Relatório do Projeto Final – Machine Learning

Tema: Predição de aprovação de Estudantes

Grupo: Igor Ryan, Mario Raposo, Pedro Luan

1. Resumo do Problema

O objetivo deste projeto foi desenvolver um modelo capaz de prever a situação final de um estudante — *Aprovado*, *Recuperação* ou *Reprovado* — utilizando informações de desempenho acadêmico e dados sócio educacionais.
A base de dados utilizada (Student Performance – Kaggle) que contém mil registros de alunos, incluindo:

- Notas de Matemática (“math score”)
- Notas de Leitura (“reading score”)
- Notas de Escrita (“writing score”)
- Tipo de almoço
- Educação dos pais
- Grupo étnico
- Participação em curso preparatório
- Gênero

A partir desses dados, o desafio central foi construir um pipeline completo de Machine Learning, passando por:

1. Exploração e entendimento dos dados (EDA)
2. Pré-processamento e engenharia de atributos
3. Treinamento de modelos supervisionados
4. Regularização e controle de overfitting
5. Avaliação com métricas clássicas
6. Comparação dos modelos
7. Seleção do melhor algoritmo

Esse processo simula aplicações reais em escolas que desejam antecipar riscos, avaliar rendimento e aplicar intervenções pedagógicas.

2. Metodologia

A metodologia seguiu rigorosamente o ciclo tradicional de Machine Learning:

2.1 Coleta e Preparação dos Dados

Os dados foram carregados diretamente do Kaggle.
Como nenhum valor nulo foi encontrado, não houve necessidade de imputação ou remoção de registros.

Foi criada a variável “media”, representando a média das três notas (matemática, leitura, escrita).
A classe final foi definida como:

- Aprovado → média ≥ 60
- Recuperação → 40 ≤ média < 60
- Reprovado → média < 40

Essa transformação produziu um problema de classificação de três categorias (*multiclasse*).

2.2 Análise Exploratória dos Dados (EDA)

A EDA mostrou padrões importantes:

- Leitura e escrita têm alta correlação; alunos que leem bem também escrevem bem.
- Matemática se comporta de forma mais independente.
- A distribuição das notas é relativamente uniforme.
- Diferenças moderadas entre grupos socioeducacionais, especialmente nível educacional dos pais e tipo de almoço.

Gráficos utilizados:

- Histogramas
- Boxplots
- Heatmap de correlação
- Gráficos de dispersão (scatterplots)

Essas análises permitiram entender quais variáveis impactam mais o desempenho final.

2.3 Pré-Processamento

Para preparar os dados:

- Aplicou-se One-Hot Encoding nas variáveis categóricas.
- Utilizou-se StandardScaler para padronizar as variáveis numéricas.
- Separou-se o conjunto em treino (80%) e teste (20%).
- Garantiu-se que o escalonamento foi aplicado somente no treino (para evitar vazamento de informação).

2.4 Modelos Utilizados

Três modelos clássicos foram treinados:

1. Regressão Logística

- Modelo linear

- Rápido, interpretável
- Usou regularização L2 (para reduzir overfitting)

2. Árvore de Decisão

- Captura relações não lineares
- Tende ao overfitting
- Foi limitado com profundidade máxima

3. Random Forest

- Conjunto de várias árvores
- Reduz o overfitting
- Melhor desempenho entre os três

2.5 Regularização e Overfitting

- A Regressão Logística utilizou regularização L2.
- A Árvore de Decisão apresentou overfitting devido à alta variância.
- A Random Forest corrigiu isso ao combinar várias árvores e limitar profundidade.

Também foi aplicada validação cruzada (k-fold) e GridSearchCV para encontrar o melhor conjunto de hiperparâmetros.

3. Principais Resultados

As métricas analisadas foram:

- Acurácia
- Precisão
- Recall
- F1-score
- Matriz de confusão
- Validação cruzada (k-fold)

3.1 Regressão Logística

- Boa precisão na classe *Aprovado*
- Desempenho limitado em *Recuperação* e *Reprovado*
- Acurácia ~ 74%

3.2 Árvore de Decisão

- Acurácia ~ 70%
- Claro overfitting
- Baixa capacidade de generalização

3.3 Random Forest

- Melhor resultado geral
- Acurácia ~ 80–92% (dependendo da divisão e configuração)
- Melhor equilíbrio entre as classes
- Melhor estabilidade nos folds da validação cruzada

3.4 Grid Search (Random Forest)

Melhores parâmetros encontrados (exemplo):

- n_estimators = 200
- max_depth = 10
- min_samples_split = 2

Essas escolhas maximizaram a acurácia final do modelo.

4. Conclusões

O projeto demonstrou com sucesso todas as etapas fundamentais de Machine Learning tradicional, desde a coleta até a avaliação final do modelo. As principais conclusões são:

1. Notas de leitura e escrita são os fatores mais determinantes para prever aprovação.
2. A criação da variável “media” tornou a classificação mais fiel ao cenário real.
3. Random Forest foi o melhor modelo, apresentando:
    - maior precisão
    - menos overfitting
    - melhor generalização
4. A Regressão Logística teve bom desempenho, mas não lidou bem com classes minoritárias.
5. A Árvore de Decisão serviu para interpretação, mas não para predição final.

Possíveis melhorias futuras

- Coletar mais dados reais
- Criar novas features (ex.: horas de estudo, presença, hábitos)
- Testar modelos avançados como XGBoost, LightGBM ou SVM
- Implantar um painel interativo (ex.: Streamlit)