Aplicação da regressão linear na previsão de salários com base em anos de experiência

Ângelo Carlos Avelino Feitosa[1], Paulo Anderson Gonçalves de Lima[1], Ester Miriã de Sousa Freire, Lucas Fontinele Alves Sousa[1], Simone Azevedo Bandeira de Melo Aquino[2], Daniel Duarte Costa[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – Campus Imperatriz (IFMA)

CEP 65.906-335 – Imperatriz – MA – Brasil

[2]Departamento de Ensino Superior e Tecnologia - Instituto Federal do Maranhão, IFMA

[2]UFMA: Centro de Ciências de Imperatriz - Universidade Federal do Maranhão.

{angelo.a, paulo.anderson, emiria, sousafontineli}@acad.ifma.edu.br, simonebandeira@ifma.edu.br, daniel.dc@ufma.br

Abstract. *This work addresses salary prediction based on years of experience, using linear regression with gradient descent on a Kaggle dataset (experience and salary). Preprocessing included exploratory analysis, handling of missing data, normalization, and data splitting, followed by manual implementation of the algorithm and evaluation with MSE, RMSE, MAE, and R^2. The results, with the obtained R^2, show good adherence and highlight the study's contribution to the practical understanding of machine learning algorithms from the ground up.*

Resumo. *O trabalho aborda a previsão salarial com base nos anos de experiência, usando regressão linear com gradiente descendente em dados do Kaggle (experiência e salário). Foram realizados pré-processamento (exploração, tratamento de ausentes, normalização e divisão), implementação manual do algoritmo e avaliação com MSE, RMSE, MAE e R^2. Os resultados, com R^2 obtido, demonstram boa aderência e evidenciam a contribuição do estudo para o entendimento prático de algoritmos de aprendizado de máquina desde os fundamentos.*

1. Introdução

A estrutura salarial permite diferenciar e decompor salários entre regiões [Barros, Corseuil & Mendonça, 1999]. A previsão salarial é relevante para contratações e planejamento, sendo a regressão linear uma abordagem possível. Essa modelagem descreve a relação entre duas variáveis, considerando apenas dependências estáticas [Rodrigues, Medeiros & Gomes, 2013] [Chein, 2019].

Nesse contexto, o trabalho apresenta uma implementação educacional de regressão linear para modelar a relação entre anos de experiência e salário. Também busca demonstrar todo o fluxo de trabalho em *machine learning*, do pré-processamento à avaliação, sem uso de bibliotecas prontas para o algoritmo principal. O artigo organiza-se em: fundamentação teórica (seção 2), metodologia (seção 3), resultados (seção 4) e considerações finais (seção 5).

2. Metodologia

A etapa de análise da base de dados é essencial, pois envolve exploração e sumarização para compreensão de características, propriedades e estrutura [Aithal, 2023]. Assim, utilizou-se a base "Salary Data" do Kaggle, a qual fornece um conjunto simples e ideal para a regressão linear, além de ser composta por 30 observações e duas colunas: *YearsExperience* (variável independente numérica) e *Salary* (variável dependente numérica).

No pré-processamento, realizou-se análise exploratória, verificando ausência de valores nulos e relação linear entre variáveis. Para fins de programação defensiva, implementou-se o algoritmo de regressão linear utilizando gradiente descendente, para o tratamento de dados não numéricos ou nulos e aplicou-se padronização z-score. Os dados foram divididos em 70% (21 amostras) para treino e 30% (9 amostras) para teste. A implementação[1] contemplou a função de custo, correspondente ao Erro Quadrático Médio (MSE), gradiente descendente e função de predição.

3. Resultados

O intervalo dos valores compreendidos na base coletada corresponde a 30 valores numéricos do tipo *float64* não nulos. Tal formato é adequado para regressão linear e envolve tanto os anos de experiência (*YearsExperience*) quanto o salário (*Salary*).

Foram descritos dados como contagem, média, desvio padrão, mínimo, máximo e variância, utilizados no cálculo dos erros quadráticos e absoluto médio, conforme Tabela 1. Os anos de experiência variam de 1,1 a 10,5, com média de 5,3. Já os salários variam de US$37.700,00 a US$122.400,00 anuais, com média de US$76.000,00. O alto desvio padrão do salário (aproximadamente US$27.400,00) indica grande dispersão.

Tabela 1. Parâmetros para fins de cálculo envolvendo partes da regressão linear.

	YearsExperience	**Salary**
count	30.00	30.00
mean	5.31	76003.00
std	2.83	27414.42
min	1.10	37731.00
max	10.50	122391.00

Os *boxplots* (diagramas de caixa), conforme figura 1, mostram que não há *outliers* extremos em nenhuma das variáveis. A mediana do salário (aproximadamente $65.000,00) está abaixo da média, indicando uma leve assimetria positiva.

[1] https://github.com/Pucapuka/RegressaoLinear/blob/main/SalaryYearsOfExperience.py

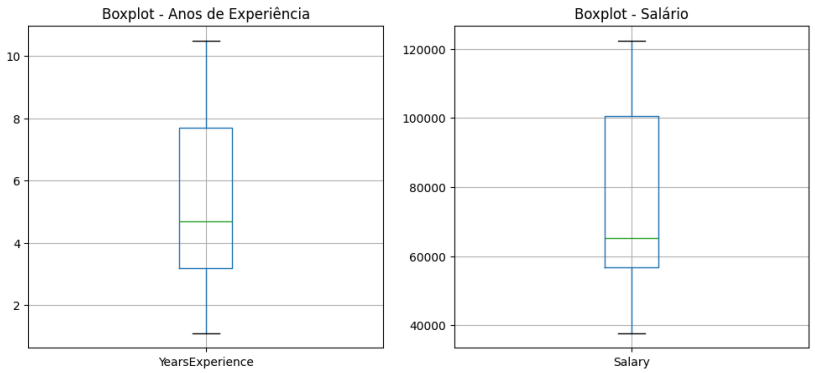

Figura 1. *Boxplots* de anos de experiência e salário.

O *scatter plot* (gráfico de dispersão), destacado na figura 2, detalha uma relação linear clara entre anos de experiência e salário, justificando o uso da regressão linear.

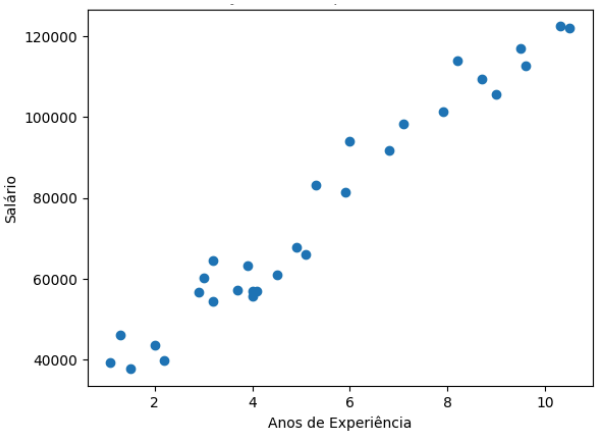

Figura 2. *Scatter plot da relação* entre anos de experiência e salário.

No gráfico de perda (MSE), demonstrado na figura 3, a curva mostra que o erro decresce rapidamente nas primeiras iterações e depois estabiliza, indicando que a taxa de aprendizado (lr = 0.1) foi bem escolhida.

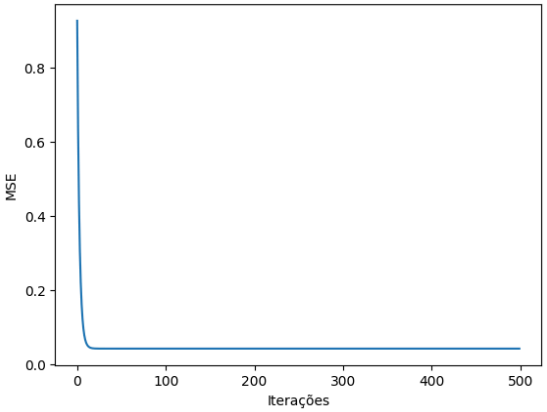

Figura 3. Convergência do Gradiente Descendente.

Em relação ao desempenho do modelo baseado em suas respectivas métricas, treinos e testes, foram obtidos os seguintes resultados. O desempenho da métrica MSE apresentou um valor aproximado de 0.2, considerado um erro quadrático médio baixo em escala normalizada, já a Raiz Quadrada do MSE (RMSE) obteve aproximadamente 0.45 como raiz

do erro quadrático, algo equivalente a próximo de R$12.300 na escala original. Já a Média de Valores Absolutos dos Erros (MAE) obteve erro absoluto médio de 0.35, equivalente a aproximadamente R$9600,00 na escala original, por fim, o coeficiente de determinação (R^2) explica 85% da variância pelos anos de experiência, sendo considerado um ótimo desempenho.

No gráfico de regressão linear, destacado na figura 4, a linha vermelha (predições) ajusta-se bem aos dados reais (pontos azuis). Apesar de alguns pontos terem erros maiores, a tendência geral se mostra bem precisa.

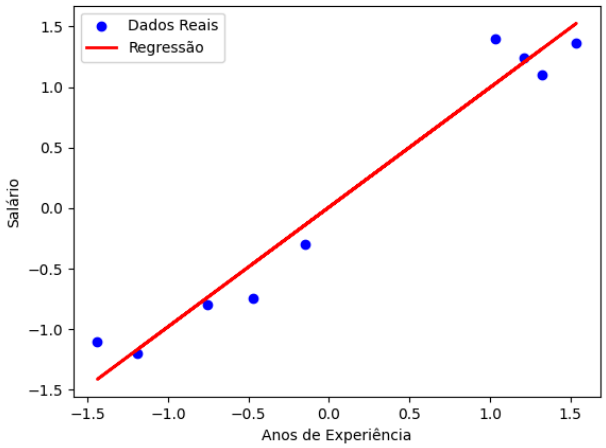

Figura 4. Regressão linear: Predição de salário (normalizado).

4. Considerações finais

O trabalho implementou um modelo de regressão linear com gradiente descendente para previsão salarial, eficaz, mas limitado pelo pequeno conjunto de dados e pelo uso de apenas uma variável preditora. Para futuras pesquisas, propõe-se adotar modelos mais complexos, incluir mais variáveis e comparar com outras técnicas de regressão, a fim de reduzir limitações e ampliar a aplicabilidade.

Referências:

Barros, R. P. D., Corseuil, C. H. L., & Mendonça, R. S. P. D. (1999). Uma análise da estrutura salarial brasileira baseada na PPV.

Chein, F. (2019). Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas.

Rodrigues, R. L., De Medeiros, F. P., & Gomes, A. S. (2013). Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)* (Vol. 24, No. 1, p. 607).

Tariq, D. T. H. S., & Aithal, P. S. (2023). Visualization and explorative data analysis. *International Journal of Enhanced Research in Science, Technology & Engineering*, *12*(3), 11-21.