Universidade Nove de Julho

Thiago Teixeira

ANALISA DE DADOS DE PONTUAÇÃO ANUAL DOS ALUNOS NA DISCIPLINA DE HISTÓRIA E GEOGRAFIA

Thiago Vinicius Fernandes Teixeira

ANALISA DE DADOS DE PONTUAÇÃO ANUAL DOS ALUNOS NA DISCIPLINA DE HISTÓRIA E GEOGRAFIA

Trabalho Final de Aprendizado de Máquina apresentado à Universidade Nove de Julho (Uninove), no Curso de Análise e Desenvolvimento de Sistemas.

Orientador: LEANDRO OLIVEIRA DA SILVA

RELATÓRIO

Primeiramente, busquei uma base de dados para tratar as informações de acordo com o objetivo do trabalho. O próximo passo foi compreender as informações obtidas para definir a melhor abordagem de análise. A base continha dados sobre a pontuação anual dos alunos em disciplinas de História e Geografia em escolas públicas de ensino médio na Argélia.

Inicialmente, examinei o número de alunos de ambos os gêneros e calculei a média e o desvio padrão das notas por gênero. As alunas tiveram um desempenho da média anual superior em comparação com os alunos, já o desvio padrão (dispersão) se encontra com uma leve vantagem para os alunos.

Em seguida, defini as variáveis independentes (X) e a variável dependente (y). Dividi os dados em conjuntos de treino e teste e, então, criei um modelo de regressão logística. Após o treinamento do modelo, realizei previsões e atingi uma acurácia de 98%. A matriz de confusão revelou a contagem de previsões corretas e incorretas, ajudando a verificar o desempenho do modelo.

Por fim, ao avaliar os coeficientes resultantes da regressão logística, cheguei a interpretações sobre as influências das variáveis no desempenho dos alunos.

ANÁLISE DE DADOS

As colunas foram dividias em gênero Masculino e Feminino.

Para determinar o resultado anual das notas, foram verificadas as colunas das notas obtidas em testes e exames. São elas: teste 1, teste 2, teste, 3 e exame 1, exame 2 e exame 3.

Obtendo assim o resultado.

Exam_02 (1.08): Esta variável tem o maior coeficiente, sugerindo que é a mais influente para prever uma maior nota anual. Isso significa que, para cada unidade de aumento no Exam_02, a probabilidade de a nota anual ser classificada como "alta" aumenta consideravelmente em relação às outras variáveis.

Exam_01 (0.92) e Exam_03 (0.90): Também têm coeficientes altos, indicando que os exames são fatores significativos na previsão da nota anual.

Test_01 (0.46), Test_02 (0.48), e Test_03 (0.49): Embora ainda sejam positivos, esses testes têm um impacto relativamente menor na probabilidade de uma alta nota em comparação com os exames.

A conclusão ao interpretar os coeficientes ajudou a entender como cada variável contribuiu para a previsão da variável dependente.

Variável	Coeficiente
Test_01	0.46
Exam_01	0.92
Test_02	0.48
Exam_02	1.08
Test_03	0.49
Exam_03	0.90

Relação com as Probabilidades

Com os resultados de Odds Ratios obtidos, realizei uma interpretação mais clara sobre o impacto relativo de cada variável na probabilidade de uma alta nota anual:

Exames:

- Exam_02 (2.96): a chance de a nota anual ser alta é multiplicada aproximadamente por 2.96. Esse é o maior Odds Ratio, indicando que Exam_02 é o fator mais forte para prever um desempenho alto.
- Exam_01 (2.52) e Exam_03 (2.47): Ambos também têm Odds Ratios elevados. Isso indica que, com o aumento dessas variáveis, há uma grande probabilidade de que a nota anual seja alta.

Testes:

- Test_03 (1.64), Test_02 (1.63) e Test_01 (1.60): Embora com impacto menor que os exames, os testes ainda têm uma influência positiva. Para cada aumento de uma unidade em qualquer um dos testes, a chance de um desempenho alto na nota anual aumento entre 1.6 e 1.64 vezes.

resultado:	Variável Odds_Ratio
0 Test_01	1.599941
2 Test_02	1.626596
4 Test_03	1.635175
1 Exam_01	2.520858
3 Exam_02	2.959795
5 Exam_03	2.465619

Resumo das Conclusões

- Exames são Mais Impactantes: Os Odds Ratios dos exames são significativamente maiores do que os dos testes, indicando que o desempenho nos exames tem um efeito mais forte sobre a probabilidade de uma nota alta no final do ano.
- Influência dos Testes: Embora os testes tenham um impacto positivo, eles são menos influentes em comparação com os exames. Isso pode indicar que os exames, possivelmente por cobrirem um conteúdo mais amplo ou serem mais decisivos, contribuem mais para o resultado.

Implicações

- Refinar Preparação para Exames: Como os exames têm o maior impacto, investir em estratégias de ensino e revisão focadas nos tópicos que caem nos exames pode ser eficaz.
- Avaliação de Peso de Componentes: Se os exames estão influenciando fortemente as notas anuais, pode-se avaliar se essa distribuição está adequada para os objetivos educacionais ou se seria mais justo equilibrar o peso com os testes.

Tomada de Decisão: Essa análise pode ajudar educadores e administradores a focarem mais nas variáveis que têm um impacto maior nas notas, como os exames, ao considerar mudanças curriculares ou intervenções pedagógicas.

Por último foi feita uma classificação de desempenho entre alto e baixo (0 = baixa, 1 = alta)

GRÁFICOS

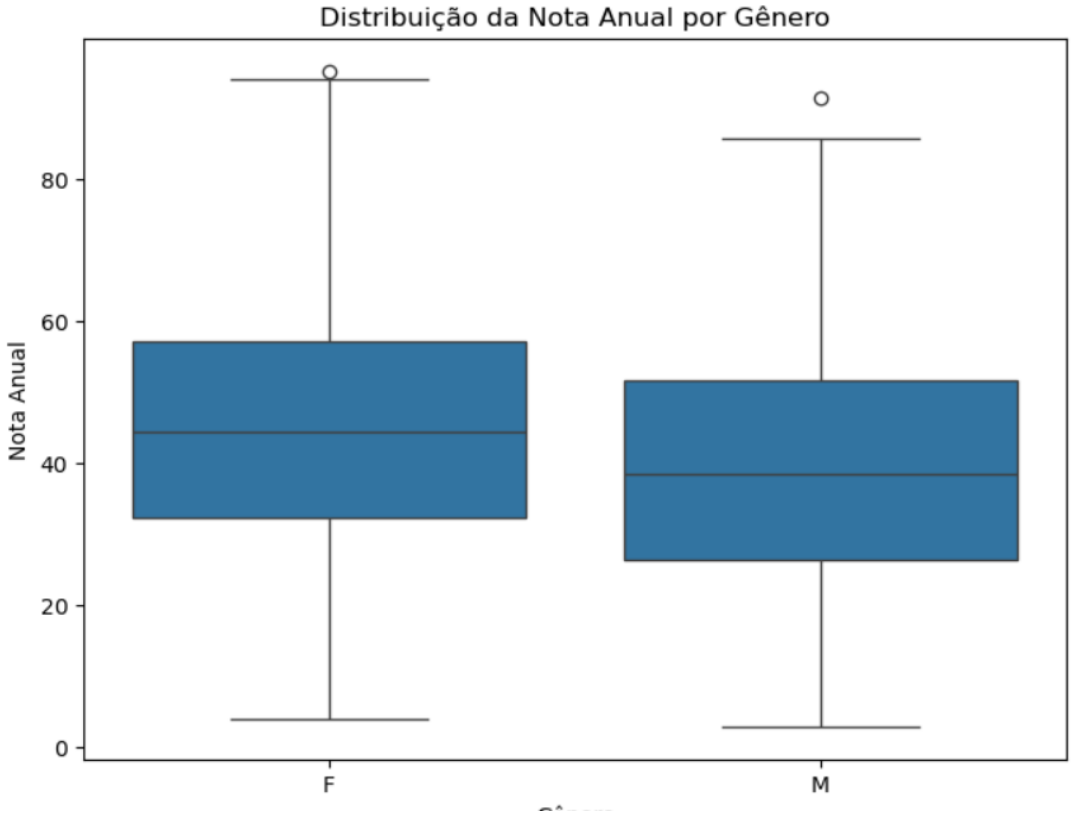

Figura 1 Distribuição da Nota Anual por Gênero

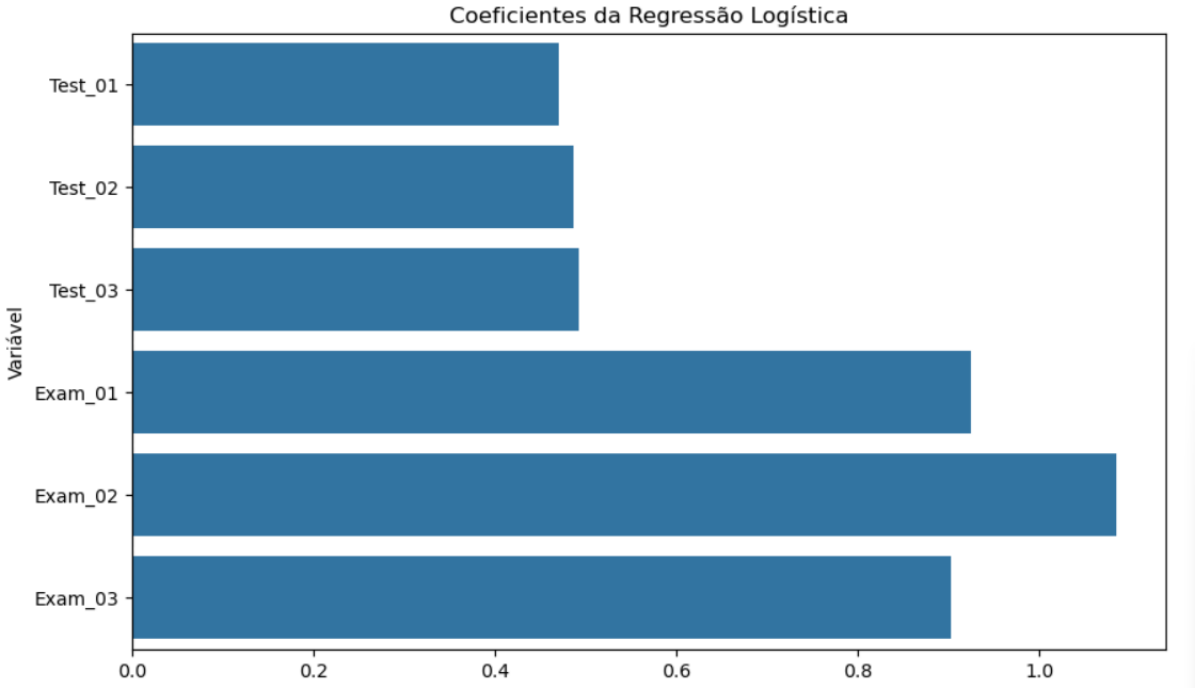

Figura 2 Coeficientes da Regressão Logística

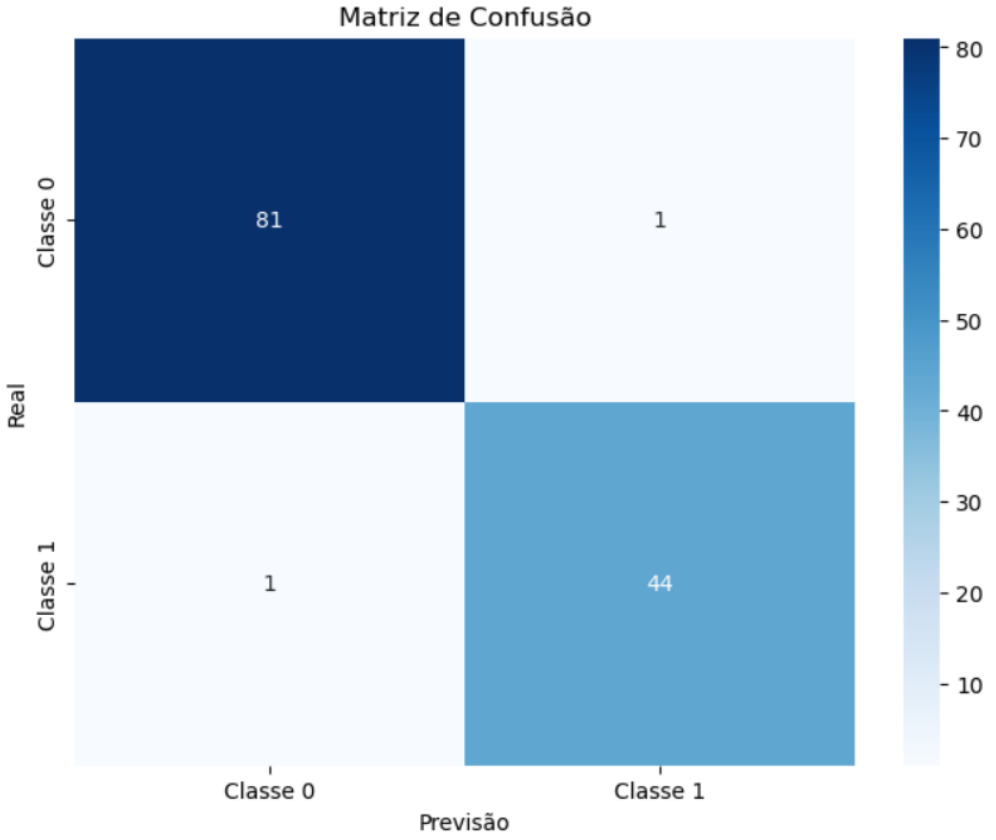

Figura 3 Matriz de Confusão

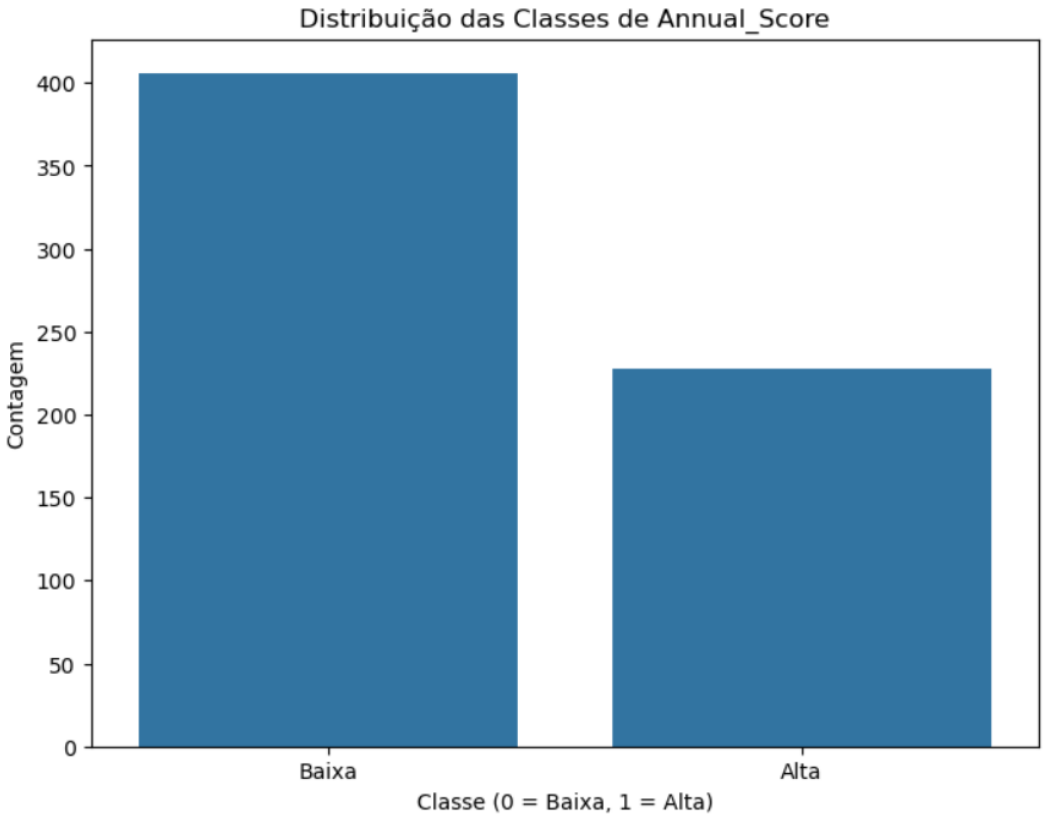

Figura 4 Distribuição das Classes de Annual_Score

RESULTADO MÉTRICAS E METODOLOGIAS

Primeiramente encontrar uma base de dados solida, simples e de fácil compreensão. Muitos dados e planilhas divulgas são extremamente técnicas com metodologias de alto complexidade, dificultando o entendimento e consequentemente a forma de tratar com cada situação, pois grande parte das informações são para profissionais pelas quais já tem um grau de familiaridade, tanto para os dados como ferramentas de aprendizado de máquina e seus códigos.

O resultado foi de maneira tranquila até chegar no objetivo proposto. Mais uma vez, os dados obtidos são de fácil compreensão, facilitando no tratamento e na tomada de decisão.

Busquei encontrar o coeficiente na regressão logística na qual indicam como cada variável preditora (X)i, influencia a probabilidade de a variável dependente assumir um determinado valor. Encontrado assim os coeficientes positivos, quando tem o efeito positivo na probabilidade de um desempenho maior, e um coeficiente negativo indica que, a variável preditora tem um efeito negativo na probabilidade.

A matriz de confusão avalia o desempenho do modelo. Ela mostra a contagem de previsões corretas e incorretas.

Foi feita também a interpretação da acurácia, medir o desempenho do modelo, onde determina previsões corretas de acordo com o total de previsões realizadas. Uma acurácia alta indica que a maioria das previsões estão corretas.

Essas foram as métricas em aprendizado de máquina utilizados para tratar, filtrar e analisar cada dados, possibilitando uma análise completa na tomada de decisão.