Relatório Final - Projeto de Machine Learning (AP1)

1. Introdução

Este projeto tem como objetivo aplicar técnicas estatísticas e modelos de Machine Learning utilizando a linguagem R. O dataset utilizado é o "Students Performance in Exams", disponível no Kaggle. Com ele, buscamos compreender como variáveis sociodemográficas influenciam o desempenho escolar de estudantes, e também prever se o aluno participou de um curso preparatório com base em suas notas.

---

2. Dataset

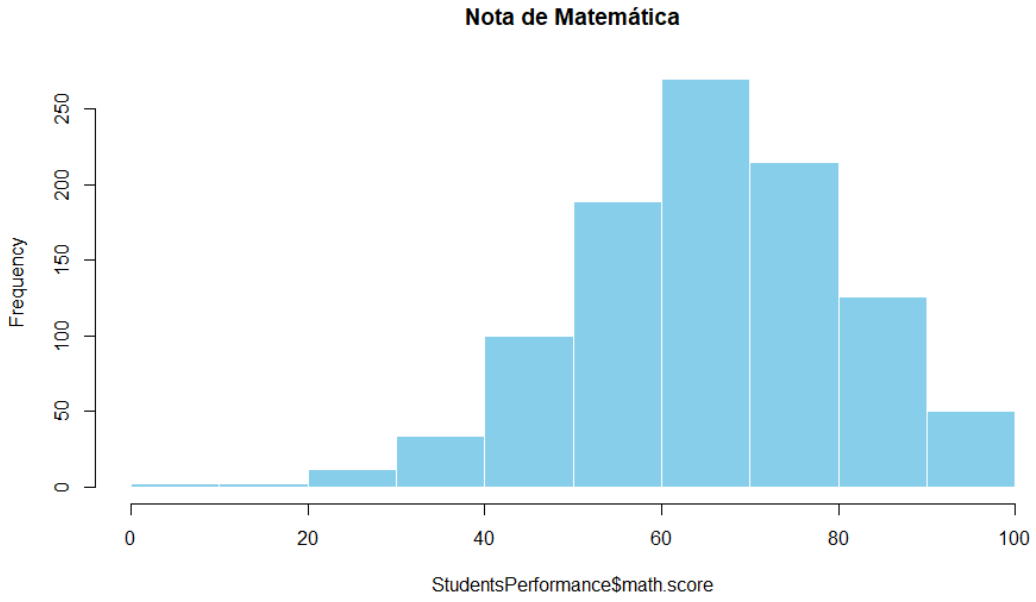
Fonte: Kaggle - Students Performance in Exams
Observações: 1000 linhas
Variáveis:

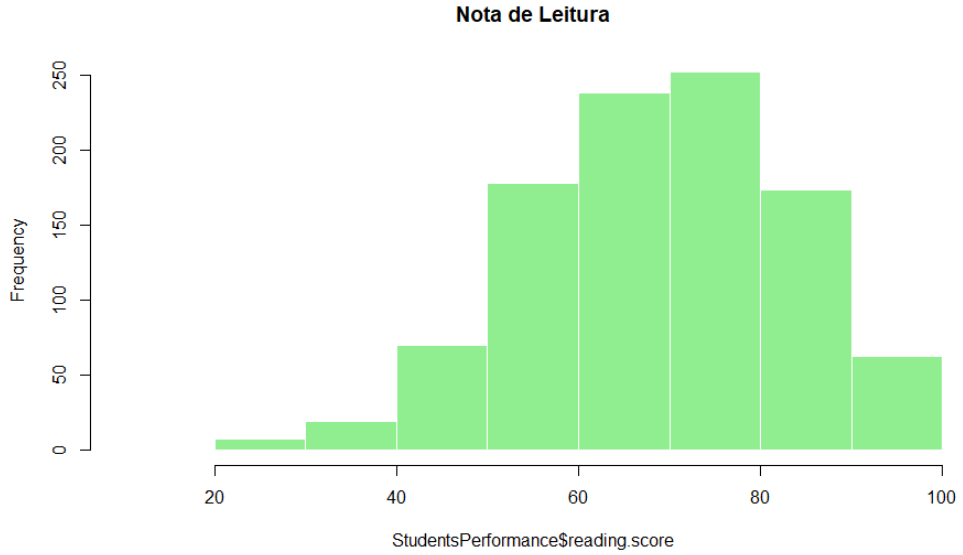
- Gênero
- Grupo étnico
- Nível de escolaridade dos pais
- Tipo de almoço
- Curso preparatório (binária)
- Notas em Matemática, Leitura e Escrita

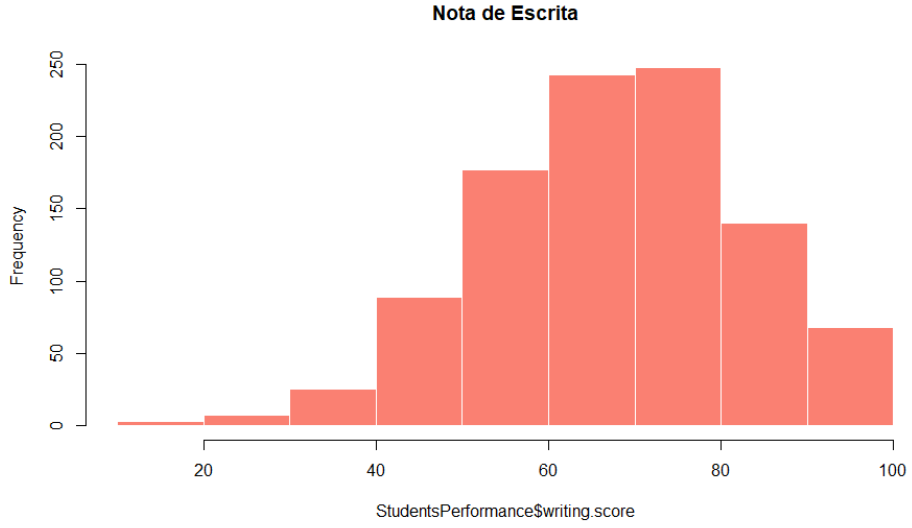
---

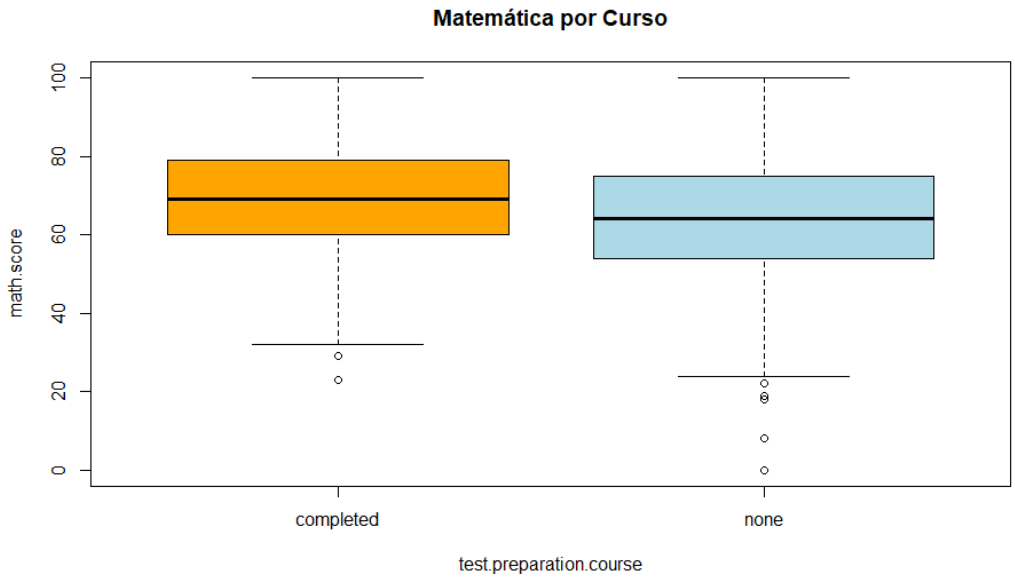
3. Análise Exploratória

- **Verificação de dados ausentes**: Nenhum valor ausente encontrado.
- **Estatísticas descritivas**: Notas variam de 0 a 100. Distribuições relativamente simétricas.
- **Visualizações**:


Nota de Matemática
Frequency
0
50
100
150
200
250
0
20
40
60
80
100
StudentsPerformance$math.score


Nota de Leitura
Frequency
0
50
100
150
200
250
20
40
60
80
100
StudentsPerformance$reading.score


Nota de Escrita
Frequency
0
50
100
150
200
250
20
40
60
80
100
StudentsPerformance$writing.score



---

4. Testes Estatísticos

4.1. Testes de Normalidade

Para avaliar a normalidade das variáveis numéricas (math.score, reading.score e writing.score), foi adotado o teste de Shapiro-Wilk. A escolha se baseia em sua reconhecida eficácia em amostras pequenas e moderadas, sendo ideal para conjuntos com menos de 5000 observações – como é o caso deste dataset, que possui 1000 linhas.

Resultados:

- **math.score:** Shapiro-Wilk ($p = 0.00015$) → não normal
- **reading.score:** Shapiro-Wilk ($p = 0.0001$) → não normal
- **writing.score:** Shapiro-Wilk ($p < 0.0001$) → não normal

Apesar das não-normalidades observadas, os modelos de regressão linear foram mantidos, considerando a robustez do método frente a leves desvios de normalidade e o tamanho razoável da amostra.

4.2. Correlação de Pearson

- **Matemática x Leitura:** $r = 0.817$ ($p < 0.001$)

- **Matemática x Escrita:** $r = 0.803$ ($p < 0.001$)
- **Leitura x Escrita:** $r = 0.955$ ($p < 0.001$)

---

5. Modelagem - Regressão Linear Multivariada

Objetivo:

Prever as três notas (matemática, leitura, escrita) com base nas variáveis categóricas:

- Gênero
- Grupo étnico
- Escolaridade dos pais
- Tipo de Almoço

Resultados:

Foram ajustados três modelos com lm(). Todos mostraram significância estatística de alguns preditores, especialmente **escolaridade dos pais**, **grupo étnico** e **tipo de almoço**. Os R^2 variaram entre 0.11 e 0.17, sugerindo que fatores adicionais também influenciam o desempenho.

Exemplo:

Para uma aluna do grupo C com pais que fizeram "some college" e almoço "standard", o modelo retorna:

- Nota prevista em Matemática: 62.9
- Nota prevista em Leitura: 72.1
- Nota prevista em Escrita: 72.1

---

6. Modelagem - Regressão Logística

Objetivo:

Prever a probabilidade de um aluno ter feito o curso preparatório com base em suas **notas**.

Variáveis utilizadas:

- Nota de Matemática
- Nota de Leitura

- Nota de Escrita

Resultados:

O modelo glm() indicou:

- Leitura tem efeito negativo (coef = -0.094)
- Escrita tem efeito positivo significativo (coef = 0.151)

Avaliação:

- **Acurácia**: 68.9%
- **Sensibilidade**: 84.1% | **Especificidade**: 41.6%
- **Classe prevista (exemplo)**: notas 75, 80, 78 → Prob. = 40.9% → Classe = none

---

7. API REST com Plumber

Foi desenvolvida uma API REST em R utilizando o pacote plumber.

Endpoints:

- **/predicao**: Recebe gênero, grupo étnico, escolaridade dos pais e tipo de almoço, e retorna as três notas previstas.
- **/classificacao**: Recebe as três notas e retorna a probabilidade de curso preparatório (e a classe prevista).

A API permite testar o modelo facilmente via navegador ou ferramentas como Postman.

---

8. Conclusão

O projeto mostrou como variáveis demográficas influenciam o desempenho escolar e como é possível prever participação em programas de apoio com base no desempenho. A publicação via API REST demonstra um passo em direção à produção de modelos prontos para consumo externo.