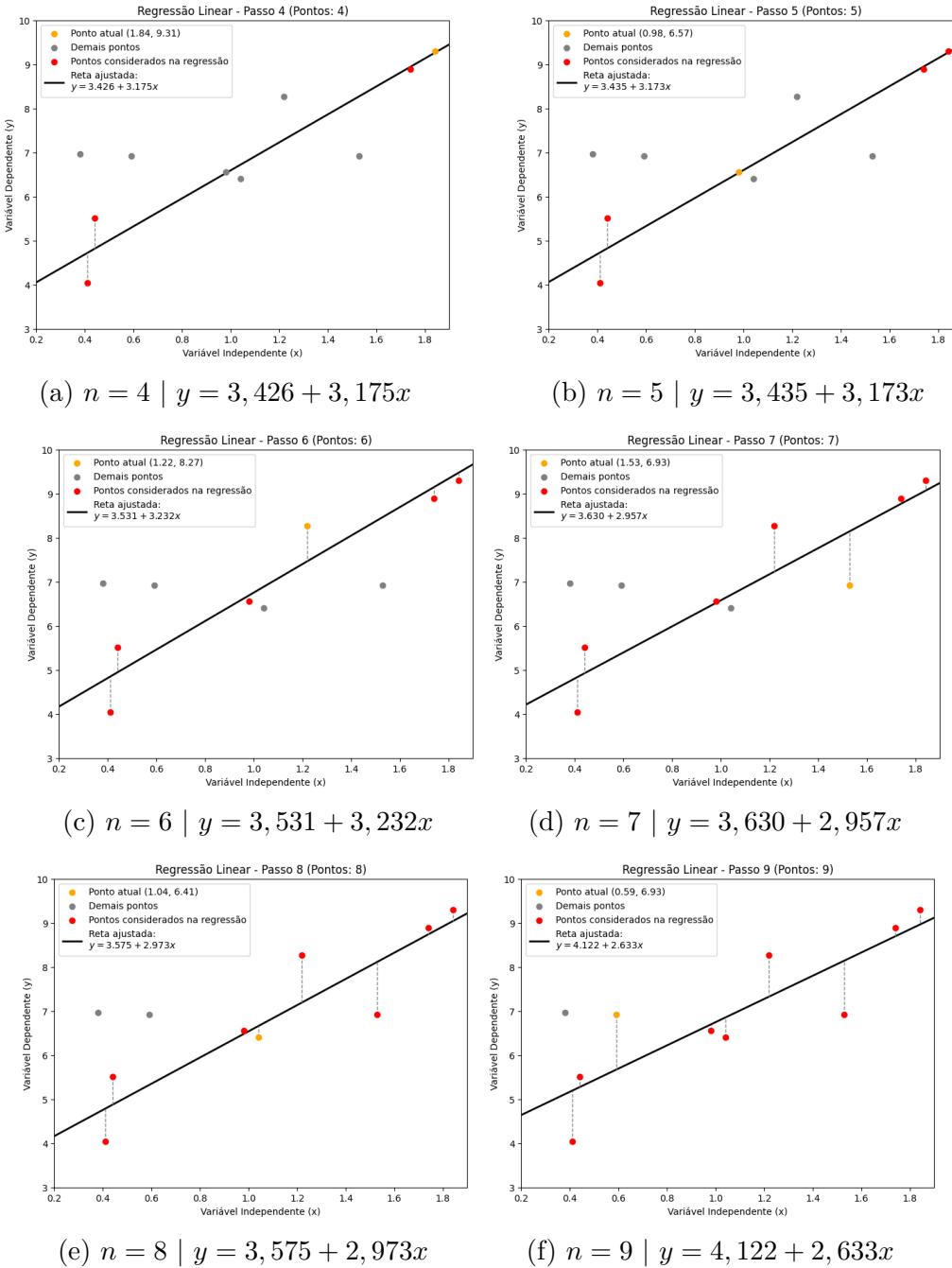
Figura 9 – Passo 3 da regressão linear pelo método MQO.



Fonte: Autor.

Para os demais passos, o mesmo cálculo é realizado, onde a média amostral e os coeficientes são recalculados a cada iteração, conforme os pontos são adicionados.

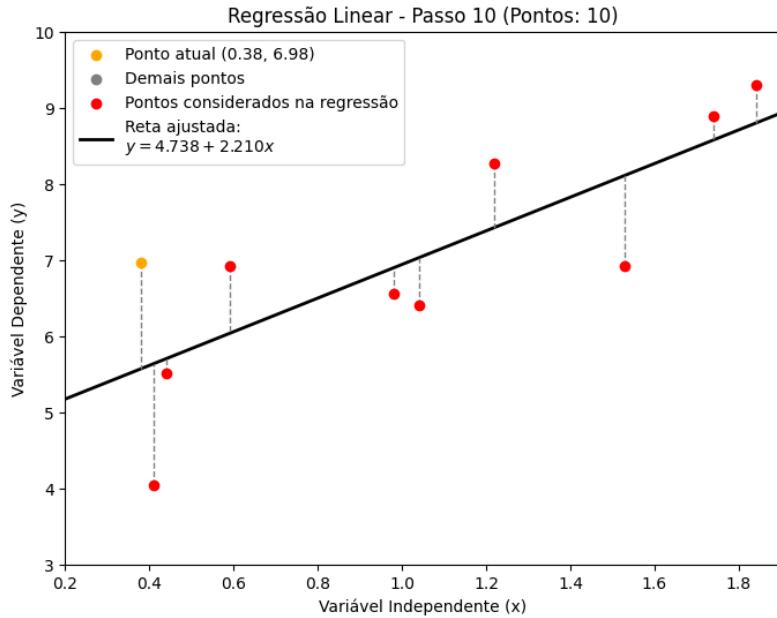
Figura 10 – Iterações da aplicação do método MQO



Fonte: Autor.

Assim, a cada ponto selecionado da amostra, é calculada a derivada parcial da função estimada, determinando novos coeficiente que melhor descrevem a reta entre as amostras. Dado que não é possível estimar uma reta que passe sobre todos os pontos amostrados, os resíduos representados pelas linhas tracejadas na Figura 11 são definidos de tal modo que o somatório dos seus quadrados seja o menor possível.

Figura 11 – Passo 10 da regressão linear pelo método MQO.



Fonte: Autor.

2.6 MODELO RIDGE

A Regressão Ridge é uma técnica de regularização estatística amplamente utilizada em modelos de regressão linear para abordar problemas de sobreajuste (ou *overfitting*, quando o modelo se ajusta demais aos dados de treinamento, perdendo a capacidade de prever dados fora da base passada para o modelo) e multicolinearidade entre variáveis preditoras (MCDONALD, 2009). Essa técnica, também conhecida como regularização L2, é particularmente eficaz em cenários onde o número de variáveis preditoras é grande ou quando essas variáveis apresentam alta correlação, o que pode levar a estimativas de coeficientes instáveis e de baixa generalização.

Conforme visto em 2.4, na regressão linear tradicional, o objetivo é minimizar o valor de RSS, que mede a diferença entre os valores observados e os valores previstos pelo modelo. A função de perda é dada por:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

onde y_i são os valores observados, \hat{y}_i são os valores previstos, e n é o número de observações. No entanto, em cenários com multicolinearidade (alta correlação entre variáveis preditoras) ou um grande número de preditores, o modelo pode se ajustar excessivamente aos dados de treinamento, resultando em alta variância e baixa performance em dados não vistos. A

Regressão Ridge resolve esse problema ao adicionar um termo de penalidade à função de perda, proporcional à soma dos quadrados dos coeficientes de regressão.

A função objetivo da Regressão Ridge é:

$$RSS_{L2} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p \beta_j^2 \quad (13)$$

No contexto da biblioteca *Scikit-learn* em Python, a classe Ridge implementa a Regressão Ridge. O parâmetro *alpha* define a força da regularização: valores maiores de α resultam em coeficientes mais próximos de zero, enquanto valores menores permitem que o modelo se aproxime da regressão linear ordinária. A escolha adequada de α é crucial para evitar tanto o *overfitting* quanto o *underfitting* (quando o modelo é muito simples para capturar os padrões dos dados). (SCIKIT-LEARN DEVELOPERS, 2025b).

Além de impactar a função RSS, a Regressão Ridge altera o processo de busca pelos coeficientes β que definem a reta (ou hiperplano, no caso de múltiplas variáveis) que melhor representa a regressão.

Na regressão linear tradicional, os coeficientes β são encontrados minimizando-se apenas a soma dos quadrados dos resíduos, resultando em uma solução exata obtida pela equação normal:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (14)$$

onde \mathbf{X} é a matriz de variáveis independentes e \mathbf{y} é o vetor de variáveis dependentes (MONTGOMERY; PECK; VINING, 2012).

Entretanto, quando há multicolinearidade ou grande número de variáveis, a matriz $\mathbf{X}^\top \mathbf{X}$ pode se tornar quase singular, causando coeficientes instáveis e de grande magnitude. A Regressão Ridge resolve isso ao adicionar o termo de penalização $\alpha \sum_{j=1}^p \beta_j^2$, modificando a equação normal para:

$$\hat{\beta}_{ridge} = (\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \quad (15)$$

(SCIKIT-LEARN DEVELOPERS, 2025b).

Geometricamente, a regularização imposta pela Ridge faz com que a busca pela solução deixe de se restringir apenas à reta ou hiperplano definido pelo menor erro de previsão. Em vez disso, a solução passa a ser buscada dentro de uma região esférica ao redor da origem, delimitada pela penalização L2.

Com isso, mesmo que múltiplas combinações de coeficientes possam gerar ajustes semelhantes aos dados (especialmente em casos de multicolinearidade), a regressão *Ridge* prefere aquelas soluções onde os coeficientes são menores, evitando oscilações extremas.

Portanto, o modelo *Ridge* não apenas minimiza o erro de ajuste aos dados da RSS, mas também impõe um “encolhimento” dos coeficientes em direção ao zero, resultando

em retas ou hiperplanos mais estáveis, com menor variância e melhor capacidade de generalização para dados não vistos ao qual se deseja ser previsto (MCDONALD, 2009).

3 DESENVOLVIMENTO

Este capítulo descreve as etapas fundamentais para a preparação dos dados utilizados no treinamento do modelo de previsão do nível do Lago Guaíba, utilizando dados meteorológicos e a técnica de Regressão Ridge. A preparação adequada dos dados é essencial para garantir a qualidade e a confiabilidade das previsões, especialmente em um contexto de alta variabilidade climática. O capítulo está estruturado em quatro seções principais: a Seção 3.1 detalha a obtenção de dados meteorológicos e dos níveis dos rios, estabelecendo a base de informações para o modelo; a Seção 3.2 aborda a concatenação e formatação dos dados para garantir consistência; a Seção 3.3 descreve as técnicas para tratar valores ausentes e *outliers* (valores discrepantes, que destoam do comportamento típico que se deseja considerar nos dados coletados), assegurando a integridade das informações; e a Seção 3.4 explica a seleção de variáveis relevantes, otimizando a eficiência do modelo.

Na sequência, a Seção 3.5 apresenta os métodos de transformação e normalização dos dados, a fim de adequar as diferentes escalas das variáveis e garantir que o modelo de aprendizado de máquina opere de forma eficiente. A Seção 3.6 discute a divisão do conjunto de dados em subconjuntos de treinamento e teste, garantindo a capacidade de generalização do modelo. Por fim, a Seção 3.7 detalha a aplicação prática do modelo de Regressão Ridge, incluindo a configuração dos hiperparâmetros, o processo de treinamento e a avaliação dos resultados obtidos por meio do script desenvolvido na linguagem de programação *Python*. Cada seção contribui para a construção de um conjunto de dados robusto, permitindo que o modelo de Regressão *Ridge* capture padrões sazonais e climáticos com maior precisão, fundamental para a previsão eficaz do nível do rio e a mitigação de impactos de enchentes.

3.1 COLETA DE DADOS

Considerando a premissa do trabalho, em que a previsão do nível do rio será dada a partir de dados meteorológicos da cidade de Porto Alegre, junto aos dados de monitoramento do nível dos rios que constituem a bacia do Guaíba, duas fontes de dados foram utilizadas. Para os dados meteorológicos, o portal do INMET (Instituto Nacional de Meteorologia)¹ foi utilizado, onde foram coletadas as informações de temperatura, umidade relativa do ar, precipitação e velocidade do vento. Já os dados de monitoramento dos rios foram coletados na página do SEMA-RS (Secretaria do Meio Ambiente e Infraestrutura do Rio Grande do Sul)² da internet (Sala de situação), onde foram coletados os dados de nível do Lago Guaíba, Caí, Jacuí, Sinos e Gravataí. Os dados meteorológicos foram coletados em formato CSV (Comma Separated Values (Valores Separados por Vírgula)), com arquivos separados por ano de monitoramento, com frequência horária. Os dados dos

¹ <https://portal.inmet.gov.br/>

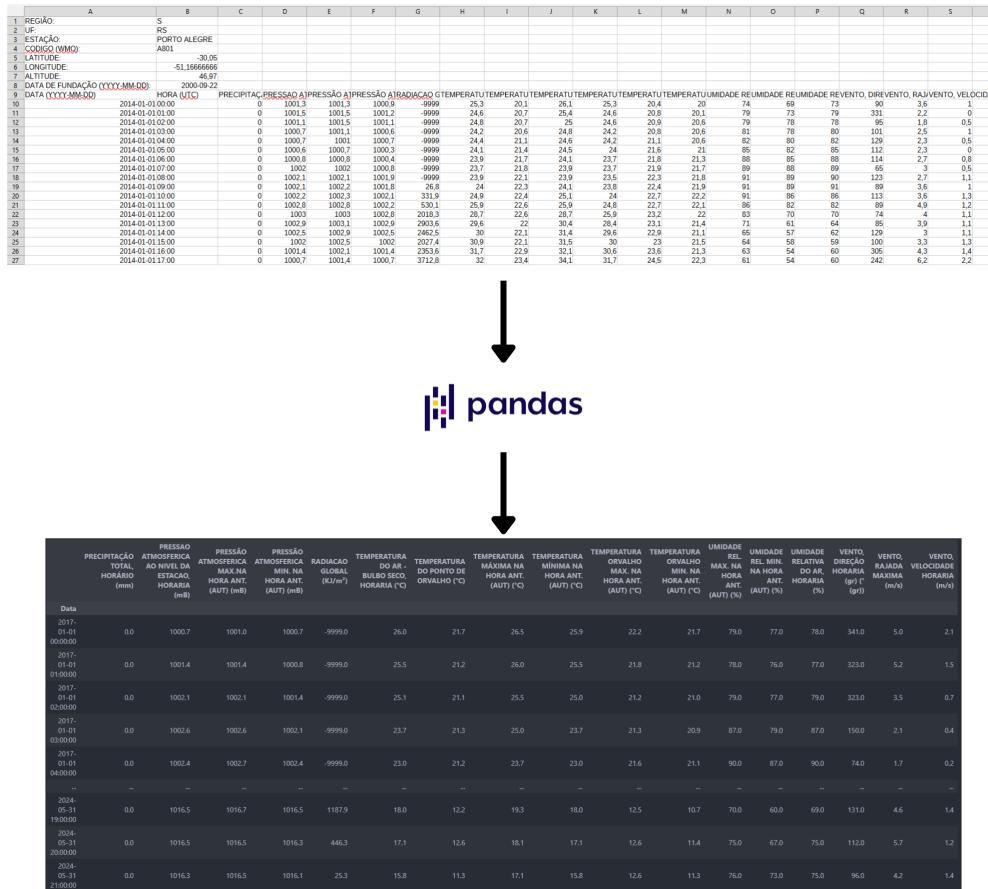
² <https://www.saladesituacao.rs.gov.br/dados>

níveis dos rios foram coletados na forma de uma planilha eletrônica em formato .xlsx, com histórico completo de amostragem das informações em frequência de 15 minutos.

3.2 PRÉ PROCESSAMENTO DOS DADOS

Antes de seguir para a etapa de limpeza dos dados mostrado na Figura ??, os dados coletados necessitam de um pré-processamento específico para cada uma das fontes utilizadas. Para os dados meteorológicos, devido às informações estarem separadas por ano de monitoramento, foi necessário concatenar os arquivos de cada ano em um único arquivo. Isso foi feito utilizando a função *concat* da biblioteca *Pandas*, removendo o cabeçalho de informações geográficas da estação, ilustrado nas linhas 1 a 8 da Figura 12. Além disso, foi necessário combinar as duas primeiras colunas e converter o formato de data e hora para o padrão *datetime*, utilizando a função *to_datetime* da mesma biblioteca.

Figura 12 – Dados meteorológicos.



Fonte: Autor.

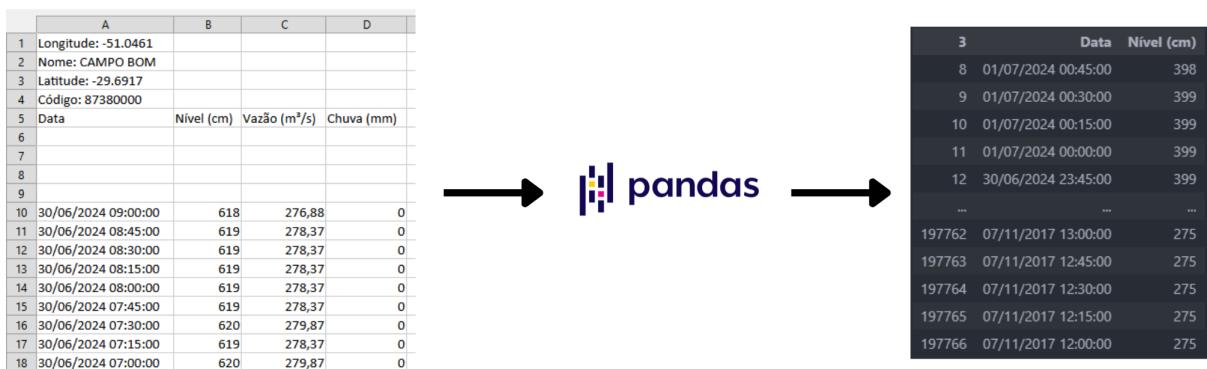
Analisando as colunas e os seus respectivos tipos de dados, tem-se a seguinte tabela:

Coluna	Tipo
Precipitação Total (mm)	float64
Pressão Atmosférica ao Nível da Estação (mB)	float64
Pressão Atmosférica Máx. na Hora Anterior (mB)	float64
Pressão Atmosférica Mín. na Hora Anterior (mB)	float64
Radiação Global (kJ/m ²)	float64
Temperatura do Ar - Bulbo Seco (°C)	float64
Temperatura do Ponto de Orvalho (°C)	float64
Temperatura Máxima na Hora Anterior (°C)	float64
Temperatura Mínima na Hora Anterior (°C)	float64
Temperatura Orvalho Máx. na Hora Anterior (°C)	float64
Temperatura Orvalho Mín. na Hora Anterior (°C)	float64
Umidade Relativa Máx. na Hora Anterior (%)	float64
Umidade Relativa Mín. na Hora Anterior (%)	float64
Umidade Relativa do Ar (%)	float64
Vento - Direção Horária (° (gr))	float64
Vento - Rajada Máxima (m/s)	float64
Vento - Velocidade Horária (m/s)	float64

Tabela 2 – Tabela de tipos de dados da base de informações meteorológicas.

Para os dados dos níveis dos rios, também foi necessário remover o cabeçalho com dados geográficos da estação de monitoramento, como mostra a Figura 13, junto da conversão do formato de data e hora para o padrão *datetime* da primeira coluna.

Figura 13 – Dados do nível do rio.



Fonte: Autor.

3.3 LIMPEZA DOS DADOS

Após a coleta dos dados, o próximo passo é a limpeza dessas informações, que consiste em remover dados duplicados, corrigir erros de formatação e lidar com valores

ausentes. Existem diferentes abordagens para tratar valores inconsistentes no dados coletados. Uma abordagem comum, principalmente em casos onde não há uma linearidade ou tendência clara, é o preenchimento com zero (COUSINEAU; CHARTIER, 2010), como mostra a Figura 14, ou com a média dos dados na coluna a ser limpa.

Figura 14 – Limpeza dos dados coletados.

	col1	col2	col3	col4	col5
1	2	5.0	3.0	6	?
2	9	?	9.0	0	7.0
3	19	17.0	?	9	?

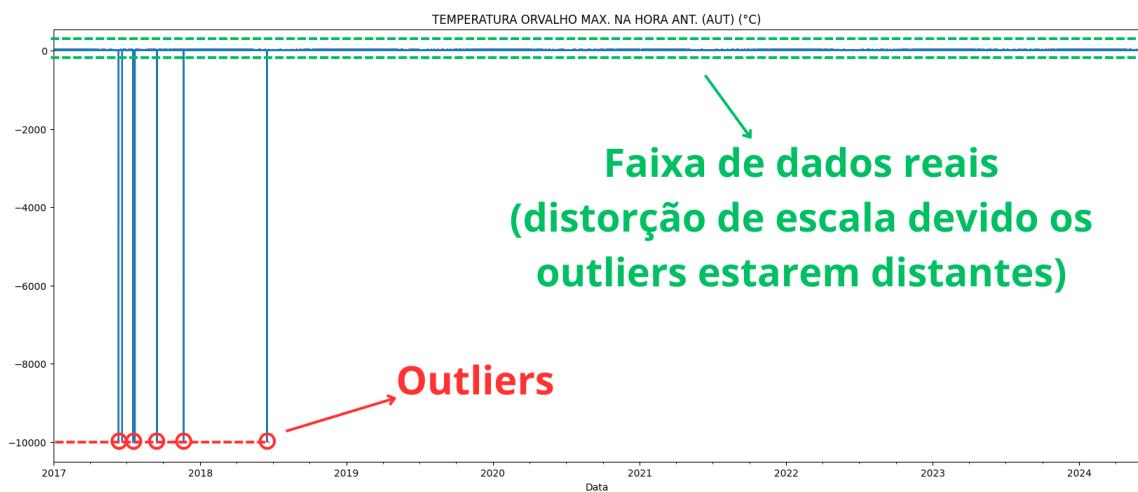
	col1	col2	col3	col4	col5
1	2	5.0	3.0	6	0.0
2	9	0.0	9.0	0	7.0
3	19	17.0	0.0	9	0.0

Fonte: (AI, 2023)

Nessa etapa, na base de dados meteorológicos, optou-se por preencher os dados ausentes, representados por “-”, e os dados com valor igual a “-9999” por zero, já que dados meteorológicos tendem a ser mais voláteis e não apresentam uma tendência clara. Desse modo, a abordagem de aproximação linear é descartada, a fim de não comprometer a análise do modelo de previsão.

Tomando como exemplo a coluna *TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT)* ($^{\circ}\text{C}$), na Figura 15, observa-se os valores ausentes representados por “-” e “-9999”, evidenciados na Figura pelos pontos distantes do restante dos dados.

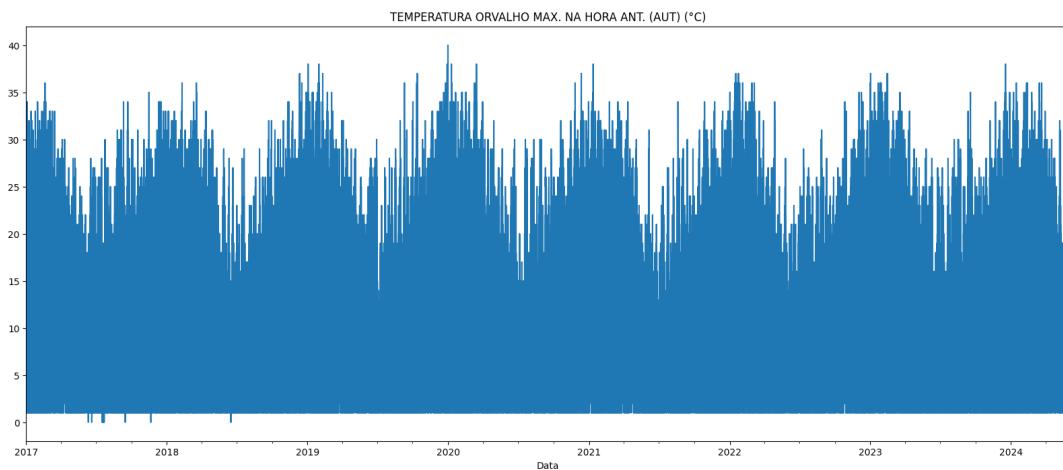
Figura 15 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior ($^{\circ}\text{C}$) não tratado.



Fonte: Autor.

Após a substituição dos valores ausentes por zero, o gráfico da mesma coluna, mostrado na Figura 16, apresenta uma distribuição mais uniforme, sem os picos de dados ausentes.

Figura 16 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior ($^{\circ}\text{C}$) tratado.

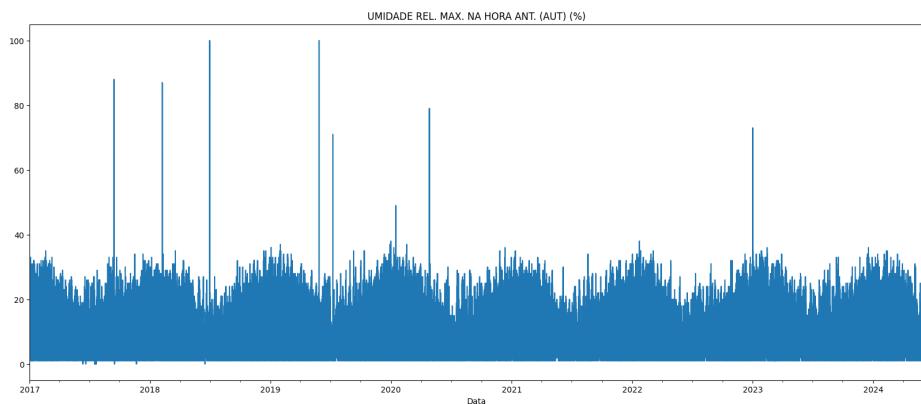


Fonte: Autor.

Em casos onde os dados apresentam *outliers* (dados que se distanciam significativamente do restante da coluna analisada), foi aplicado o método IQR (Interquartile Range (Intervalo Interquartil)) para identificar e remover esses valores. A partir da diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) de um conjunto de dados, multiplicado por um fator de tolerância, o método determina limites superiores e inferiores para o conjunto de dados analisado, removendo ou substituindo os valores que estão fora dessa faixa, para manter a consistência das informações.

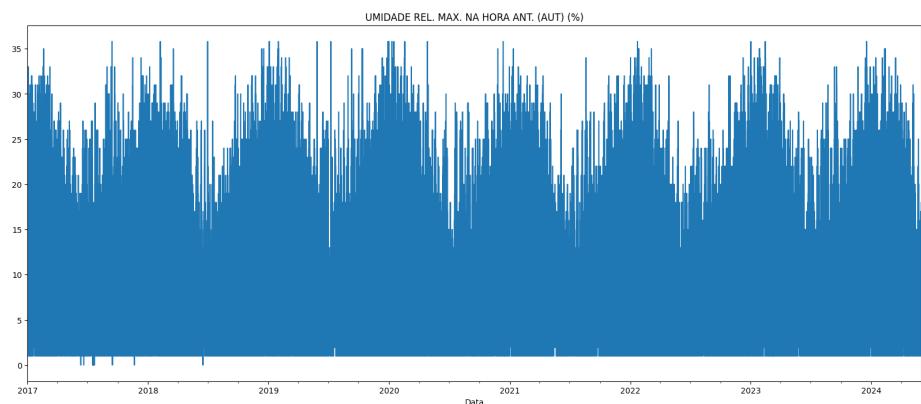
Por exemplo, na coluna *UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)*, as Figuras 17 e 18 mostram a diferença entre os dados após a aplicação do método IQR, com um fator de tolerância de 1.2 entre os interquartis Q1 e Q3. Para esta coluna, o cálculo do IQR resultou em um intervalo de -11,8% a 35,8%, limitando os *outliers* superiores vistos no gráfico para o valor máximo calculado.

Figura 17 – Gráfico de Umidade Relativa Máxima na Hora Anterior (%) sem tratamento.



Fonte: Autor.

Figura 18 – Gráfico de Umidade Relativa Máxima na Hora Anterior (%) com tratamento.



Fonte: Autor.

Na base de dados dos níveis dos rios, embora a dinâmica que representa os respetivos comportamentos não seja linear, o uso do conceito de aproximação linear permite preencher os dados ausentes por meio da interpolação entre os valores que cercam as células sem informação em determinado período. Essa abordagem é válida, uma vez que a variação do nível do rio entre dois pontos de amostragem próximos tende a ser linear, já que o nível do rio não apresenta oscilações bruscas em curtos períodos de tempo.

Para realizar a interpolação dos valores ausentes na base de dados dos níveis dos rios, utilizou-se a função *interpolate* da biblioteca Pandas com o argumento *method = 'linear'*. Esse método preenche os valores ausentes calculando uma interpolação linear entre os dois dados adjacentes mais próximos, ou seja, o valor ausente é estimado como uma média ponderada dos valores imediatamente anteriores e posteriores no conjunto de dados, considerando a distância temporal entre eles. Especificamente, a interpolação linear utiliza

apenas os dois pontos adjacentes (um antes e um depois do valor ausente) para determinar o valor interpolado, assumindo uma variação linear entre esses pontos.

Ainda, observou-se nas bases dos níveis dos rios a ocorrência de valores ausentes no início ou no final do período de amostragem, inviabilizando a interpolação. Para esses casos, o preenchimento foi feito a partir da repetição do primeiro ou do último valor disponível, respectivamente. É importante ressaltar que tal limpeza só pode ser feita se o preenchimento ocorra em um intervalo de tempo curto, visando não afetar a análise feita pelo modelo de previsão.

3.4 REDUÇÃO DOS DADOS

A etapa de redução dos dados visa eliminar informações redundantes ou irrelevantes, mantendo apenas os dados que contribuem para a análise e treinamento do modelo de previsão. Considerando que serão utilizados dados de monitoramento do nível de 4 rios para a previsão do nível do Lago Guaíba, os dados meteorológicos coletados precisam ser filtrados de modo a garantir tanto que as informações sejam relevantes para a previsão, quanto que os dados de monitoramento dos rios estejam alinhados com os dados meteorológicos, evitando assim a inclusão de dados desnecessários no treinamento.

Voltando para a Tabela 2, nota-se a presença de 17 colunas, das quais:

- 6 colunas referem-se a dados de temperatura;
- 3 colunas referem-se a dados de pressão atmosférica;
- 3 colunas referem-se a dados do comportamento do vento;
- 3 colunas referem-se a dados de umidade relativa do ar;
- 1 coluna refere-se a dados de radiação solar;
- 1 coluna refere-se a dados de precipitação.

Desse modo, para o primeiro passo de redução da base, foram mantidas apenas uma coluna de cada tipo de medição meteorológica, filtrando 6 colunas no total, conforme a Tabela 3.

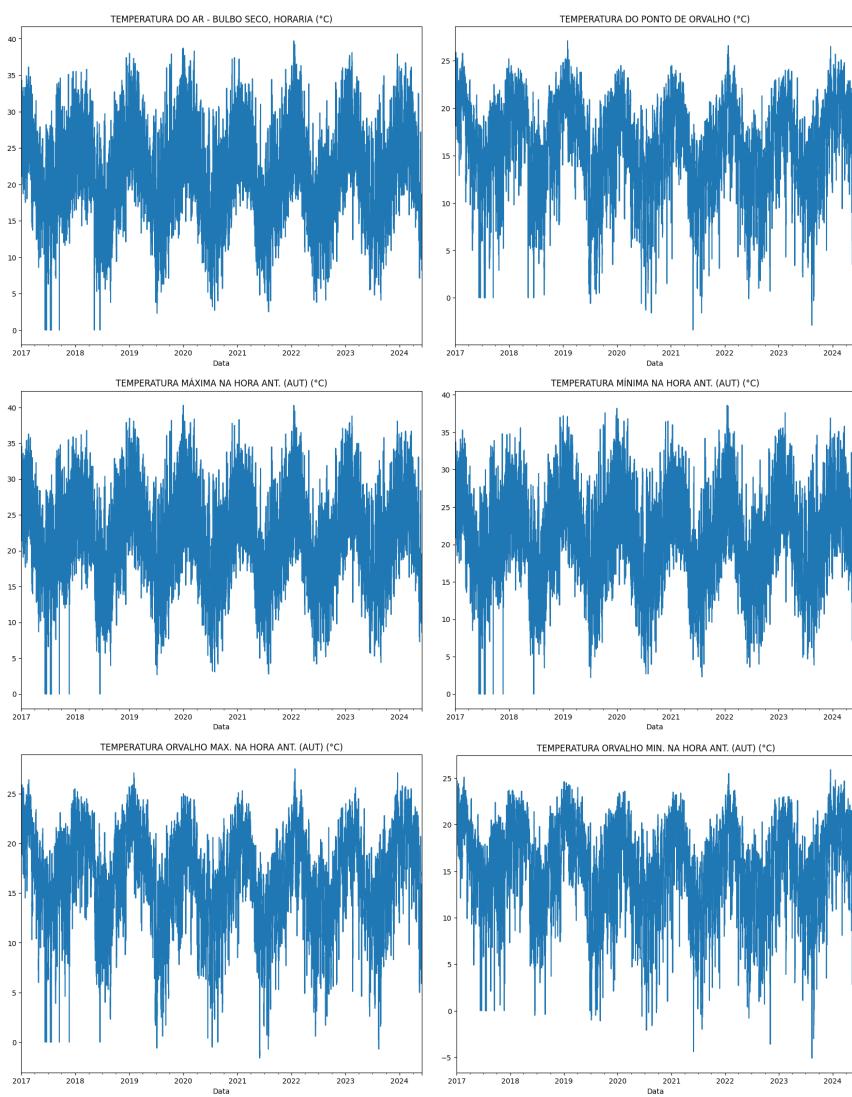
Coluna	Unidade de medida
Temperatura do Ar - Bulbo seco	(°C)
Pressão Atmosférica ao Nível da Estação	(mB)
Vento - Velocidade Horária	(m/s)
Umidade Relativa do Ar	(%)
Radiação Global	(kJ/m ²)
Precipitação Total	(mm)

Tabela 3 – Tabela de dados meteorológicos reduzidos - 1^a Filtragem

Para a escolha das colunas a serem mantidas, foram levados em consideração os seguintes critérios:

- As colunas *Temperatura do Ar - Bulbo seco*, *Pressão Atmosférica ao Nível da Estação* e *Umidade Relativa do Ar* foram escolhidas por serem medidas diretas dos respectivos fenômenos atmosféricos, enquanto as demais colunas referem-se a medidas derivadas ou não diretamente observáveis, que não são tão relevantes para a previsão do nível do rio. Além disso, observou-se que, para essas unidades de medida derivadas, o comportamento dos dados era semelhante, descartando a necessidade de manter mais de uma coluna para cada tipo de medição, como mostra a Figura 19 em relação às medidas de temperatura.

Figura 19 – Comparativo de gráficos de temperatura em diferentes tipos de medição



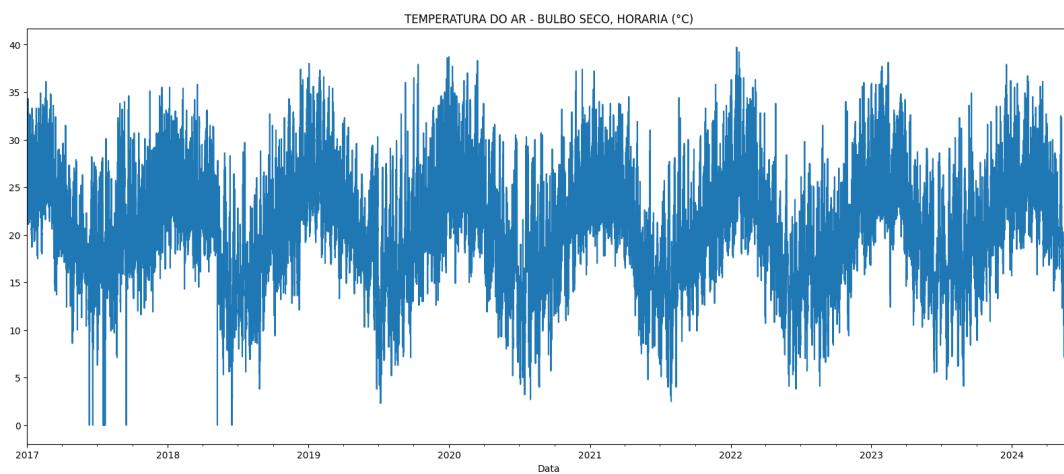
Fonte: Autor.

- A coluna *Precipitação Total* foi mantida por ser um dos principais fatores que influenciam o nível do rio.

- As colunas *Radiação Global* e *Umidade Relativa do Ar* foram mantidas por serem fatores importantes para a evaporação da água, que também influenciam, mesmo que de forma indireta, no nível do rio.

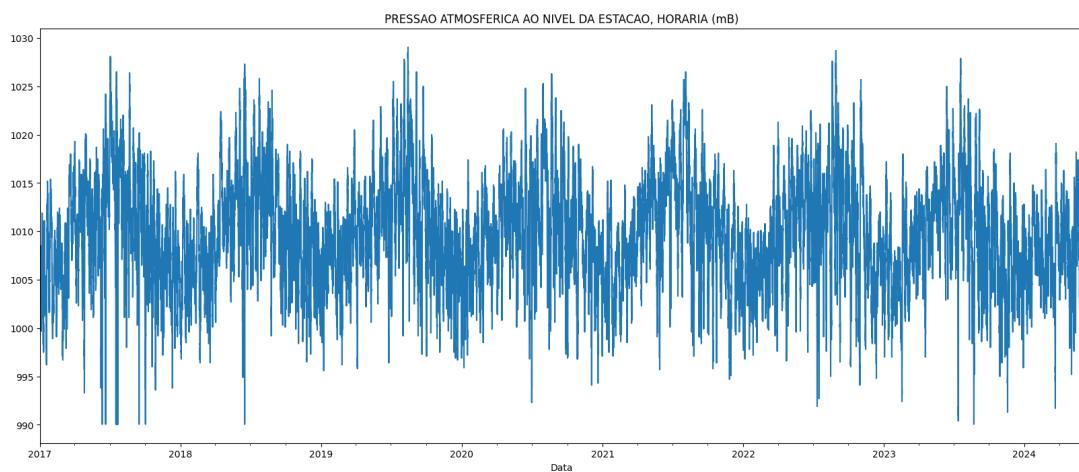
Após essa redução inicial, os gráficos e seus dados de cada coluna foram analisados, a fim de verificar se as informações eram consistentes e poderiam contribuir para o treinamento do modelo. Nas Figuras 20, 21, 22, 23, 24 e 25, são apresentados os gráficos de cada uma das colunas mantidas.

Figura 20 – Gráfico de Temperatura do Ar - Bulbo Seco



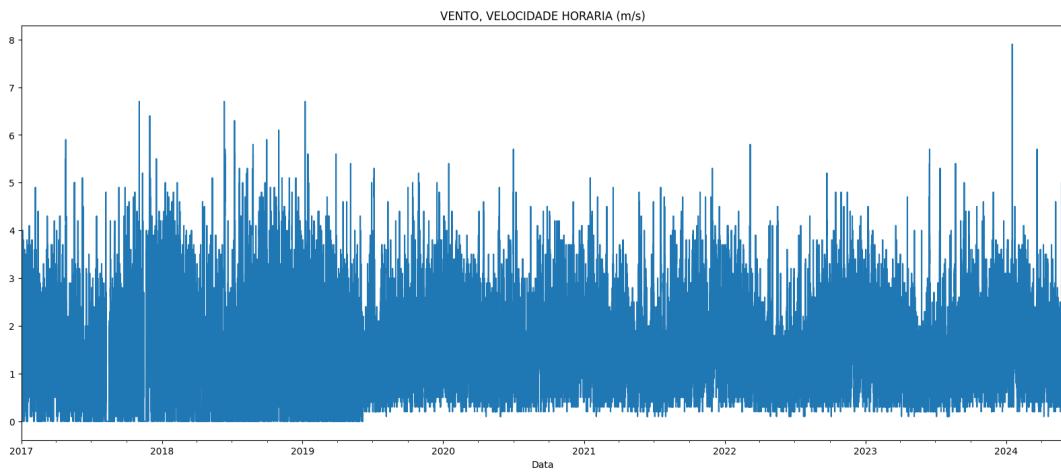
Fonte: Autor.

Figura 21 – Gráfico de Pressão Atmosférica ao Nível da Estação.



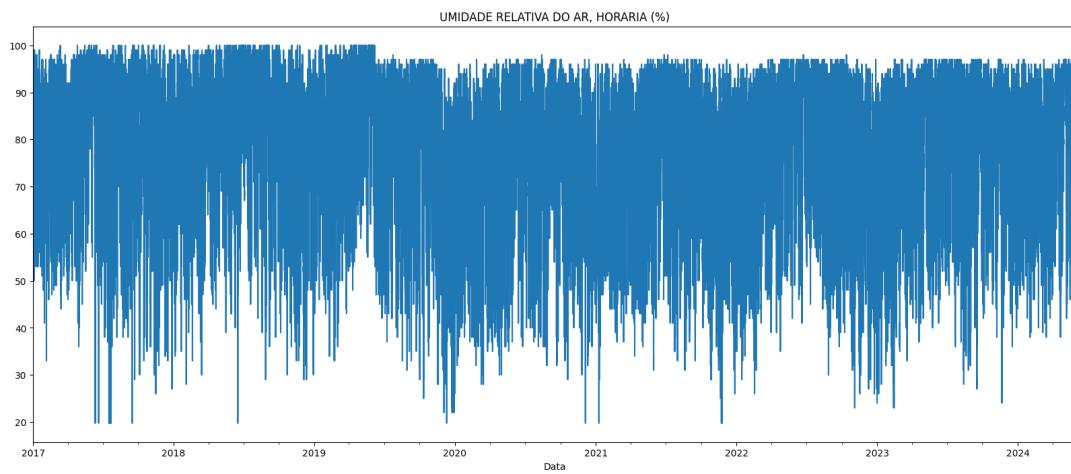
Fonte: Autor.

Figura 22 – Gráfico de Vento - Velocidade Horária.



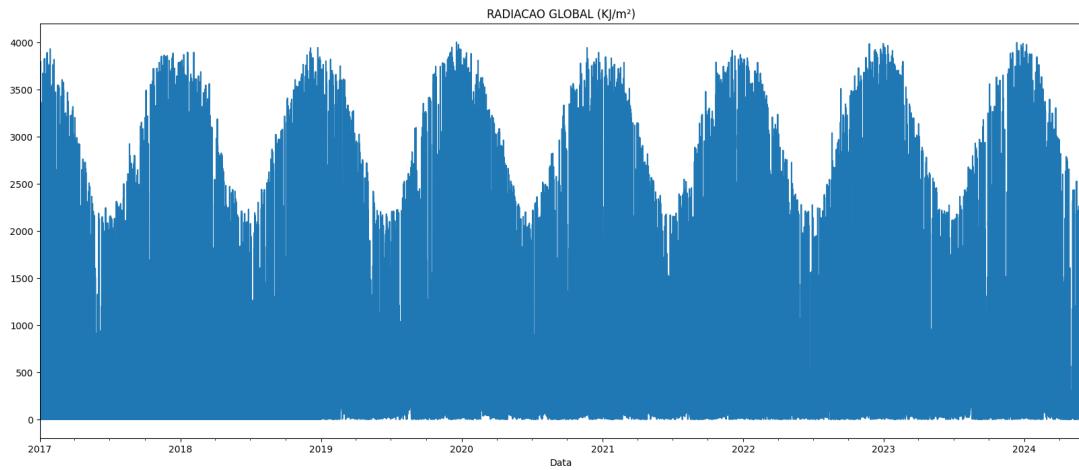
Fonte: Autor.

Figura 23 – Gráfico de Umidade Relativa do Ar



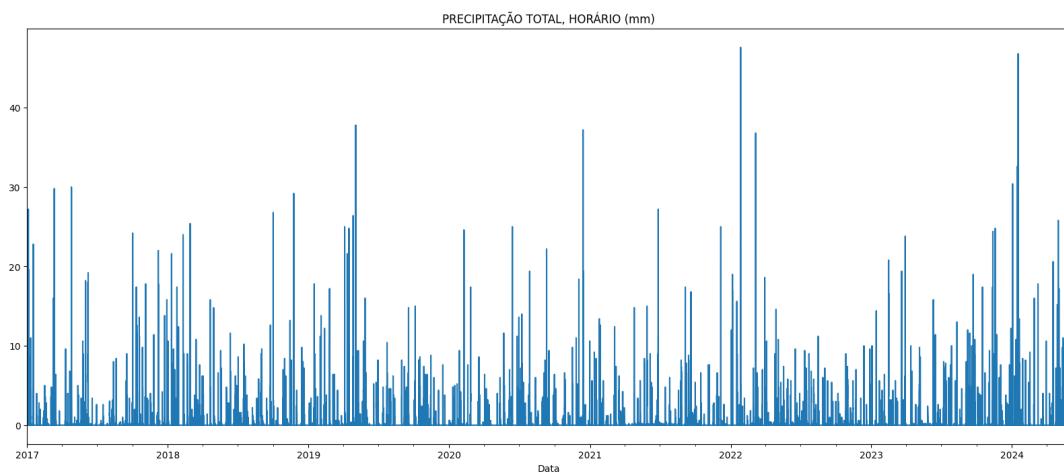
Fonte: Autor.

Figura 24 – Gráfico de Radiação Global



Fonte: Autor.

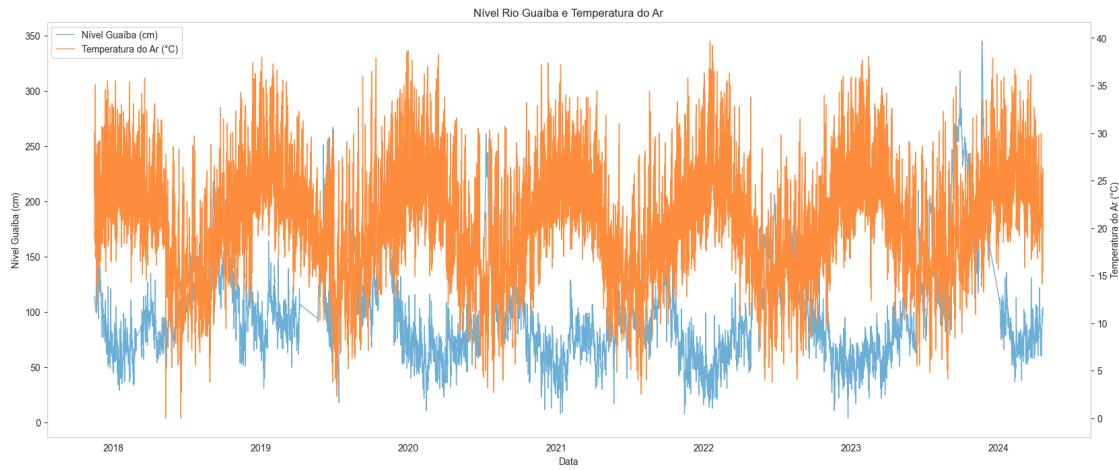
Figura 25 – Gráfico de Precipitação Total



Fonte: Autor.

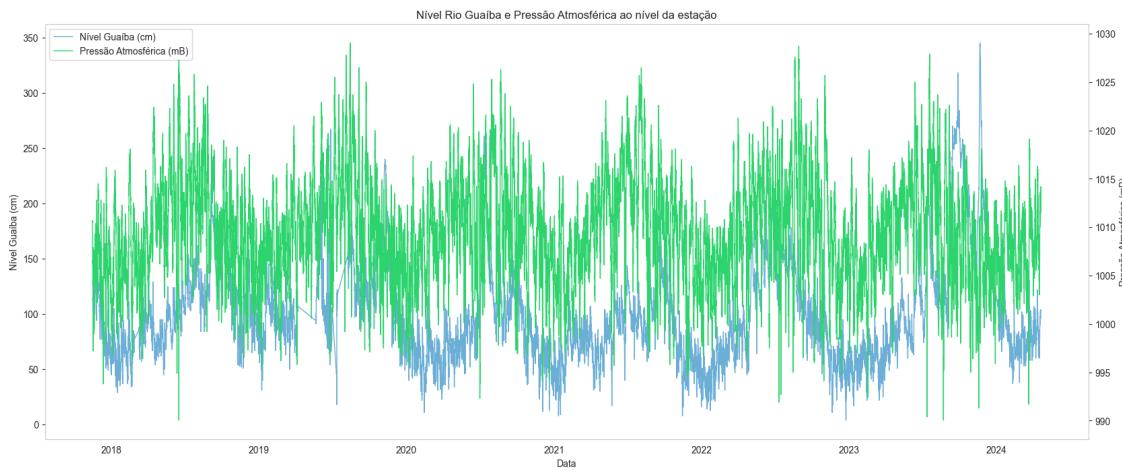
As Figuras 20, 21 e 24 ilustram um padrão sazonal nos dados, com aumentos e diminuições periódicas dos valores ao longo do ano, indicando uma correlação entre as variáveis meteorológicas e as estações do ano. Desse modo, torna-se justificável a manutenção dessas colunas, uma vez que o modelo de previsão pode se beneficiar dessa sazonalidade para melhorar a acurácia das previsões, considerando que o nível do rio também apresenta a mesma dinâmica, ilustrado nas Figuras 26, 27 e 33.

Figura 26 – Gráfico comparativo de temperatura e nível quanto a sazonalidade



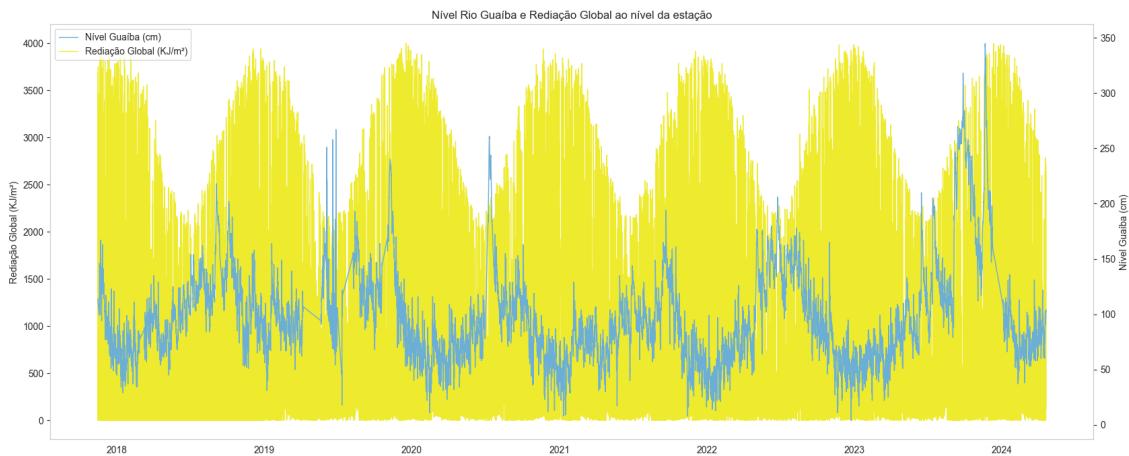
Fonte: Autor.

Figura 27 – Gráfico comparativo de pressão atmosférica e nível quanto a sazonalidade



Fonte: Autor.

Figura 28 – Gráfico comparativo de radiação global e nível quanto a sazonalidade



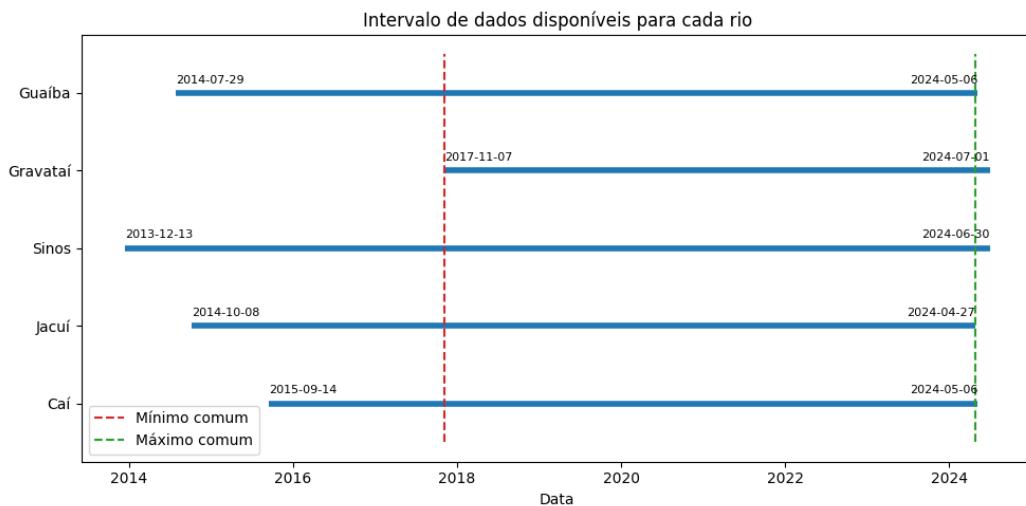
Fonte: Autor.

Em relação às colunas *Vento - Velocidade Horária, Umidade Relativa do Ar* e *Precipitação Total*, embora estas não apresentem um padrão sazonal tão evidente quanto as plotadas acima, estas medidas estão diretamente ligadas aos fenômenos de previsão do tempo, noticiadas em jornais e revistas, e portanto, possuem influência direta no nível dos rios onde estes dados são monitorados. Sendo assim, as colunas também foram consideradas para o treinamento do modelo.

Definidas as colunas a serem mantidas, o próximo passo consistiu em alinhar os dados meteorológicos com os dados de monitoramento dos rios, garantindo que as informações estivessem na mesma frequência de amostragem. Os dados de monitoramento do nível do Lago Guaíba possuem maior frequência de amostragem entre os rios analisados, com intervalo de 15 minutos, enquanto os dados meteorológicos possuem frequência de 1 hora (60 minutos). Assim, a frequência de amostragem das informações do nível dos rios foi reduzida para 1 hora, aplicando um filtro simples que mantém apenas as linhas cujo *timestamp* tem minuto igual a 0.

Por fim, alinhados os dados meteorológicos com os dados de monitoramento dos rios, o *dataframe* resultante passou por um nivelamento superior e inferior quanto a data inicial e final, visando manter um período de amostragem em que todas as colunas possuíssem informações de seus respectivos níveis, garantindo a mesma quantidade de linhas preenchidas, como mostra a Figura 29. Embora na etapa de limpeza dos dados tenha sido aplicado o preenchimento de valores ausentes, verificou-se que as estações de monitoramento dos rios não possuem medições que se iniciam ou finalizam na mesma data.

Figura 29 – Nivelamento superior e inferior dos dados dos rios.



Fonte: Autor.

Extraindo os dados mostrados na Figura 29, a Tabela 4 apresenta o período de amostragem de cada uma das fontes de dados dos rios.

Rio	Data Mínima	Hora Mínima	Data Máxima	Hora Máxima
Rio dos Sinos	2013-12-13	05:00:00	2024-06-30	09:00:00
Rio Caí	2015-09-14	13:00:00	2024-05-06	14:00:00
Rio Gravataí	2017-11-07	12:00:00	2024-07-01	00:00:00
Rio Jacuí	2014-10-08	20:00:00	2024-04-27	01:00:00
Lago Guaíba	2014-07-29	14:00:00	2024-05-06	14:00:00

Tabela 4 – Datas mínima e máxima disponíveis para cada rio analisado

Quanto à base de dados meteorológicos, o site do INMET disponibiliza dados desde o início dos anos 2000 e mantém a base atualizada até os dias atuais, não sendo, portanto, um limitador para a definição do período do dataframe final. Assim, o período de amostragem final pode ser definido a partir de 07 de novembro de 2017, que é a data mínima do rio Gravataí, até 06 de maio de 2024, que é a data máxima do rio Caí, com um dataframe de 56366 linhas.

3.5 TRANSFORMAÇÃO DOS DADOS

A transformação dos dados é uma etapa fundamental para garantir que as informações estejam no formato correto para o treinamento do modelo de previsão. Nesta fase, os dados são convertidos em um formato numérico, adequado para algoritmos de aprendizado de máquina, e normalizados para assegurar que todas as variáveis tenham a mesma escala.

Com os dados meteorológicos e de monitoramento dos rios alinhados, a normalização dos dados é aplicada visando principalmente equilibrar os dados dos níveis dos rios com os

dados meteorológicos, tendo em vista que as escalas entre essas informações apresentam uma discrepância maior, pela diferença de unidade de medida entre elas.

Ademais, outro fator que deve ser considerado é se as colunas do *dataframe* são compostas de dados categóricos (geralmente representados através de textos) ou numéricos. Por se tratar de medições aferidas por sensores meteorológicos e/ou geográficos, todas as colunas da base de dados estudada são numéricas.

A partir dessa premissa, a normalização dos dados foi feita utilizando o método *StandardScaler* da biblioteca *Scikit-learn*, que transforma os dados de uma coluna para que a média de seus valores seja zero e o desvio padrão seja igual a um. A pontuação padrão de uma amostra x é dada por:

$$z = \frac{x - \mu}{\sigma} \quad (16)$$

onde μ é a média da amostra e σ é o desvio padrão. Muitos elementos usados em funções objetivas de ML assumem que os dados passados para o modelo estão normalizados. No caso do Modelo Ridge, que utiliza um regularizador L2 em sua função objetivo, esse requisito também é necessário, logo, a normalização pelo método *StandardScaler* garante que os recursos estão centrados em torno uma média igual a zero e variação de mesma ordem (DEVELOPERS, 2025a). Caso uma amostra apresente uma variância de magnitude maior do que outros, ele pode dominar a função objetivo e fazer o estimador incapaz de aprender com outros recursos corretamente como esperado.

O método *StandardScaler* também é sensível a *outliers*, que afetam a média e o desvio padrão calculados para a definição dos valores normalizados. Por esse motivo, a aplicação do método IQR para remoção dos *outliers*, descrito na seção 3.3, além de limpar os dados durante aquela etapa de preparação, garante que os dados estivessem mais homogêneos e não fossem distorcidos por valores extremos na normalização desta etapa.

Com os dados normalizados, o *dataframe* está pronto para ser dividido em conjuntos de treinamento e teste, chegando ao estágio final de preparação, e seguindo para o treinamento do modelo de previsão.

3.6 DIVISÃO DOS DADOS EM CONJUNTOS DE TREINAMENTO E TESTE

No processo de implementação de um modelo de previsão, a divisão da base de dados em conjuntos de treino e teste constitui a última etapa antes da aplicação do algoritmo, com o objetivo de avaliar a capacidade de generalização do modelo. De maneira geral, a base de dados é dividida em duas partes: uma maior, utilizada para o treinamento do modelo, e uma menor, destinada a testar seu desempenho em dados não observados durante o treinamento. Após a etapa de treinamento com o primeiro conjunto, é possível realizar previsões e compará-las com o segundo conjunto, cujos dados são conhecidos

apenas pelo pesquisador. Essa comparação permite a avaliação da performance do modelo por meio do levantamento de métricas de desempenho.

Na divisão dos dados em conjuntos de treinamento e teste, diferentes proporções podem ser utilizadas, dependendo das características do conjunto de dados e do modelo de aprendizado de máquina. Exemplos comuns incluem 80% dos dados para treinamento e 20% para teste (80-20), 70% para treinamento e 30% para teste (70-30), e 60% para treinamento e 40% para teste (60-40) (SUPRI *et al.*, 2023).

Não há uma regra geral que defina a melhor divisão entre os conjuntos de treinamento e teste, pois essa escolha pode variar de acordo com o tipo de dado, o modelo utilizado e outros fatores específicos de cada conjunto de informações. No presente trabalho, foi inicialmente adotada a proporção 80:20, com 80% dos dados para treinamento e 20% para teste, que produziu bons resultados em trabalhos relacionados (SUPRI *et al.*, 2023). Para comparar o desempenho do modelo com outras proporções, também foram testadas as divisões 70:30 e 60:40, a fim de verificar se as métricas de desempenho apresentavam o mesmo comportamento observado no estudo de previsão de preços de ações.

3.7 APLICAÇÃO DO MODELO DE PREVISÃO

Realizada toda a preparação dos dados, com a divisão em conjuntos de treino e teste e a padronização das variáveis independentes para garantir que todas tenham a mesma escala, o modelo *Ridge* é instanciado em um script em Python, com a possibilidade de ajuste de alguns parâmetros para fazer o treinamento e as previsões.

A Regressão *Ridge*, implementada na biblioteca *Scikit-learn* em Python, possui parâmetros ajustáveis que influenciam o desempenho e a robustez do modelo. O principal parâmetro é o alpha, que controla a intensidade da penalização L2, descrita na Seção 2.6. Valores maiores de alpha aumentam a regularização, reduzindo a magnitude dos coeficientes $\hat{\beta}_i$, o que pode levar a *underfitting* ao simplificar excessivamente o modelo. Por outro lado, valores menores aproximam o modelo da regressão linear padrão, com maior risco de *overfitting*.

O parâmetro `fit_intercept` determina se o modelo inclui um termo de intercepto β_0 na equação de regressão, permitindo que a reta ou hiperplano ajustado não passe necessariamente pela origem (0,0), ou seja, que $y = x$ quando $x = 0$. Quando `fit_intercept=True`, o modelo calcula o intercepto para melhor ajustar os dados.

O parâmetro `solver` define o método de otimização (como auto, svd, ou cholesky), impactando a eficiência computacional, enquanto `random_state` garante reproduzibilidade em solvers que utilizam aleatoriedade. A Figura 30 apresenta um trecho de código Python que ilustra a configuração desses parâmetros na implementação da Regressão *Ridge*.

Figura 30 – Exemplo de código Python para configuração da Regressão Ridge



```
from sklearn.linear_model import Ridge

# Configuração do modelo Ridge
model = Ridge(alpha=1.0, fit_intercept=True, solver='auto', random_state=42)

# Treinamento do modelo com dados normalizados
model.fit(X_train, y_train)

# Realizando previsão com o modelo treinado
y_pred = model.predict(X_test)
```

Fonte: Autor.

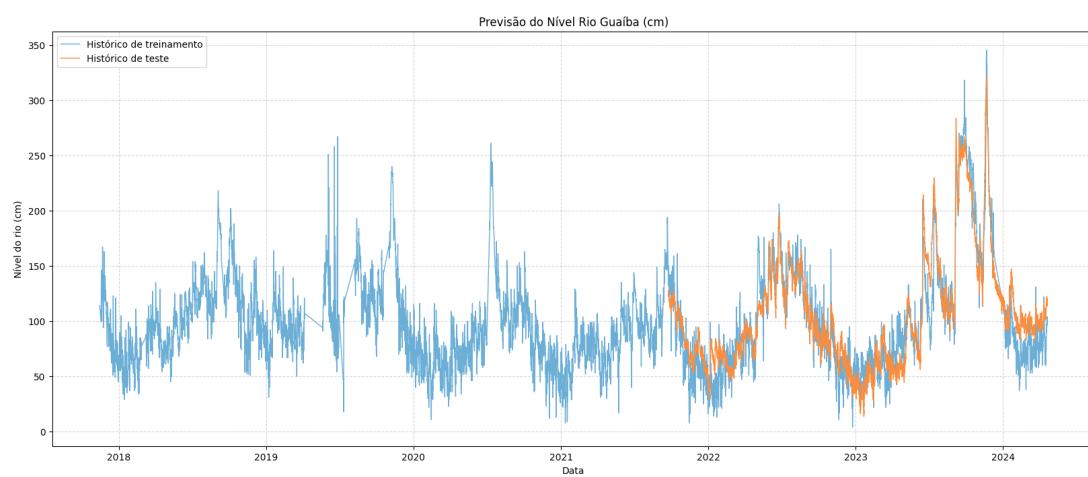
A implementação prática envolve importar a classe *Ridge*, configurar os hiperparâmetros, treinar o modelo com o método `fit` e realizar previsões com o método `predict`. Em (AND, s.d.), o autor aborda sobre o termo de regularização L2 que penaliza os coeficientes do modelo, explorando o impacto do parâmetro `alpha` em vários contextos. Em sua tese, é destacado que valores de `alpha` são geralmente selecionados por validação cruzada, com intervalos típicos variando de 0.01 a 1000, dependendo da escala dos dados e do grau de multicolinearidade. O artigo também menciona que, em aplicações práticas, valores como 0.1, 1 e 10 são frequentemente testados como pontos de partida.

Partindo desse pressuposto, foram testados valores de `alpha` de 0.001, 0.01, 0.1, 1, 10 e 100, com o objetivo de avaliar o desempenho do modelo em diferentes níveis de regularização. Os resultados obtidos para cada valor de `alpha` considerando as diferentes proporções treinamento e teste são descritos no Capítulo 4.

4 ANÁLISE DOS RESULTADOS

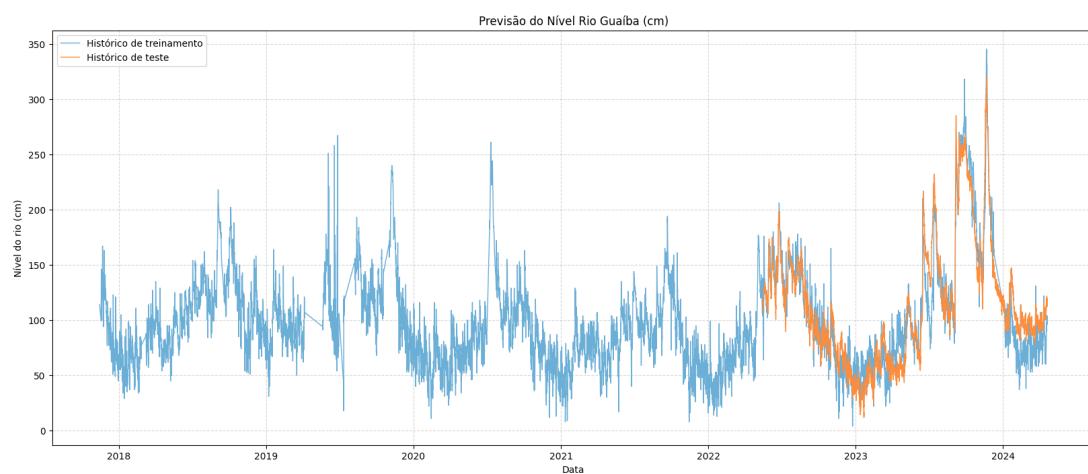
A seguir, são apresentados os resultados obtidos através do desenvolvimento descrito no Capítulo 3. Para cada valor de α , são exibidas as tabelas de desempenho, considerando as três proporções de divisão dos dados em conjuntos de treinamento e teste: 80:20, 70:30 e 60:40, junto às Figuras X, Y e Z, que ilustram o desempenho do modelo para cada split de treinamento e previsão.

Figura 31 – Gráfico modelo de previsão de nível do Lago Guaíba com $\alpha = 1$ e split de treinamento e teste 60:40



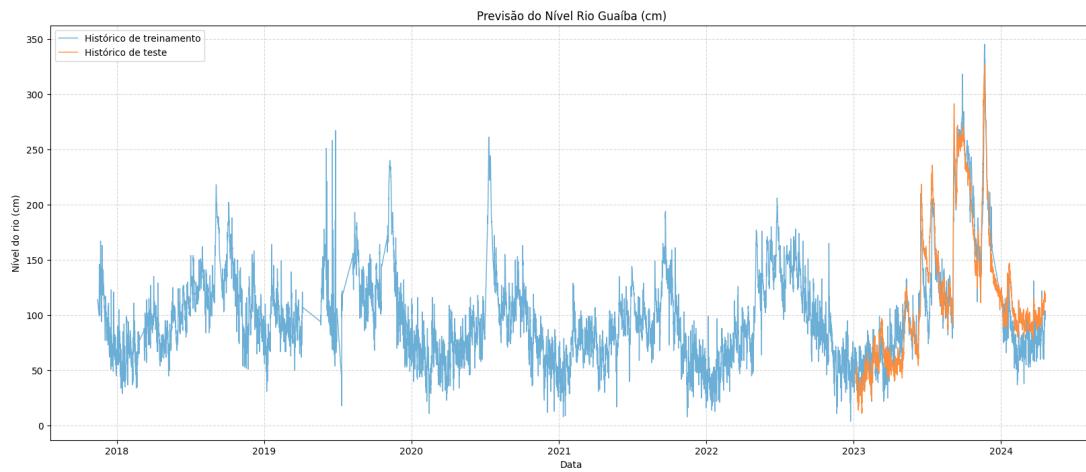
Fonte: Autor.

Figura 32 – Gráfico modelo de previsão de nível do Lago Guaíba com $\alpha = 1$ e split de treinamento e teste 70:30



Fonte: Autor.

Figura 33 – Gráfico modelo de previsão de nível do Lago Guaíba com $\alpha = 1$ e split de treinamento e teste 80:20



Fonte: Autor.

Nos valores de α testados, vale ressaltar que o modelo apresentou desempenho semelhante, com pequenas variações nas métricas de avaliação.

Alpha	0.001			
	Split	MSE	RMSE	MAE
80:20	456.72	21.27	16.81	0.89
70:30	372.44	19.30	15.00	0.88
60:40	379.64	19.48	15.14	0.87

Tabela 5 – Tabela de avaliação de desempenho do modelo com alpha 0.01

Alpha	0.001			
	Split	MSE	RMSE	MAE
80:20	456.72	21.27	16.81	0.89
70:30	372.44	19.30	15.00	0.88
60:40	379.64	19.48	15.14	0.87

Tabela 6 – Tabela de avaliação de desempenho do modelo com alpha 0.1

Alpha	0.001			
	Split	MSE	RMSE	MAE
80:20	456.72	21.27	16.81	0.89
70:30	372.44	19.30	15.00	0.88
60:40	379.65	19.48	15.14	0.87

Tabela 7 – Tabela de avaliação de desempenho do modelo com alpha 1

Alpha	0.001				
	Split	MSE	RMSE	MAE	R2
80:20	456.72	21.27	16.81	0.89	
70:30	372.44	19.30	15.00	0.88	
60:40	379.64	19.48	15.14	0.87	

Tabela 8 – Tabela de avaliação de desempenho do modelo com alpha 10

Alpha	0.001				
	Split	MSE	RMSE	MAE	R2
80:20	456.73	21.27	16.81	0.89	
70:30	372.45	19.30	15.00	0.88	
60:40	379.65	19.48	15.14	0.87	

Tabela 9 – Tabela de avaliação de desempenho do modelo com alpha 100

Uma das formas de avaliar o desempenho de um modelo de aprendizado de máquina é através da análise das métricas de desempenho, que indicam a capacidade do modelo de generalizar e prever corretamente os dados. As métricas de avaliação utilizadas neste trabalho incluem o erro quadrático médio (MSE), a raiz do erro quadrático médio (RMSE), o erro absoluto médio (MAE) e o coeficiente de determinação (R^2), onde o "erro" considerado nessas métricas refere-se à diferença entre os valores previstos pelo modelo e os valores reais observados nos dados. Cada métrica possui uma interpretação específica, sendo elas:

- Erro Quadrático Médio (MSE): Mede a média dos quadrados dos erros, penalizando erros maiores. É útil para entender a magnitude do erro, porém não é intuitiva em termos da escala dos dados, neste caso medido em centímetros.
- Raiz do Erro Quadrático Médio (RMSE): É a raiz quadrada do MSE, expressa na mesma unidade dos dados, facilitando a interpretação.
- Erro Absoluto Médio (MAE): Mede a média dos erros absolutos, sendo menos sensível a *outliers* do que o MSE.
- R^2 (Coeficiente de Determinação): Indica a proporção da variância dos dados explicada pelo modelo. Varia de 0 a 1, onde 1 significa ajuste perfeito.

Com isso, as tabelas apresentadas mostram que o MSE e o RMSE diminuem conforme o *split* de treinamento diminui e o *split* de teste aumenta. Tal comportamento indica que, à medida que mais dados são utilizados para treinamento, o modelo passa a ter um sobreajuste (*overfitting*), perdendo a capacidade de generalizar e, em vez disso, se adequando estritamente ao conjunto de dados de treinamento.

Quanto ao R^2 , o valor obtido possui uma baixa variação entre os ensaios com diferentes valores de *alpha* e *split*, ficando entre 0.87 e 0.89. Isso indica que 87% a 89% da variância dos dados da variável dependente é explicada pelo modelo. Em outras palavras, o modelo captura a maior parte dos padrões nos dados, deixando apenas 11% a 13% da

variância como não explicada (devido a ruído, variáveis omitidas ou outros fatores). Um coeficiente de determinação próximo a 1 sugere que o modelo é robusto e confiável para o conjunto de dados analisado. Boa parte dessa robustez se dá pelo tratamento e limpeza dos dados, somada à a correlação entre as variáveis preditoras e a variável dependente.

Em relação ao parâmetro **alpha**, os resultados mostram que o modelo apresenta um desempenho semelhante para os valores de 0.001, 0.01, 0.1 e 1, com pequenas variações nas métricas de desempenho. Existem alguns fatores que levam a esse comportamento, como a baixa sensibilidade à regularização nos dados, intervalo de **alpha** insuficiente para notar diferenças significativas ou baixa multicolinearidade entre as variáveis preditoras. Alguns outros fatores como escala dos dados ou tamanho e qualidade do conjunto foram descartados, uma vez que ao longo do trabalho, evidenciam-se os procedimentos realizados para evitar este tipo de falha no modelo.

Uma forma de testar se o modelo possui uma sensibilidade maior à regularização seria aumentar o intervalo de valores de **alpha** e/ou expor o modelo a mais dados de teste. Tomando um valor muito maior de **alpha** em relação ao intervalo utilizado anteriormente, foram obtidos os seguintes resultados:

Alpha	1000000				
	Split	MSE	RMSE	MAE	R2
80:20	469.82	21.68	16.93	0.88	
70:30	379.54	19.48	15.00	0.88	
60:40	391.98	19.80	15.31	0.87	

Tabela 10 – Tabela de avaliação de desempenho do modelo com alpha 1000000.

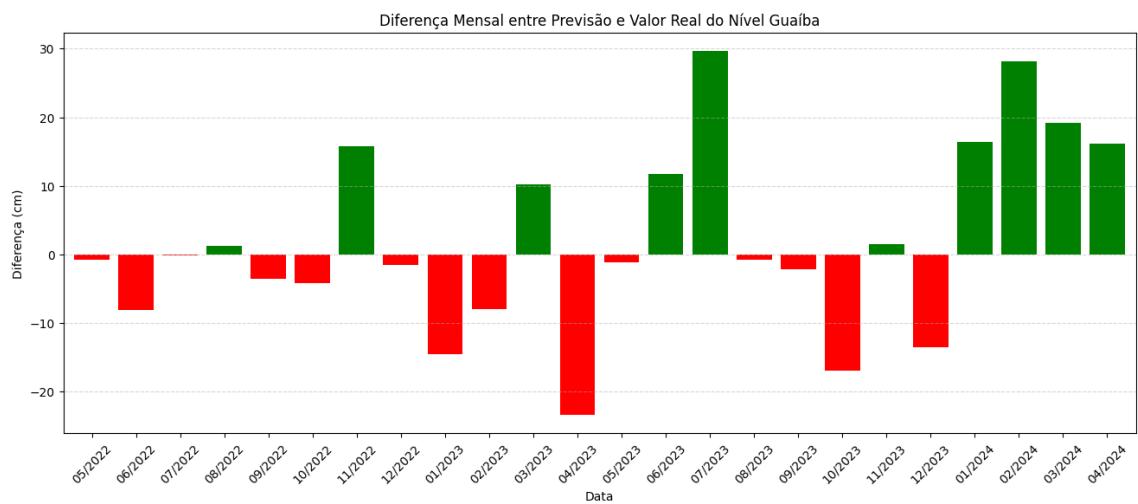
A partir das métricas de avaliação apresentadas, nota-se novamente que o aumento do valor de **alpha** resultou em um desempenho similar ao observado anteriormente, com erros um pouco maiores, e coeficientes de determinação próximos. Assim, em casos onde uma grande variação de **alpha** permanece com resultados semelhantes, tal comportamento indica que o modelo já atingiu seu ponto de saturação, obtendo um desempenho ótimo para baixos termos de regularização.

Considerando que a base de dados utilizada possui uma frequência temporal em horas, o modelo realiza uma previsão do nível do Lago Guaíba para o próximo período de uma hora, com base nos dados meteorológicos e do monitoramento dos rios daquela amostra temporal. Pensando em uma previsão com frequência diária, o modelo foi ajustado para ter um intervalo de previsão de 24 horas. Assim, ao prever o nível do lago na hora "00:00", a próxima previsão desconsidera os registros das 23 horas seguintes, realizando uma nova previsão no dia seguinte, com dados meteorológicos e dos demais rios também somente do outro dia. A Tabela 11 mostra um resumo mensal das previsões do modelo, agregandos as previsões diárias de cada mês como uma média e comparando os valores reais e previstos do nível do lago, além da diferença absoluta e percentual entre eles.

Mês/Ano	Atual(cm)	Previsto(cm)	Diferença(cm)	Diferença(%)
05/2022	126.62	125.83	-0.79	-0.62
06/2022	157.60	150.06	-7.54	-4.78
07/2022	139.67	140.09	+0.42	+0.30
08/2022	137.95	139.59	+1.64	+1.19
09/2022	102.62	99.25	-3.37	-3.28
10/2022	85.76	81.89	-3.87	-4.51
11/2022	64.00	79.35	+15.35	+23.98
12/2022	53.69	51.27	-2.42	-4.51
01/2023	51.08	36.35	-14.73	-28.84
02/2023	62.91	55.49	-7.42	-11.79
03/2023	58.01	67.77	+9.76	+16.82
04/2023	78.50	56.32	-22.18	-28.25
05/2023	94.21	93.28	-0.93	-0.99
06/2023	113.47	124.27	+10.80	+9.52
07/2023	138.53	162.85	+24.32	+17.56
08/2023	115.04	111.91	-3.13	-2.72
09/2023	229.53	218.44	-11.09	-4.83
10/2023	220.16	200.87	-19.29	-8.76
11/2023	201.79	202.08	+0.29	+0.14
12/2023	149.75	134.52	-15.23	-10.17
01/2024	95.98	110.42	+14.44	+15.04
02/2024	67.41	91.07	+23.66	+35.10
03/2024	76.78	90.61	+13.83	+18.01
04/2024	86.68	98.16	+11.48	+13.24

Tabela 11 – Diferença entre os valores reais e previstos do nível do Lago Guaíba

Figura 34 – Diferenças entre os valores reais e previstos do nível do Lago Guaíba



Fonte: Autor.

Com base na Tabela 11 e na Figura 34, é possível observar que, embora os erros

não se acumulem a cada mês, eles se tornam maiores de forma geral, conforme a previsão se afasta do início do teste. A o maior erro ocorreu em julho de 2024, com uma diferença de 24.32 cm em relação ao valor real do nível do lago, o que representa uma diferença percentual de 17.56%. Por outro lado, o menor erro ocorreu em julho de 2022, com uma diferença de apenas 0.42 cm, ou 0.30% em relação ao valor real.

Outro ponto a ser destacado é que, 14 dos 24 meses analisados, o modelo apresentou uma previsão com erro negativo, ou seja, o valor previsto foi menor que o valor real do nível do lago. Isso pode indicar que o modelo tende a subestimar os níveis do lago, sendo um fator de risco no alerta para enchentes e inundações, uma vez que a previsão de níveis mais baixos pode levar a uma resposta inadequada em situações de emergência. Por outro lado, mesmo com a subestimação, 8 dos 14 meses com erro negativo apresentaram uma diferença percentual inferior a 5%.

Utilizando outras pesquisas de previsão do nível de rios para fins comparativos, um estudo que implementou um modelo de Rede Neural do tipo Long Short-Term Memory (LSTM), para a previsão do rio Tisza, na Hungria. De forma geral, o modelo obteve um erro absoluto médio que variou de 4,2 cm no 1º dia de previsão, até 34,7 cm no 7º dia. Outra análise feita no estudo mostrou que entre 68,5% e 76,1% das previsões ficaram dentro dos intervalos de precisão exigidos para os diferentes dias da previsão de 7 dias à frente. Estes intervalos eram progressivos, sendo o primeiro dia com tolerância de 5cm, a partir do 3º dia com tolerância de 15cm, no 5º dia com tolerância de 25cm e por fim no 7º dia com tolerância de 35cm (VIZI *et al.*, 2023).

Comparando com este Artigo, o modelo *Ridge*, embora tenha oscilado nas previsões, gerando um grau de incerteza no desempenho, apresentou resultados gerais melhores em termos de erro absoluto médio, especialmente nos dias mais distantes de previsão.

Ainda, em (GUO *et al.*, 2021), os autores apresentam a aplicação de quatro modelos de aprendizado de máquina (SVR (Support Vector Regression (Regressão por Vetores de Suporte)), RFR (Random Forest Regression (Regressão por Floresta Aleatória)), MLPR (Multilayer Perceptron Regression (Regressão por Perceptron Multicamadas)) e LGBMR (Light Gradient Boosting Machine Regression (Regressão por Máquina de Reforço de Gradiente Leve))) para prever o nível do rio Lan-Yang, em Taiwan, com horizontes de 1 a 6 horas. Os resultados indicam que o modelo LGBMR obteve o melhor desempenho, com erro médio absoluto de pico de nível de água de 0,22 metros e erro de tempo até o pico de 0,12 horas, especialmente na estação Kavalan, influenciada pela maré do rio estudado. As métricas de avaliação, incluindo R^2 , MAE, RMSE e NSE (Nash-Sutcliffe Efficiency (Eficiência de Nash-Sutcliffe)), mostraram valores de NSE acima de 0,72 para previsões de 1 hora, com redução de precisão em horizontes mais longos. O modelo LGBMR também apresentou eficiência computacional, com tempos de treinamento de 156 segundos e validação de 0,03 segundos, destacando-se como uma ferramenta viável para previsões em tempo real.

Por fim, no artigo (HASAN *et al.*, 2023), é realizado um estudo para levantamento geral de técnicas, desafios e soluções para a previsão de inundações. Nele, é visto que tal previsão pode ser realizada com robustez e confiança por meio de modelos de aprendizado profundo, como redes neurais recorrentes (RNNs) e de longa memória curta (LSTMs). Esses modelos utilizam dados hidrológicos (precipitação, fluxo de rios, derretimento de neve), meteorológicos (temperatura, umidade, vento), espaciais (imagens de satélite, topografia, uso do solo) e de sensores (níveis de água, umidade do solo) para capturar padrões temporais e espaciais. A robustez das previsões depende da qualidade, quantidade e integração desses dados, que devem ser preprocessados para remover ruídos e tratar valores ausentes. A avaliação do desempenho dos modelos é conduzida por métricas como erro médio absoluto (MAE), erro quadrático médio (RMSE) e coeficiente de correlação, que verificam a precisão e a confiabilidade das previsões, possibilitando sua aplicação em gestão de riscos de inundações.

5 CONCLUSÃO

O presente trabalho teve como objetivo geral desenvolver um modelo de previsão do nível do Lago Guaíba, utilizando dados meteorológicos por meio da técnica de Rgressão Ridge. Para alcançar esse propósito, foram estabelecidos objetivos específicos que envolveram desde a coleta e pré-processamento dos dados, até a implementação do modelo e análise dos resultados.

No desenvolvimento do projeto, procedeu-se à coleta de dados meteorológicos disponibilizados pelo INMET, assim como dos níveis hidrométricos dos rios da bacia do Guaíba, através do SEMA-RS. Esses dados passaram por um pré-processamento, que incluiu a limpeza, tratamento de valores ausentes, remoção de outliers e normalização, garantindo maior qualidade ao conjunto utilizado para treinamento. Foi também realizada a redução da dimensionalidade, mantendo apenas variáveis com maior relevância para a previsão, como temperatura do ar, pressão atmosférica, radiação global, velocidade do vento, umidade relativa e precipitação.

Na etapa de implementação, o modelo de Rgressão Ridge foi escolhido por sua capacidade de lidar com multicolinearidade e reduzir o risco de sobreajuste, característica importante considerando a volatilidade das variáveis climáticas. Foram testados diferentes valores do parâmetro de regularização (*alpha*), variando também as proporções de divisão entre conjuntos de treinamento e teste (80:20, 70:30, 60:40). Para avaliar o desempenho da previsão do modelo, as métricas de desempenho — MSE, RMSE, MAE e R^2 foram utilizadas, visando uma análise mais objetiva e estatística da implementação.

Os resultados evidenciaram a robustez do modelo e a correlação significativa entre variáveis meteorológicas e o nível do Lago Guaíba. Apesar de pequenas variações nos indicadores de desempenho, não se observou impacto expressivo da alteração dos valores de alpha, sugerindo que o modelo atinge um platô de desempenho para os intervalos testados. Tal constatação indica que ajustes adicionais de regularização, ou a inclusão de novas variáveis externas, poderiam ser explorados em trabalhos futuros para ganhos incrementais de performance. Além disso, cabe a comparação com outros modelos de previsão existentes, ou até mesmo uma rede neural para comparar o desempenho em relação ao modelo Ridge.

Assim, conclui-se que o objetivo proposto foi atingido. O modelo desenvolvido mostrou-se eficiente e tecnicamente sólido para prever o nível do Lago Guaíba a partir dos dados meteorológicos utilizados. Entretanto, vale ressaltar que os dados utilizados são históricos, ou seja, compõem medidas meteorológicas aferidas por instrumentos no momento da medição. Para uma aplicação real de previsão do nível do rio, visando um alerta antecipado de enchente, o modelo seria expostos a medidas meteorológicas previstas ao invés de medidas, adicionando um grau de incerteza que pode gerar previsões com um grau de erro maior. Por isso, é necessário realizar testes com dados previstos, avaliar

novas métricas de desempenho, para então considerar o uso do modelo em casos reais de previsão do nível do rio.

REFERÊNCIAS

AI, Pecan. **Data Preparation for Machine Learning: The Ultimate Guide to Doing It Right.** [S.l.: s.n.], 2023. Disponível em: <https://www.pecan.ai/blog/data-preparation-for-machine-learning/>. Acesso em: 4 mai. 2025.

ALKAMA, Djamel *et al.* **A cultura participativa no YouTube: relação entre ídolos-fãs em canais brasileiros.** 2020. Tese (Doutorado) – Universidade Estadual Paulista (UNESP). Acesso em: 05/04/2025.

AND, Trevor Hastie. Ridge Regularization: An Essential Concept in Data Science. **Technometrics**, v. 62, n. 4, p. 426–433.

ANDRADE, Leonardo Capeleto de; RODRIGUES, Lucía Ribeiro; ANDREAZZA, Robson; CAMARGO, Flávio Anastácio de Oliveira. Lago Guaíba: uma análise histórico-cultural da poluição hídrica em Porto Alegre, RS, Brasil. **Engenharia Sanitária e Ambiental**, v. 24, n. 2, p. 229–237, 2019.

ANDRADE, Mauro M.; SCOTTÁ, Fernando C.; JR., Elírio E. Toldo; WESCHENFELDER, Jair; NUNES, José C. Hidrodinâmica do Lago Guaíba: Resultados Preliminares. In: XXII Simpósio Brasileiro de Recursos Hídricos. Porto Alegre: Associação Brasileira de Recursos Hídricos, 2017.

CARBONELL, Jaime G.; MICHALSKI, Ryszard S.; MITCHELL, Tom M. **Machine Learning: An Artificial Intelligence Approach.** Berlin: Springer, 1983. P. 619.

COUSINEAU, Denis; CHARTIER, Sylvain. Outliers detection and treatment: a review. **International Journal of Psychological Research**, Universidad de San Buenaventura, v. 3, n. 1, p. 59–68, 2010.

DEVELOPERS, scikit-learn. **sklearn.preprocessing.StandardScaler.** [S.l.: s.n.], 2025. Online documentation. Accessed: 2025-06-16. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.

G1. Acima dos 5 metros, Guaíba volta a avançar sobre ruas no dia mais frio do ano em Porto Alegre. [S.l.: s.n.], mai. 2024. Disponível em: <https://g1.globo.com/rs/rio-grande-do-sul/noticia/2024/05/14/acima-dos-5-metros->

guaiba-volta-a-avancar-sobre-ruas-no-dia-mais-frio-do-ano-em-porto-alegre.ghtml.
Acesso em: 7 jul. 2025.

G1. Santa Catarina registra enchente, queda de barreira e dia mais chuvoso de 2024. 2024. Disponível em: <https://g1.globo.com/sc/santa-catarina/noticia/2024/05/19/sc-enchente-quedas-barreira-dia-mais-chuvoso.ghtml>. Acesso em: 15 out. 2024.

GUO, Wen-Dar; CHEN, Wei-Bo; YEH, Shen-Hai; CHANG, Chih-Hsin; CHEN, Hongey. Prediction of River Stage Using Multistep-Ahead Machine Learning Techniques for a Tidal River of Taiwan. **Water**, v. 13, n. 7, p. 920, 2021.

HASAN, Md Nazmul; AHMAD, Musheer; AFZAL, Haider; AHMAD, Aisha; EUSUFZAI, A. R. Khan. The State of the Art in Deep Learning Applications, Challenges, and Future Prospects: A Comprehensive Review of Flood Forecasting and Management. **Sustainability**, MDPI, v. 15, p. 9513, 12 2023.

HELLIWELL, John F.; HUANG, Haifang; WANG, Shun; NORTON, Max. Social Environments for World Happiness. In: WORLD Happiness Report 2020. [S.l.: s.n.], 2020.

KIRCH, Wilhelm. Pearson's Correlation Coefficient. In: ENCYCLOPEDIA of Public Health. Dordrecht: Springer Netherlands, 2008. P. 1090–1091.

MCDONALD, Gary C. Ridge regression. **WIREs Computational Statistics**, v. 1, n. 1, p. 93–100, 2009. ISSN 1939-0068.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to Linear Regression Analysis**. 5. ed. Hoboken, NJ: Wiley, 2012. P. 672. ISBN 978-0-470-54281-1.

NOBRE, Carlos Afonso; YOUNG, Andrea Ferraz; ORSINI, José Antônio Marengo; SALDIVA, Paulo Hilário Nascimento; AL., et. **Vulnerabilidades das megacidades brasileiras às mudanças climáticas: Região Metropolitana de São Paulo - Relatório Final**. São Paulo, Brasil, 2011. Acesso em: julho de 2025. Disponível em: https://www.nepo.unicamp.br/publicacoes/relatorio-final/megacidades_RMSP.pdf.

PORTO IMAGEM. **O Guaíba é um lago, sim!** [S.l.: s.n.], abr. 2009. Disponível em: <https://portoimagem.wordpress.com/2009/04/23/o-guaiba-e-um-lago-sim/>. Acesso em: 7 jul. 2025.

SARAVANAN, R.; SUJATHA, P. A State of Art Techniques on Machine Learning: A Perspective of Supervised Learning Approaches in Data Classification. In: PROCEEDINGS of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018). Puducherry: IEEE, 2018.

SASSI, Cecília P.; PEREZ, Felipe G.; MYAZATO, Letícia; YE, Xiao; FERREIRA-SILVA, Paulo H.; LOUZADA, Francisco. **Modelos de Regressão Linear Múltipla Utilizando os Softwares R e STATISTICA: Uma Aplicação a Dados de Conservação de Frutas**. São Carlos, SP, Brasil, 2012. Relatórios Técnicos. Disponível em: <http://www.icmc.usp.br>.

SCIKIT-LEARN DEVELOPERS. **Linear Models: Ridge Regression**. Acesso em: 05/04/2025. 2025. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression.

SERVICES, Amazon Web. **O que é regressão linear?** 2024. Disponível em: <https://aws.amazon.com/pt/what-is/linear-regression/>. Acesso em: 15 ago. 2024.

SOTO, Timothy. Regression Analysis. In: VOLKMAR, Fred R. (Ed.). **Encyclopedia of Autism Spectrum Disorders**. New York, NY: Springer New York, 2013. P. 2538–2538.

SUPRI, B.; RUDIANTO, Abdurohim; MAWADAH, Badriatul; ALI, Helmi. Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing. **JEMSI (Jurnal Ekonomi, Manajemen, Dan Akuntansi)**, v. 9, n. 4, p. 1403–1408, 2023.

VEJA. **De 1941 a 2024: por que as enchentes são um desafio constante no Rio Grande do Sul**. 2024. Disponível em: <https://veja.abril.com.br/ciencia/de-1941-a-2024-porque-as-enchentes-sao-desafio-constante-no-rs>. Acesso em: 15 out. 2024.

VIZI, Zsolt; BATKI, Bálint; RÁTKI, Luca; SZALÁNCZI, Szabolcs; FEHÉRVÁRY, István; KOZÁK, Péter; KISS, Tímea. Water level prediction using long short-term memory neural network model for a lowland river: a case study on the Tisza River, Central Europe. **Environmental Sciences Europe**, v. 35, n. 92, 2023.