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

Título do trabalho: subtítulo (se houver)

Blumenau
2025

Cláudio Lourenço Moreira

Título do trabalho: subtítulo (se houver)

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

Título do trabalho: subtítulo (se houver)

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 mês de 2025.

Banca Examinadora:

Prof. Dr. Maiquel de Brito
Instituição xxxx

Prof. Segundo, Dr.
Instituição xxxx

Prof. Terceiro, Dr.
Instituição xxxx

Este trabalho é dedicado a todos que me ajudaram,
orientaram e contribuiram para a minha graduação, em
especial meus pais, amigos e professores.

AGRADECIMENTOS

Inserir os agradecimentos aos colaboradores à execução do trabalho.

Texto da Epígrafe. Citação relativa ao tema do trabalho. É opcional. A epígrafe pode também aparecer na abertura de cada seção ou capítulo. Deve ser elaborada de acordo com a NBR 10520. (SOBRENOME do autor da epígrafe, ano)

RESUMO

No resumo são ressaltados o objetivo da pesquisa, o método utilizado, as discussões e os resultados com destaque apenas para os pontos principais. O resumo deve ser significativo, composto de uma sequência de frases concisas, afirmativas, e não de uma enumeração de tópicos. Não deve conter citações. Deve usar o verbo na voz ativa e na terceira pessoa do singular. O texto do resumo deve ser digitado, em um único bloco, sem espaço de parágrafo. O espaçamento entre linhas é simples e o tamanho da fonte é 12. Abaixo do resumo, informar as palavras-chave (palavras ou expressões significativas retiradas do texto) ou, termos retirados de thesaurus da área. Deve conter de 150 a 500 palavras. O resumo é elaborado de acordo com a NBR 6028.

Palavras-chave: palavra-chave 1; palavra-chave 2; palavra-chave 3.

ABSTRACT

Resumo traduzido para outros idiomas, neste caso, inglês. Segue o formato do resumo feito na língua vernácula. As palavras-chave traduzidas, versão em língua estrangeira, são colocadas abaixo do texto precedidas pela expressão “Keywords”, separadas por ponto e vírgula.

Keywords: keyword 1; keyword 2; keyword 3.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LISTA DE SÍMBOLOS

SUMÁRIO

1 INTRODUÇÃO

1.1 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos deste TCC.

1.1.1 Objetivo Geral

Descrição...

1.1.2 Objetivos Específicos

Descrição...

2 FUNDAMENTAÇÃO TEÓRICA

Explicar brevemente o que será tratado como fundamentação teórica para o entendimento do contexto em que o modelo de aprendizagem de máquina será aplicado.

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 repentinas. 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 indíces pluviométricos, levando ao transbordamento de rios, deslizamentos de terra e bloqueios em diversas rodovias (G1, 2024).

2.1.2 Dinâmica do Rio Guaíba

O Rio 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. No Artigo conduzido pelos pesquisadores (ANDRADE *et al.*, 2017), a variabilidade nas descargas líquidas do Rio Guaíba revelou flutuações significativas nos volumes de descarga, variando de 407 m^3/s a 14.270 m^3/s, o que indica uma grande influência das condições climáticas sazonais e da vazão dos rios tributários, como o Jacuí, Taquarí, Caí e Sinos. Essas variações extremas foram observadas durante o período de

2014 a 2017 e reforçam a importância de monitorar continuamente o regime de águas do Guaíba para prevenir enchentes e outros desastres associados (ANDRADE *et al.*, 2017).

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 aprendizado de máquina (do inglês, *machine learning*, ou ML) se tornou realidade, com aplicações em diversos setores relacionados ou não a 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

Com pesquisas e algoritmos sendo desenvolvidos para novas aplicações e/ou aprimoramento de implementações existentes, tornou-se necessário criar categorias de ML, a fim de classificar a sua função e estipular em quais cenários o seu uso é adequado.

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

- Supervisionado: é o mais comum e 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;
- Semi supervisionado: 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 um agente interage com o ambiente e aprende a otimizar suas ações para alcançar um objetivo a partir de feedbacks recebidos.

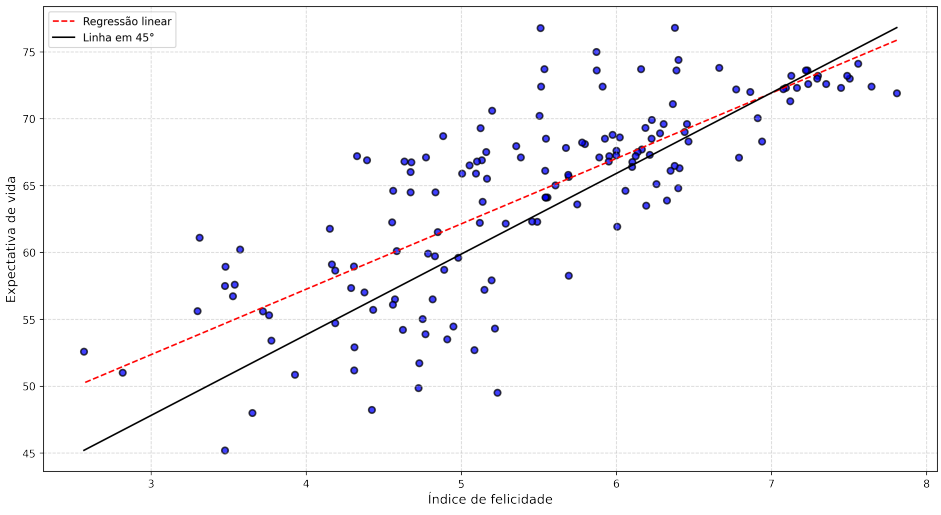
2.3 REGRESSÃO

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 excerce influência sobre y.

Na Figura 1, a fim de exemplificar um caso de regressão, é apresentada a relação entre a expectativa de vida baseada e um índice de felicidade calculado em diversos países obtidos a partir de um levantamento feito por (HELLIWELL *et al.*, 2020). Neste estudo, a variável independente é representada pelo índice de felicidade, enquanto a expectativa de vida representa a variável independente. Desse modo, uma análise visual do gráfico permite inferir uma tendência de expectativa de vida maior em países com alto índice de felicidade.

Figura 1 – 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 váriavéis

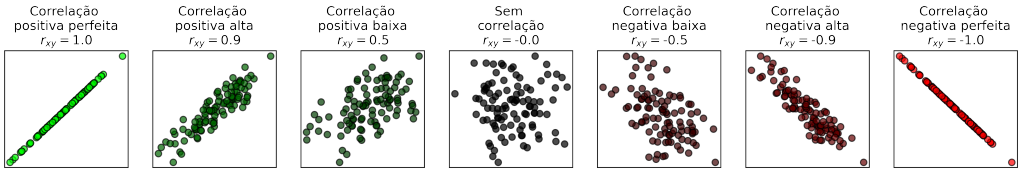
da equação, a criação de um modelo de previsão necessita de métodos que comprovem a correlação pressuposta. Para determinar as relações entre as variáveis dependentes e independentes de um sistema, coeficientes de correlação são calculados, gerando valores que medem e comprovam estatisticamente o grau 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.

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

A Figura 2 mostra alguns exemplos com gráficos de dispersão de variáveis com diferentes correlações.

Figura 2 – Diferentes correlações entre variáveis.



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

Sendo assim, com o cálculo do coeficiente indicando uma alta correlação entre os dados estudados, os métodos de regressão utilizam esta premissa entre as variáveis para estimar valores não existentes no conjunto de dados. Contudo, o coeficiente de correlação também pode mostrar variáveis que não interferem na dinâmica uma da outra no conjunto de informações analisadas, tornando necessário o uso de algoritmos robustos, que dispensam esse o fator de correlação para realizar as previsões desejadas.

2.4 REGRESSÃO LINEAR

A técnica de regressão linear é amplamente utilizada nos campos de estudo da engenharia, ciências físicas e químicas, economia, gestão, ciências biológicas e da vida, e ciências sociais. Em casos onde se deseja estabelcer uma relação entre uma variável preditora ou regressora (normalmente representada por x) e uma variável resposta (representada por y), a descrição dessa relação por meio de uma equação linear configura a implementação da técnica para a modelagem do problema estudado (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: regressão linear simples (RLS) e regressão linear múltipla (RLM) (MONTGOMERY; PECK; VINING, 2012). A RLS tem como objetivo estabelecer uma relação entre duas variáveis através de uma função, cuja definição é dada por:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2}$$

Onde y é a variável alvo, x a variável regressora, enquanto β_0 e β_1 são coeficientes calculados pela regressão, que representam o intercepto no eixo Y e a inclinação da reta, respectivamente.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{3}$$

Onde 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).

Partindo para um cenário ideal, a fim de se ter uma melhor compreensão do modelo de regressão linear, assume-se que seja possível fixar o valor da variável regressora x ao observar um valor y correspondente. Partindo dessa premissa, todos os valores do lado direito da equação (2) são conhecidos, exceto o erro ε, que passa a determinar as propriedades de y. Realizando outra suposição, onde a média e a variância de ε são iguais a zero e σ^2, respectivamente (MONTGOMERY; PECK; VINING, 2012), a resposta média para qualquer valor da variável regressora é dada por:

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

e a variância é dada por:

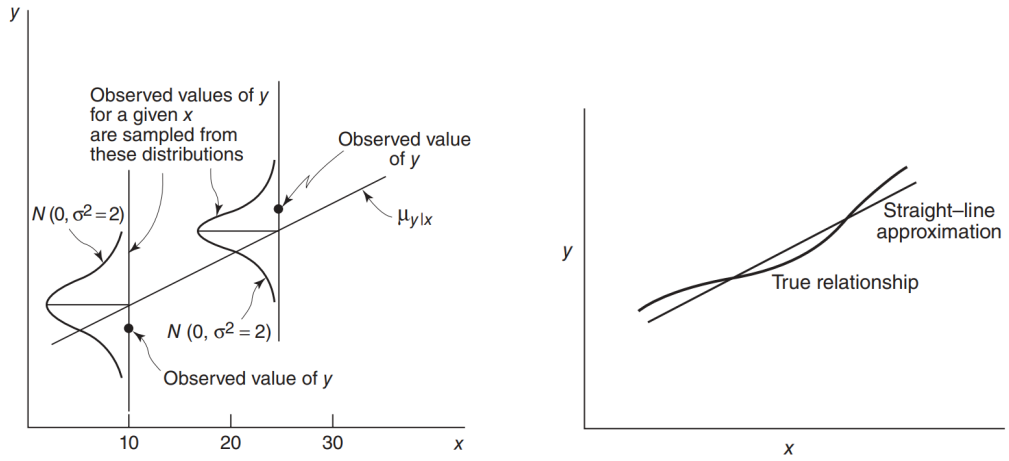
$$Var(y \mid x) = \sigma^2_{y|x} = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{5}$$

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 a suposição acima, tem-se um modelo de regressão ilustrado pela Figura 3a, onde $\mu_{y|x} = 3,5 + 2x$, com variância $\sigma^2 = 2$. Nota-se que uma distribuição normal é utilizada para descrever a variação aleatória do erro ε. Estabelecido que y é a soma de uma constante $\beta_0 + \beta_1 x$ (a média) e uma variável aleatória normalmente distribuída, é possível inferir que y também segue uma distribuição normal. No mesmo exemplo, se $x = 10$ amostras, y terá 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, enquanto uma variância maior resultará em pontos mais dispersos em relação à linha de regressão (MONTGOMERY; PECK; VINING, 2012).

A maioria dos fênomenos 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 3b, é mostrada uma relação entre as variáveis x e y relativamente complexa, mas que pode ser aproximada de por uma equação de regressão linear, com um erro relativamente baixo.

Figura 3 – Interpretação de uma regressão linear



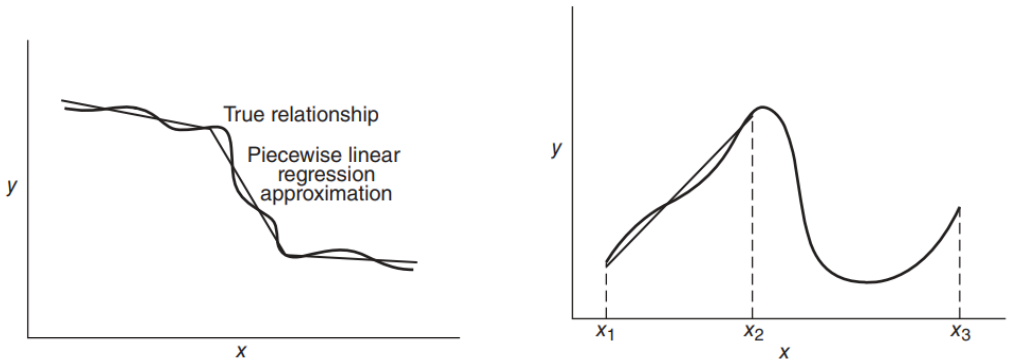
(a) Como as observações são geradas na regressão linear

(b) Regressão linear como aproximação de uma relação complexa

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 4a, utlizar 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 fênômeno é mais linear, a regressão apresenta um erro menor, como mostra a Figura 8f.

Figura 4 – Situações de inadequação da RLS



(a) Aproximação Linear complexa (b) Intervalo de aplicação 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 dos Mínimos Quadrados Orinários (MQO) 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 Soma Residual dos Quadrados (RSS), denotado por:

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

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

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

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_j 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:

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

simplificando:

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

divindindo por n:

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

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

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{9}$$

Derivada em relação à β_1:

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

substituindo β_0 na equação:

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

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:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

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.

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

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.

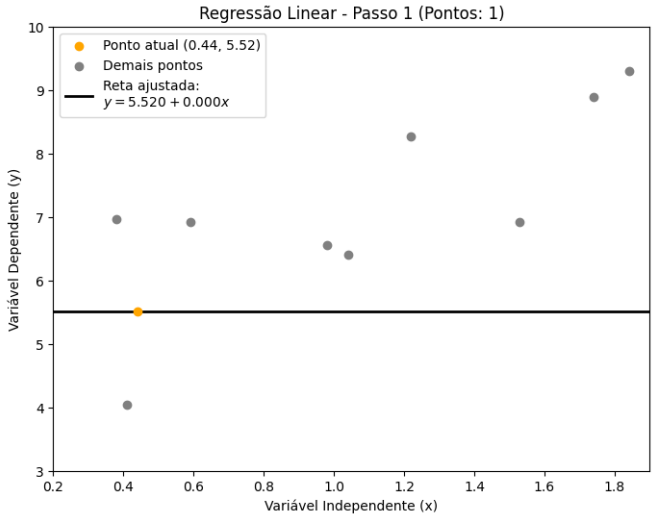
Ao aplicar o método para resolver um problema, como é o caso da Tabela 1, todos os pontos amostrados são utilizados para encontrar a reta que melhor se ajusta aos dados. Contudo, visando mostrar como a dinâmica de regressão utilizando MQO funciona, nos passos seguintes, as amostras são consideradas de 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.

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 5 – Passo 1 da regressão linear pelo método MQO.



Fonte: Autor.

Passo 2: Adiciona-se (1.74, 8.89)

n=2

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

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

$$(-0.65)(-1.685) + (0.65)(1.685) = 1.09525 + 1.09525 = 2.1905$$

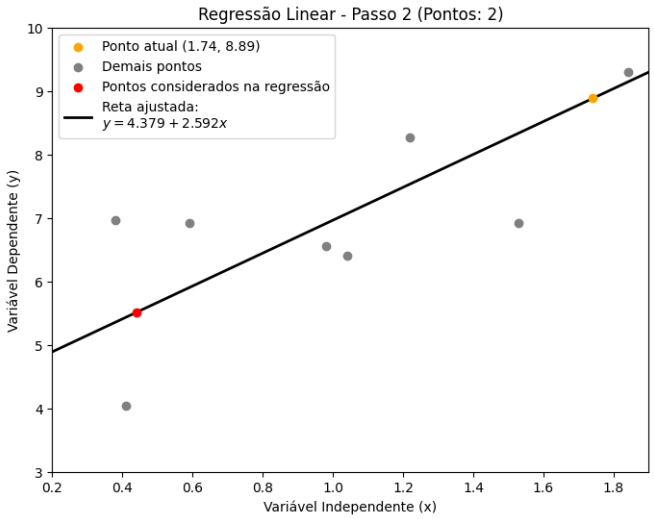
$$\sum (x_i - \bar{x})^2 = (0.44 - 1.09)^2 + (1.74 - 1.09)^2 = 0.4225 + 0.4225 = 0.845$$

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

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

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

Figura 6 – Passo 2 da regressão linear pelo método MQO.



Fonte: Autor.

Passo 3: Adiciona (0.41, 4.05)

n = 3

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

$$\sum (x_i - \bar{x})(y_i - \bar{y}) =$$

$$(0.44 - 0.8633)(5.52 - 6.1533) +$$

$$(1.74 - 0.8633)(8.89 - 6.1533) +$$

$$(0.41 - 0.8633)(4.05 - 6.1533) \approx$$

$$0.268 + 2.399 + 0.953 = 3.62$$

$$\sum (x_i - \bar{x})^2 = (0.44 - 0.8633)^2 + (1.74 - 0.8633)^2 + (0.41 - 0.8633)^2$$

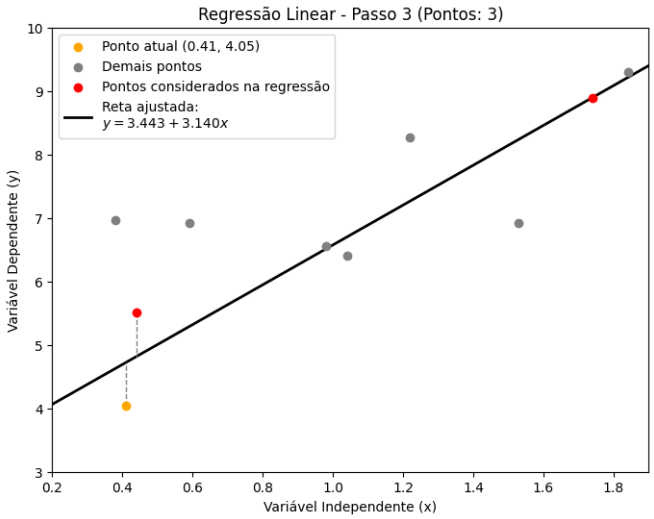
$$\approx 0.179 + 0.769 + 0.205 = 1.153$$

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

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

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

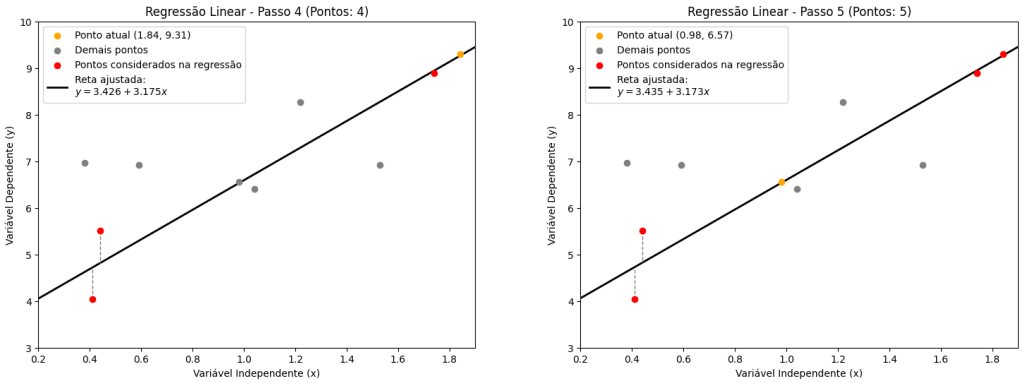
Figura 7 – Passo 3 da regressão linear pelo método MQO.



Fonte: Autor.

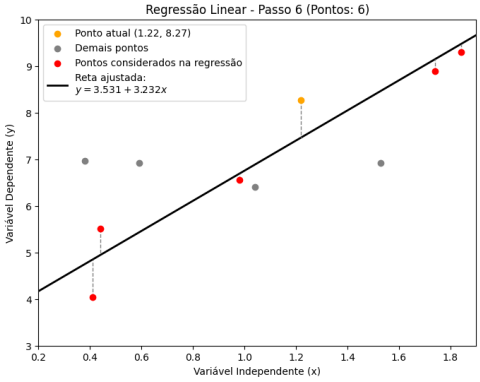
Para os demais passos, o mesmo cálculo é realizado, onde a média amostral e os coeficientes são recalculados a cada iteração, conforme os pontos são adicionados.

Figura 8 – Iterações da aplicação do método MQO

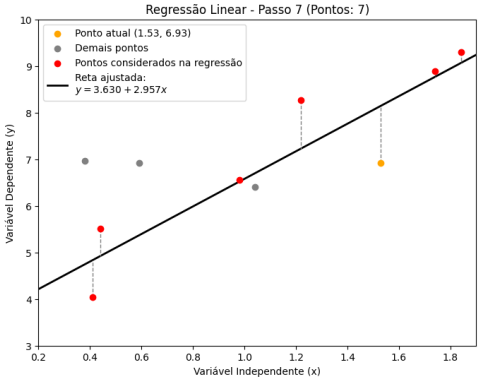


(a) $n = 4 \mid y = 3,426 + 3,175x$

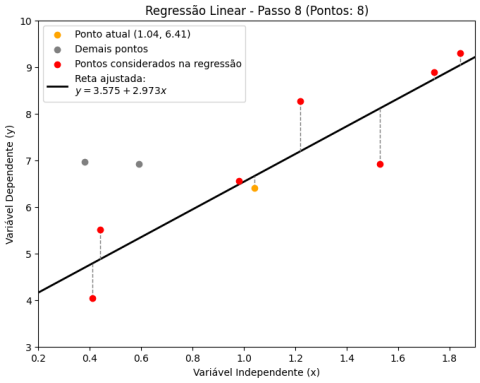
(b) $n = 5 \mid y = 3,435 + 3,173x$



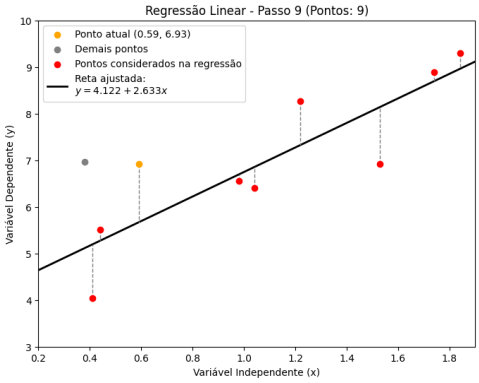
(c) $n = 6 \mid y = 3,531 + 3,232x$



(d) $n = 7 \mid y = 3,630 + 2,957x$



(e) $n = 8 \mid y = 3,575 + 2,973x$

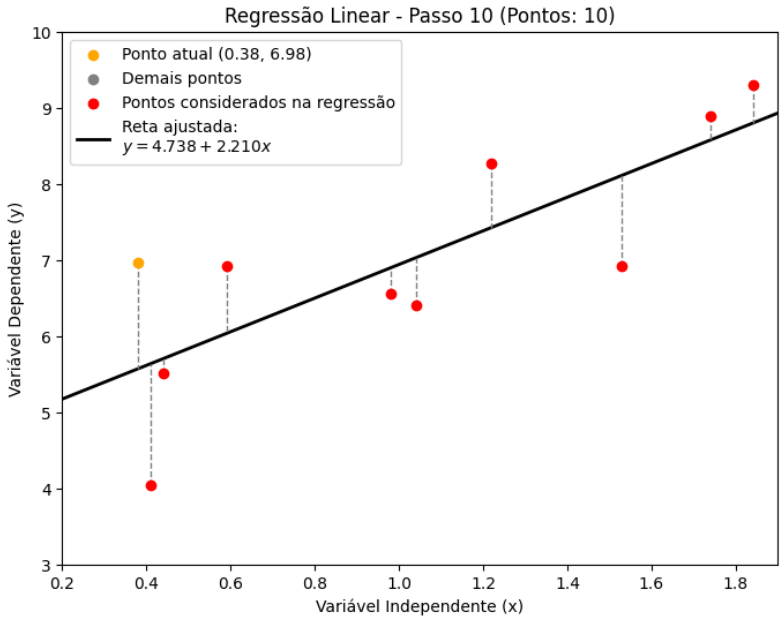


(f) $n = 9 \mid y = 4,122 + 2,633x$

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 9 são definidos de tal modo que o somatório dos seus quadrados seja o menor possível.

Figura 9 – 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 (overfitting) 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 a soma dos quadrados dos resíduos (Residual Sum of Squares, RSS), que mede a diferença entre os valores observados e os valores previstos pelo modelo. A função de perda é dada por:

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

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:

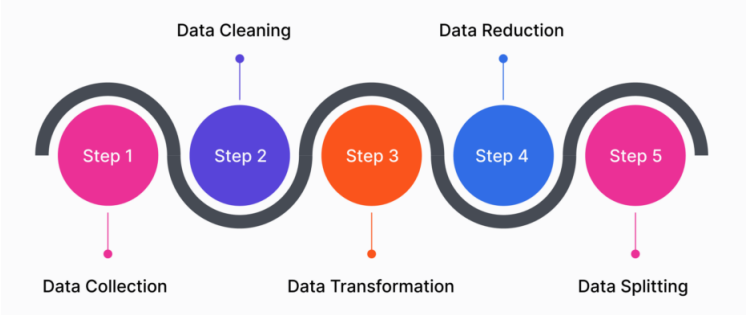
$$RSS_{L2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \tag{13}$$

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 (quando o modelo se ajusta demais aos dados de treinamento) quanto o underfitting (quando o modelo é muito simples para capturar os padrões dos dados). (SCIKIT-LEARN DEVELOPERS, 2025b).

3 PREPARAÇÃO DOS DADOS PARA O TREINAMENTO

Antes de iniciar o treinamento, exitem alguns procedimentos iniciais que garantem desde a coleta correta das informações necessárias para o modelo, até a limpeza, transformação e redução dos dados. Nesta seção, serão abordadas as etapas necessárias para preparar os dados para o treinamento do modelo, seguindo o fluxo de trabalho descrito na Figura 10.

Figura 10 – Passos para preparação dos dados.



Fonte: (AI, 2023)

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 fonte de dados foram utilizadas. Para os dados meteorológicos, o portal do *INMET*[1] (Instituto Nacional de Meteorologia) 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 no site da *SEMA-RS*[2] (Sala de situação), onde foram coletados os dados de nível do rio Guaíba,Caí, Jacuí, Sinos e Gravataí. Os dados meteorológicos foram coletados em formato *CSV*, com arquivos separados em anos de monitoramento com frequência horária, enquanto os dados dos níveis dos rios foram coletados no formato *.xlsx*, com histórico completo de amostragem das informações em frequência de 15 minutos, utilizando a biblioteca *Pandas* do Python.

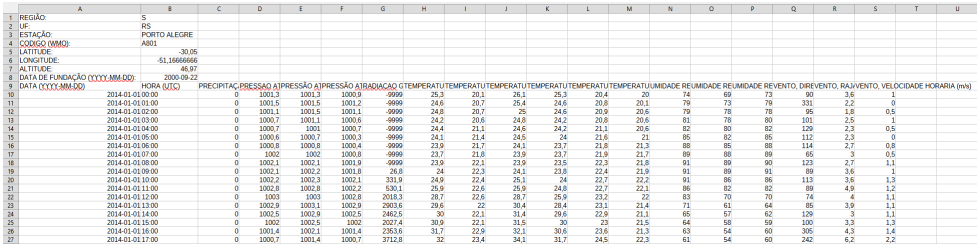
3.2 PRÉ PROCESSAMENTO DOS DADOS

Antes de seguir para o próximo passo mostrado na Figura 10, os dados coletados necessitam de um pré-processamento específico para cada uma das fontes utilizadas. Para

[1] https://portal.inmet.gov.br/
[2] https://www.saladesituacao.rs.gov.br/dados

os dados meteorológicos, devido às informações estarem separadas por ano de monitoramento, foi necessário concatenar os arquivos de cada ano em um único arquivo, utilizando a função *concat* da biblioteca *Pandas*, removendo o cabeçalho de informações geográficas da estação, ilustrado na Figura 11 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 11 – Dados meteorológicos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	REGIÃO:	S																			
2	UF:	RS																			
3	ESTAÇÃO:	PORTO ALEGRE																			
4	CODIGO (WMO):	A801																			
5	LATITUDE:	-30,05																			
6	LONGITUDE:	-51,16666666																			
7	ALTITUDE:	46,97																			
8	DATA DE FUNDAÇÃO (YYYY-MM-DD):	2000-09-22																			
9	DATA (YYYY-MM-DD)	HORA (UTC)	PRECIPITAC	PRESSAO A	PRESSÃO A	PRESSÃO A	RADIACAO G	TEMPERATU	TEMPERATU	TEMPERATU	TEMPERATU	TEMPERATU	TEMPERATU	UMIDADE R	EUMIDADE R	EUMIDADE R	EVENTO, D	IREVENTO, R	AJVENTO, V	ELOCIDADE H	ORARIA (m/s)
10	2014-01-01	00:00	0	1001,3	1001,3	1000,9	-9999	25,3	20,1	26,1	25,3	20,4	20	74	69	73	90	3,6	1		
11	2014-01-01	01:00	0	1001,5	1001,5	1001,2	-9999	24,6	20,7	25,4	24,6	20,8	20,1	79	73	79	331	2,2	0		
12	2014-01-01	02:00	0	1001,1	1001,5	1001,1	-9999	24,8	20,7	25	24,6	20,9	20,6	79	78	78	95	1,8	0,5		
13	2014-01-01	03:00	0	1000,7	1001,1	1000,6	-9999	24,2	20,6	24,8	24,2	20,8	20,6	81	78	80	101	2,5	1		
14	2014-01-01	04:00	0	1000,7	1001	1000,7	-9999	24,4	21,1	24,6	24,2	21,1	20,6	82	80	82	129	2,3	0,5		
15	2014-01-01	05:00	0	1000,6	1000,7	1000,3	-9999	24,1	21,4	24,5	24	21,6	21	85	82	85	112	2,3	0		
16	2014-01-01	06:00	0	1000,8	1000,8	1000,4	-9999	23,9	21,7	24,1	23,7	21,8	21,3	88	85	88	114	2,7	0,8		
17	2014-01-01	07:00	0	1002	1002	1000,8	-9999	23,7	21,8	23,9	23,7	21,9	21,7	89	88	89	65	3	0,5		
18	2014-01-01	08:00	0	1002,1	1002,1	1001,9	-9999	23,9	22,1	23,9	23,5	22,3	21,8	91	89	90	123	2,7	1,1		
19	2014-01-01	09:00	0	1002,1	1002,2	1001,8	26,8	24	22,3	24,1	23,8	22,4	21,9	91	89	91	89	3,6	1		
20	2014-01-01	10:00	0	1002,2	1002,3	1002,1	331,9	24,9	22,4	25,1	24	22,7	22,2	91	86	86	113	3,6	1,3		
21	2014-01-01	11:00	0	1002,8	1002,8	1002,2	530,1	25,9	22,6	25,9	24,8	22,7	22,1	86	82	82	89	4,9	1,2		
22	2014-01-01	12:00	0	1003	1003	1002,8	2018,3	28,7	22,6	28,7	25,9	23,2	22	83	70	70	74	4	1,1		
23	2014-01-01	13:00	0	1002,9	1003,1	1002,9	2903,6	29,6	22	30,4	28,4	23,1	21,4	71	61	64	85	3,9	1,1		
24	2014-01-01	14:00	0	1002,5	1002,9	1002,5	2462,5	30	22,1	31,4	29,6	22,9	21,1	65	57	62	129	3	1,1		
25	2014-01-01	15:00	0	1002	1002,5	1002	2027,4	30,9	22,1	31,5	30	23	21,5	64	58	59	100	3,3	1,3		
26	2014-01-01	16:00	0	1001,4	1002,1	1001,4	2353,6	31,7	22,9	32,1	30,6	23,6	21,3	63	54	60	305	4,3	1,4		
27	2014-01-01	17:00	0	1000,7	1001,4	1000,7	3712,8	32	23,4	34,1	31,7	24,5	22,3	61	54	60	242	6,2	2,2		

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/m²)	float64
Temperatura do Ar - Bulbo Seco (°C)	float64
Temperatura do Ponto de Orvalho (°C)	object
Temperatura Máxima na Hora Anterior (°C)	object
Temperatura Mínima na Hora Anterior (°C)	object
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 (%)	object
Umidade Relativa do Ar (%)	object
Vento - Direção Horária (° (gr))	object
Vento - Rajada Máxima (m/s)	float64
Vento - Velocidade Horária (m/s)	float64

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

A partir dela, nota-se que algumas colunas estão com o tipo de dado *object*, o que indica que os dados não estão no formato correto. Para resolver isso, foi necessário converter as colunas de temperatura e umidade relativa do ar para o tipo *float64*.

Já para os dados dos níveis dos rios, também foi necessário remover o cabeçalho com dados geográficos da estação de monitoramento, como mostra a Figura 12, junto da conversão do formato de data e hora para o padrão *datetime* da primeira coluna.

Figura 12 – Dados meteorológicos.

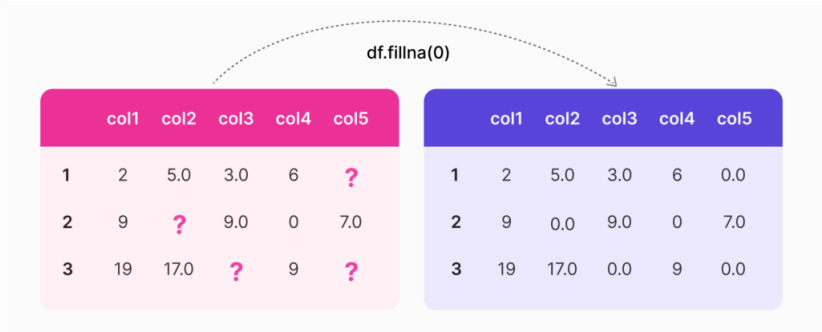
	A	B	C	D
1	Longitude: -51.0461			
2	Nome: CAMPO BOM			
3	Latitude: -29.6917			
4	Código: 87380000			
5	Data	Nível (cm)	Vazão (m³/s)	Chuva (mm)
6				
7				
8				
9				
10	30/06/2024 09:00:00	618	276,88	0
11	30/06/2024 08:45:00	619	278,37	0
12	30/06/2024 08:30:00	619	278,37	0
13	30/06/2024 08:15:00	619	278,37	0
14	30/06/2024 08:00:00	619	278,37	0
15	30/06/2024 07:45:00	619	278,37	0
16	30/06/2024 07:30:00	620	279,87	0
17	30/06/2024 07:15:00	619	278,37	0
18	30/06/2024 07:00:00	620	279,87	0

Fonte: Autor.

3.3 LIMPEZA DOS DADOS

Após a coleta dos dados, o próximo passo é a limpeza dessas informações, cujo procedimento 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, como mostra a Figura 13, ou com a média dos dados na coluna a ser limpa.

Figura 13 – Limpeza dos dados coletados.



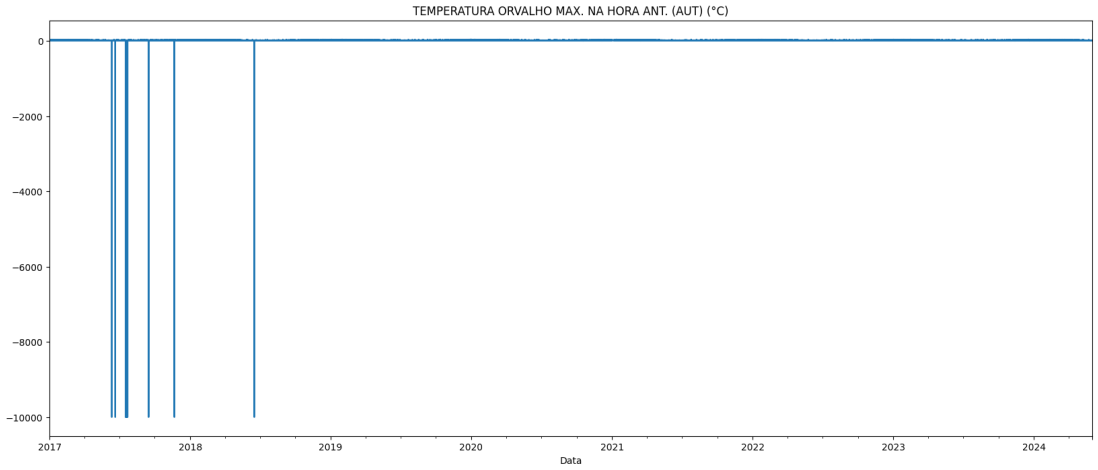
Fonte: (AI, 2023)

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) (°C)*, listada na Tabela 2, na Figura 14, observa-se os valores ausentes representados por ”-” e ”-9999”.

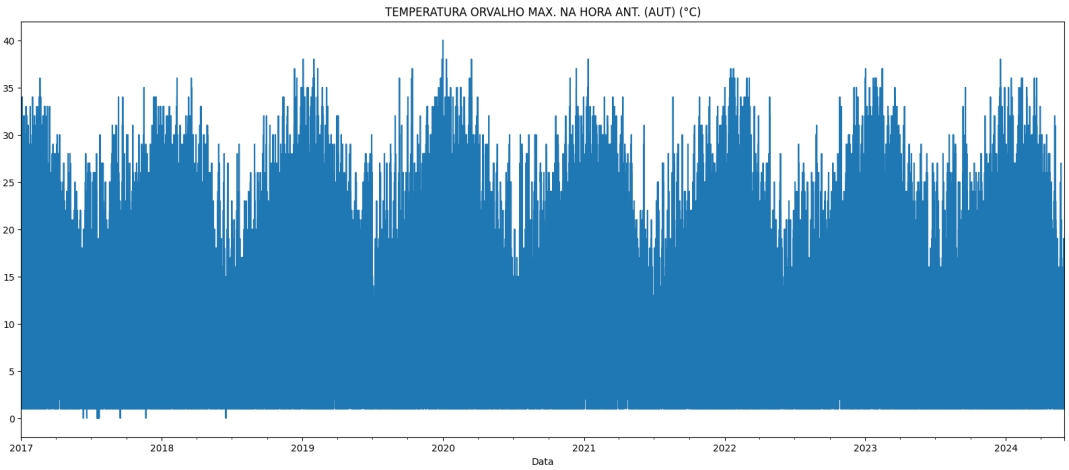
Figura 14 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior (°C) não tratado.



Fonte: Autor.

Após a substituição dos valores ausentes por zero, o gráfico da mesma coluna, mostrado na Figura 15, apresenta uma distribuição mais uniforme, sem os picos de dados ausentes.

Figura 15 – Gráfico de Temperatura do orvalho Máx. na Hora Anterior (°C) tratado.



Fonte: Autor.

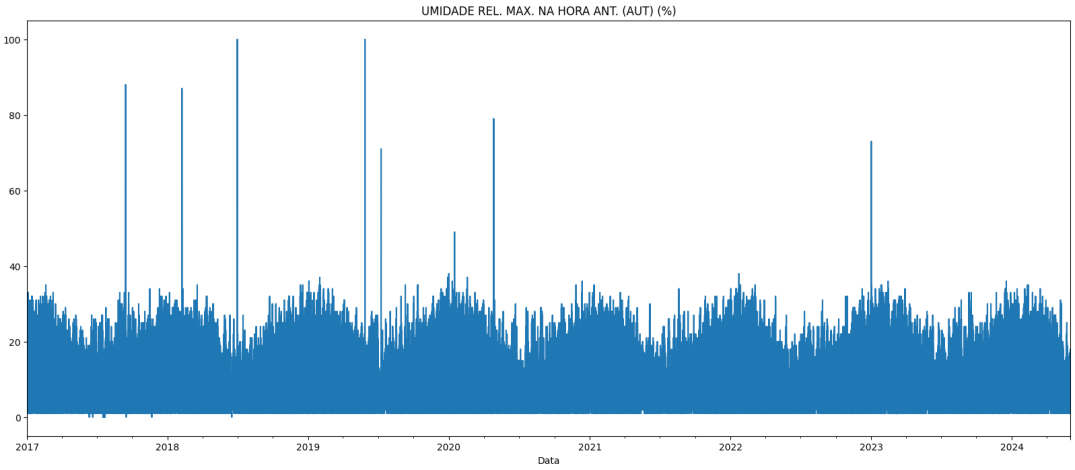
Aplicado o tratamento, a coluna *TEMPERATURA ORVALHO MAX. NA HORA*

ANT. (AUT) (°C) passa a apresentar uma distribuição mais uniforme, sem os picos de dados ausentes, com valores condizentes com a escala esperada da unidade medida, neste caso, a temperatura em graus Celsius.

Em casos onde os dados apresentam *outliers* (dados que se distanciam significativamente do restante da coluna analisada), foi aplicado o método IQR (Interquartile Range) 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.

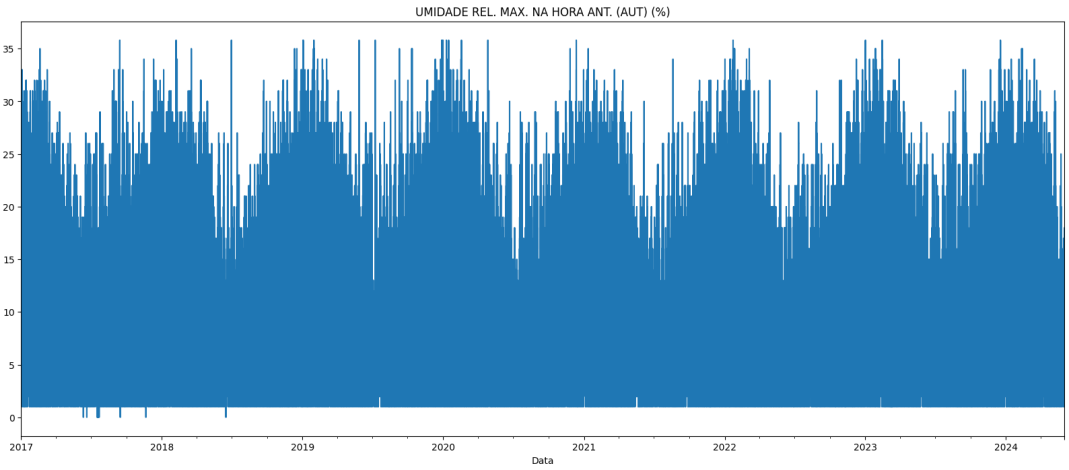
Por exemplo, na coluna *UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)*, as Figuras 16 e 17 mostram a diferença entre os dados após a aplicação do método IQR, com um fator de tolerância de 1.2 entre os interquartis Q1 e Q3.

Figura 16 – Gráfico de Umidade Relativa Máxima na Hora Anterior (%) sem tratamento.



Fonte: Autor.

Figura 17 – 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 respectivos 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, utilizou-se a função *interpolate* com o argumento *method = linear* da biblioteca *Pandas*, que preenche os valores ausentes com base nos dados adjacentes.

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 rio 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 dado, filtrando 6 colunas no total, conforme a Tabela 3.

Coluna	**Unidade de medida**
Temperatura do Ar - Bulboeco	(°C)
Pressão Atmosférica ao Nível da Estação	(mB)
Vento - Velocidade Horária	(m/s)
Umidade Relativa do Ar	(%)
Radiação Global	(kJ/m²)
Precipitação Total	(mm)

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

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

- As colunas *Temperatura do Ar - Bulboeco*, *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.
- 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. Para isso, foram utilizados gráficos de dispersão e histogramas, como os mostrados na Figura **??**.

ADICIONAR AQUI O HISTOGRAMA DOS DADOS METEOROLÓGICOS E A DECISÃO DE QUAIS COLUNAS FORAM MANTIDAS E QUAIS FORAM DESCARTADAS.

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. Para isso, foram utilizados os dados de monitoramento do nível do rio Guaíba, que possuem a 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 apnas 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 um nivelamento superior e inferior quanto a data inicial e final, de modo que o período de amostragem de todas as informações fosse o mesmo, garantindo que todas as colunas tivessem o mesmo número de linhas preenchidas. 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. Dessa forma, a fim de analisar qual o período de início de fim das informações obtidas, a tabela 4 apresenta o período de amostragem de cada uma das fontes de dados.

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
Rio 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 a 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 um valor cuja média seja zero e desvio padrão igual a um. A pontuação padrão de uma amostra x é dada por:

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

onde μ é a média da amostra e σ é o desvio padrão. Muitos elementos usados em funções objetivas de um algoritmo de aprendizagem (como o kernel RBF do Support Vector e as máquinas ou os regularizadores L1 e L2 dos modelos lineares) assumem que todos os recursos estão centrados em torno uma média igual a zero e variação de mesma ordem (DEVELOPERS, 2025a). Desse modo, 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.

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

Para exemplificar a normalização dos dados,

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

Existem diferentes abordagens para dividir os dados em conjuntos de treinamento e teste, sendo a mais comum a divisão aleatória, onde os dados são separados em duas partes, geralmente com 70% a 80% dos dados para treinamento e o restante para teste. Essa abordagem é simples e eficaz, mas pode não ser a melhor opção em todos os casos.

4 CONCLUSÃO

As conclusões devem responder às questões da pesquisa, em relação aos objetivos e às hipóteses. Devem ser breves, podendo apresentar recomendações e sugestões para trabalhos futuros.

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.

ANDRADE, Mauro M.; SCOTTÁ, Fernando C.; JR., Elírio E. Toldo; WESCHENFELDER, Jair; NUNES, José C. Hidrodinâmica do Rio 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.

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. **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.

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.

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.

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.

APÊNDICE A – Descrição

Textos elaborados pelo autor, a fim de completar a sua argumentação. Deve ser precedido da palavra APÊNDICE, identificada por letras maiúsculas consecutivas, travessão e pelo respectivo título. Utilizam-se letras maiúsculas dobradas quando esgotadas as letras do alfabeto.

ANEXO A – Descrição

São documentos não elaborados pelo autor que servem como fundamentação (mapas, leis, estatutos). Deve ser precedido da palavra ANEXO, identificada por letras maiúsculas consecutivas, travessão e pelo respectivo título. Utilizam-se letras maiúsculas dobradas quando esgotadas as letras do alfabeto.