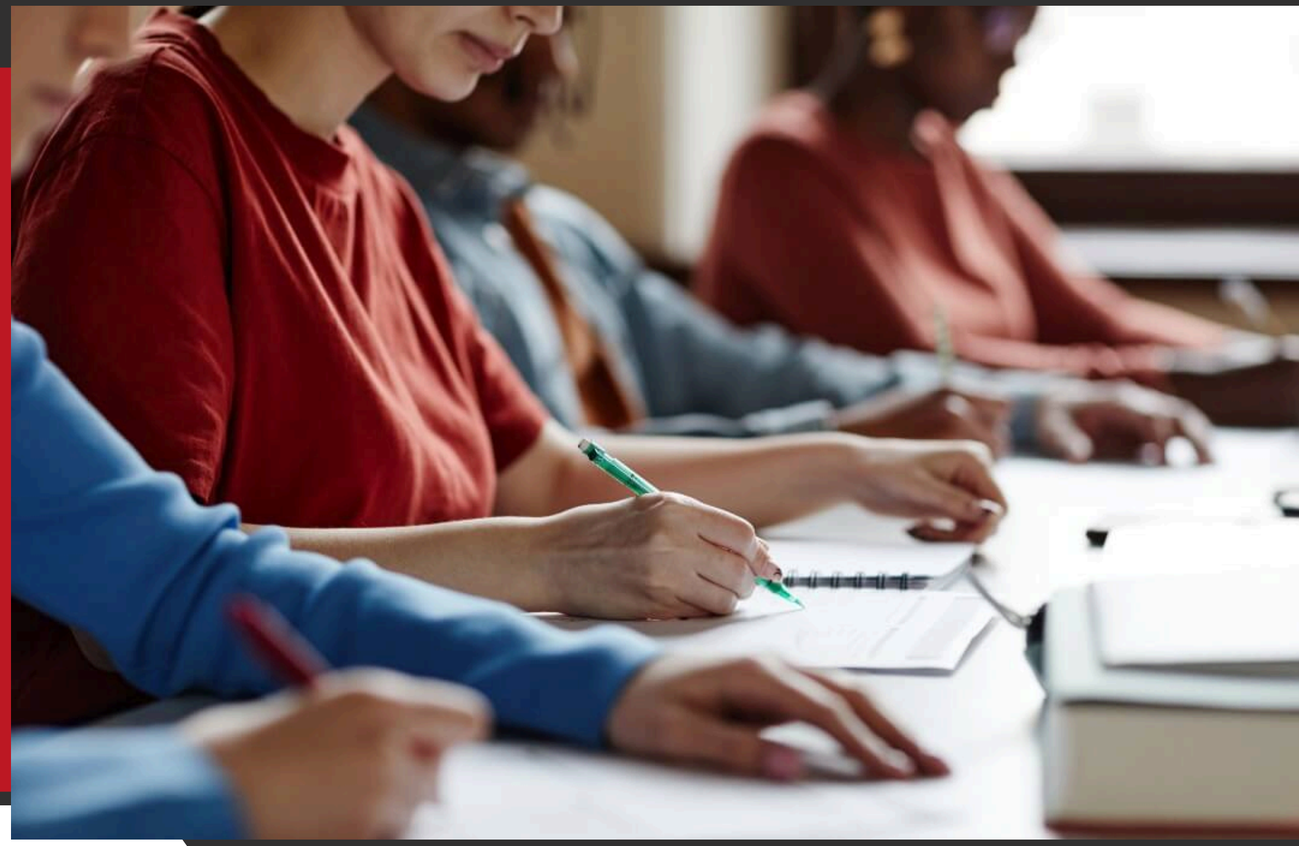
Desempenho Escolar

Fábio Leonor 202200269, Maria Oliveira 202200600 e Sara Sampaio 202200639



Introduçao

Objetivo do projeto:
Desenvolver dois modelos preditivos para análise do desempenho escolar:

- Modelo de regressão linear para previsão numérica da nota final (G3)
- Modelo de classificação binária (regressão logística) para determinar aprovação (G3 ≥ 10) vs. reprovação

Fonte dos dados:
UCI Machine Learning Repository
(Dataset ID 320 - "Student Performance")

Variável alvo:
• G3 (nota final do aluno, escala 0-20)
Características analisadas (33 variáveis após pré-processamento):
✓ **Demográficas:** sexo, idade, localização da escola
✓ **Sociais:** educação dos pais, situação conjugal familiar
✓ **Académicas:** notas anteriores (G1, G2), faltas, tempo de estudo
✓ **Comportamentais:** consumo de álcool, saúde, vida social

Análise e pré processamento de dados

Estatísticas Básicas:

- 649 amostras (estudantes)
- 30 características originais (mistas: numéricas e categóricas)
- Sem dados em falta (verificação realizada)
- Codificação one-hot aplicada às variáveis categóricas.
- Standardização aplicada às variáveis numéricas.
- Justificação: A regressão logística e linear são sensíveis à escala dos dados

Objetivo: Garantir dados completos.

Estratégia:

01

```
# Tratar dados ausentes
if X.isnull().values.any():
    X = X.fillna(X.mean(numeric_only=True))
    X = X.fillna(X.mode().iloc[0])
else:
    print("Sem dados em falta.")
```

Numéricos → Preenchidos com a média da coluna

Categóricos → Preenchidos com a moda (valor mais frequente)

02

```
# Codificação one-hot
X = pd.get_dummies(X, drop_first=True)
```

Transformação: Converte categóricas (ex: **sexo**) em colunas binárias (0/1).

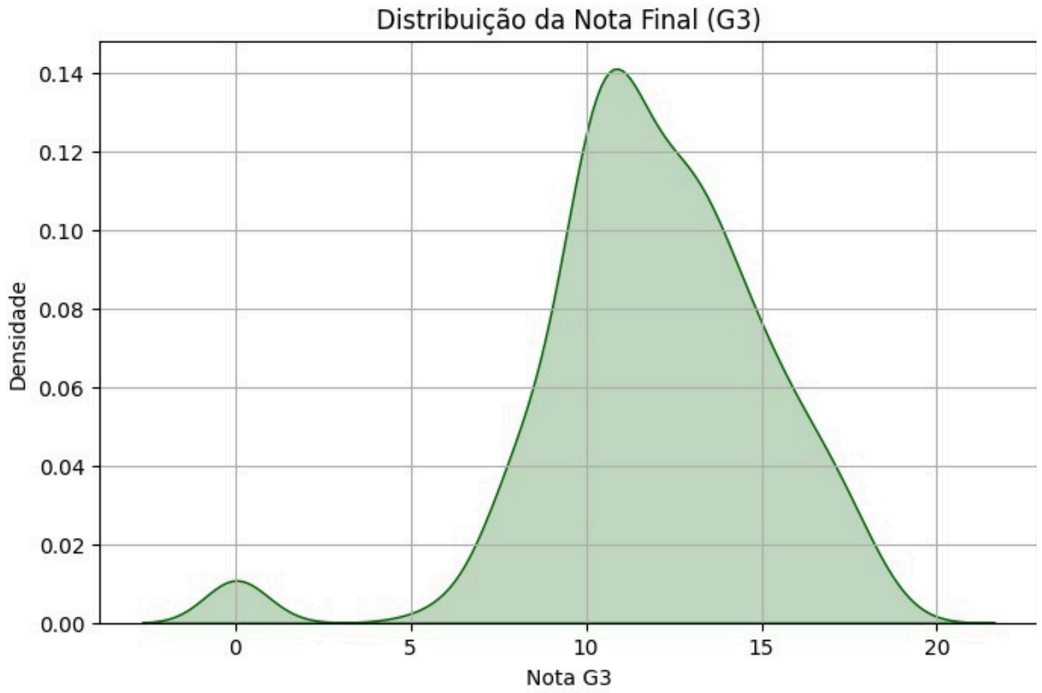
03

```
# Padronização
num_cols = X.select_dtypes(include=[np.number]).columns
scaler = StandardScaler()
X[num_cols] = scaler.fit_transform(X[num_cols])
```

Efeito: Normaliza variáveis numéricas para μ=0, σ=1.

Distribuição da Nota Final (G3)

01



Compreender a distribuição da variável alvo (G3), identificando padrões, picos e limiares naturais no desempenho dos alunos.

A distribuição revela uma concentração significativa de alunos com nota entre 10 e 12.
A existência de um limiar claro nos 10 valores (mínimo para aprovação) destaca-se.
A distribuição é assimétrica, o que pode afetar modelos lineares.

A assimetria da distribuição justificou a normalização dos dados com o StandardScaler para a regressão linear.
A concentração de valores em torno de 10 reforçou a decisão de binarizar a variável G3 (Aprovado/Reprovado) para aplicar regressão logística.

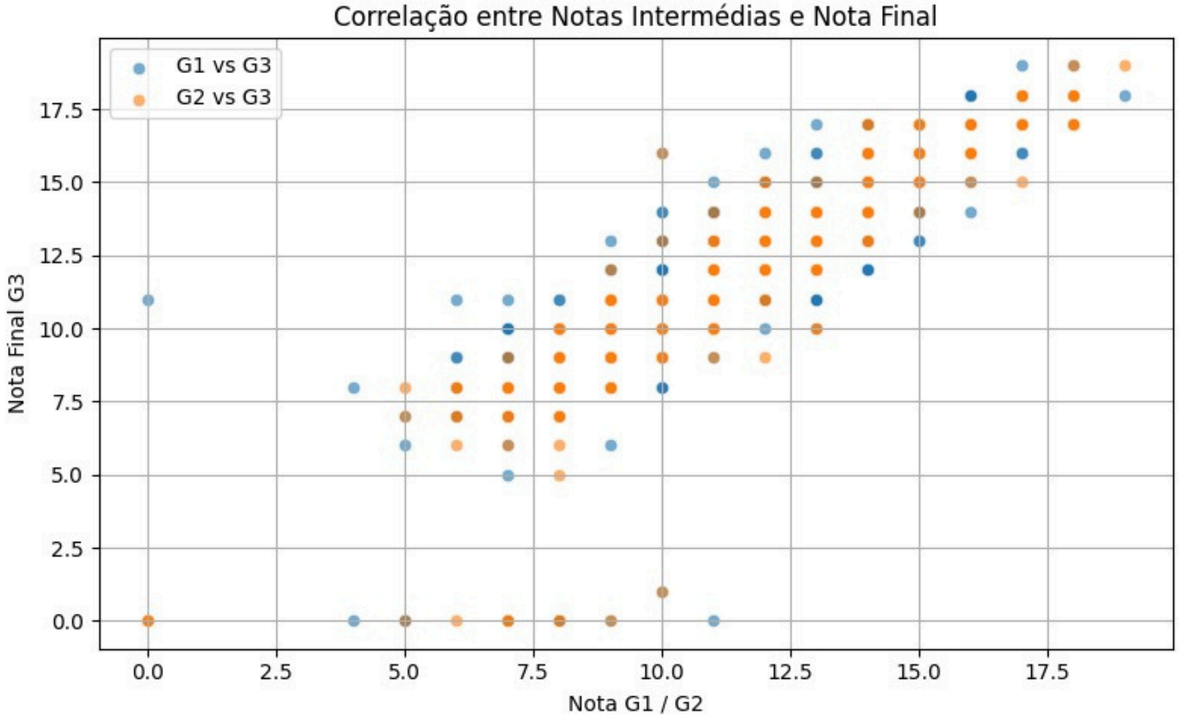
Correlação entre G1 / G2 e G3

02

Avaliar se as notas intermédias (G1 e G2) têm uma relação significativa com a nota final (G3).

Existe uma correlação muito forte entre G1/G2 e G3 (superior a 0.8).
A relação é praticamente linear, o que reforça a escolha de modelos de regressão linear.
Estes atributos explicam grande parte da variabilidade de G3.

G1 e G2 foram selecionadas como features principais nos modelos de regressão.
Considerou-se a possibilidade de redundância, que pode ser explorada em futuras otimizações (ex. remoção de G1 ou regularização).



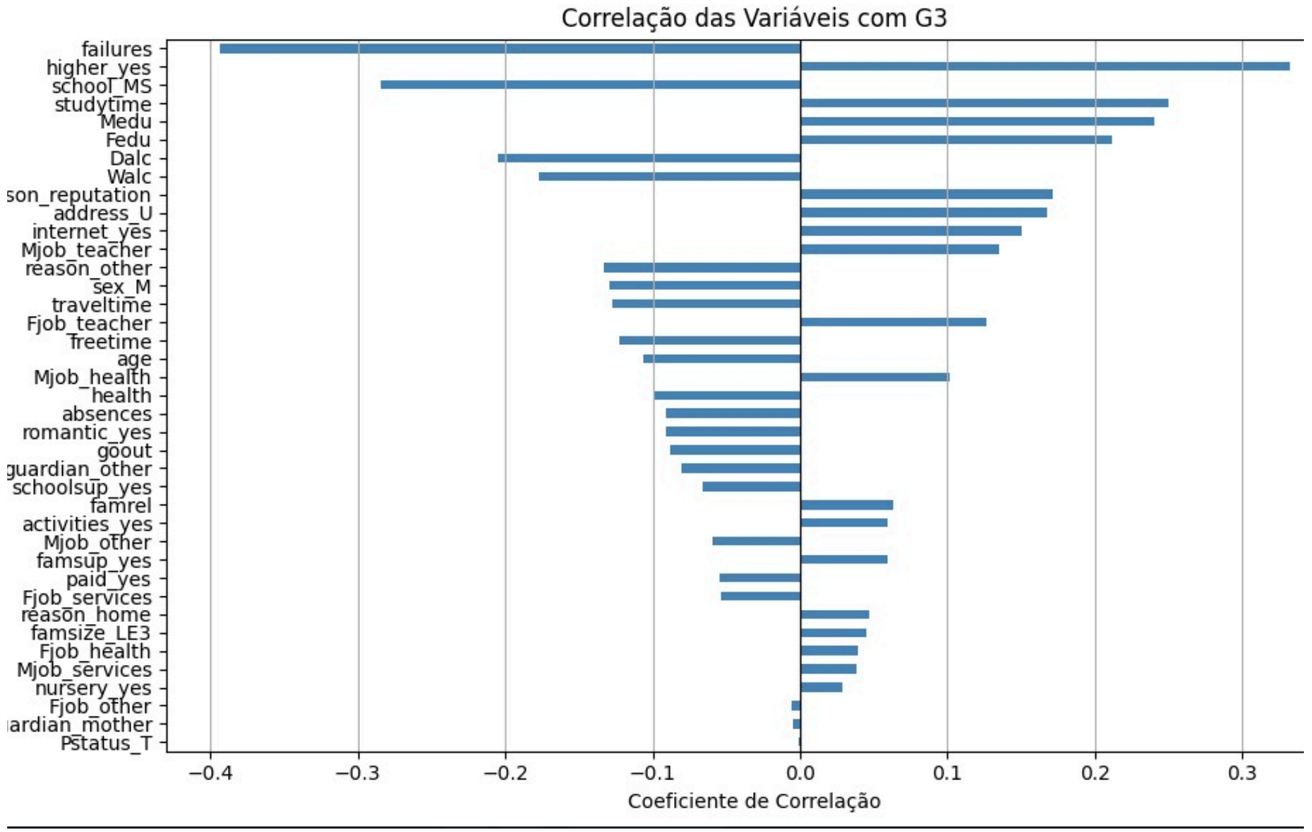
Correlação das Variáveis com G3

03

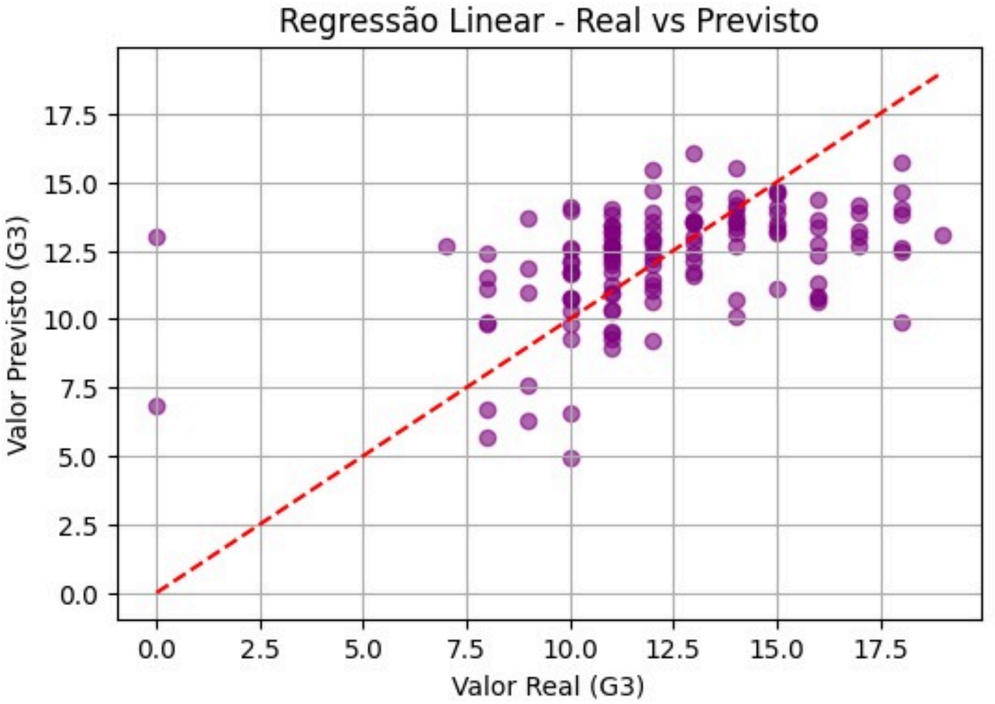
Identificar quais variáveis têm maior impacto na nota final (G3), orientando a escolha das variáveis preditoras.

Failures apresenta a maior correlação negativa com G3 — reprovações anteriores estão fortemente associadas a notas mais baixas.
higher_yes, studytime, Medu, Fedu, entre outras, mostram correlação positiva significativa.

A análise guiou a seleção das variáveis mais relevantes para os modelos.
Contribuiu para reduzir ruído e aumentar a eficácia dos modelos de regressão e classificação.
Identificou possíveis redundâncias e oportunidades para redução de dimensionalidade.



Regressão Linear



Objetivo do modelo:

- Prever a nota final dos alunos (G3) como uma variável numérica contínua, com base em características escolares, sociais e comportamentais.

Modelo:

- Regressão Linear (scikit-learn)

Conclusões e Observações:

- O modelo ajustou-se bem aos dados, especialmente devido à forte correlação entre G3, G1 e G2.
- As previsões alinham-se com os valores reais para a maioria dos alunos.
- A variabilidade dos dados pode ser explicada em grande parte pelas variáveis selecionadas.
- Algumas discrepâncias extremas podem dever-se a fatores não modelados (ex. motivação, saúde mental, ambiente familiar).

Regressão Logística

Objetivo:

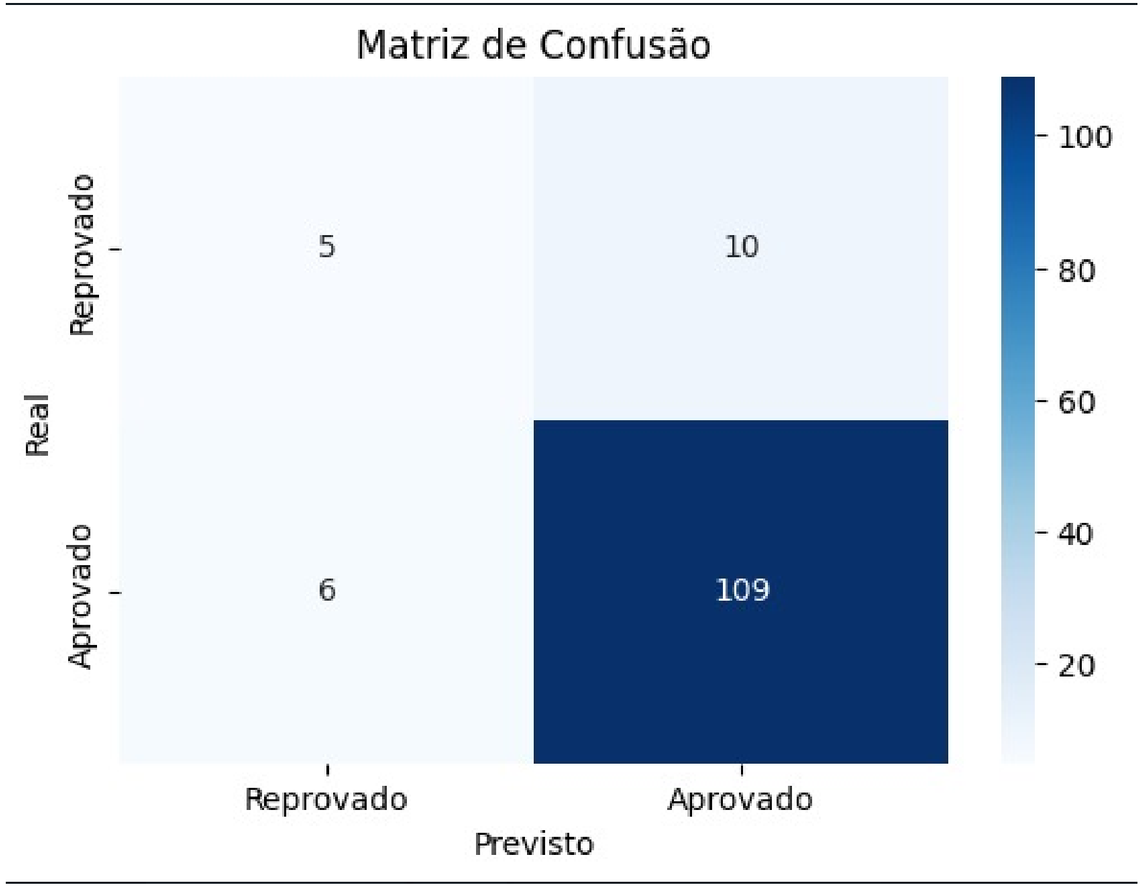
- Classificar os alunos como Aprovado (G3 ≥ 10) ou Reprovado (G3 < 10).

Modelo Utilizado:

- Regressão Logística (scikit-learn, max_iter=1000)

Métricas de Avaliação: Acurácia: 84.26%

- Matriz de Confusão: Mostra verdadeiros positivos/negativos e erros de classificação.



Resultados:

- O modelo teve bom desempenho na distinção entre aprovados e reprovados.
- Maior número de erros em alunos próximos da nota 10.
- Ideal para situações onde se pretende prever aprovação com base em histórico e perfil.

Discussão de Resultados

01

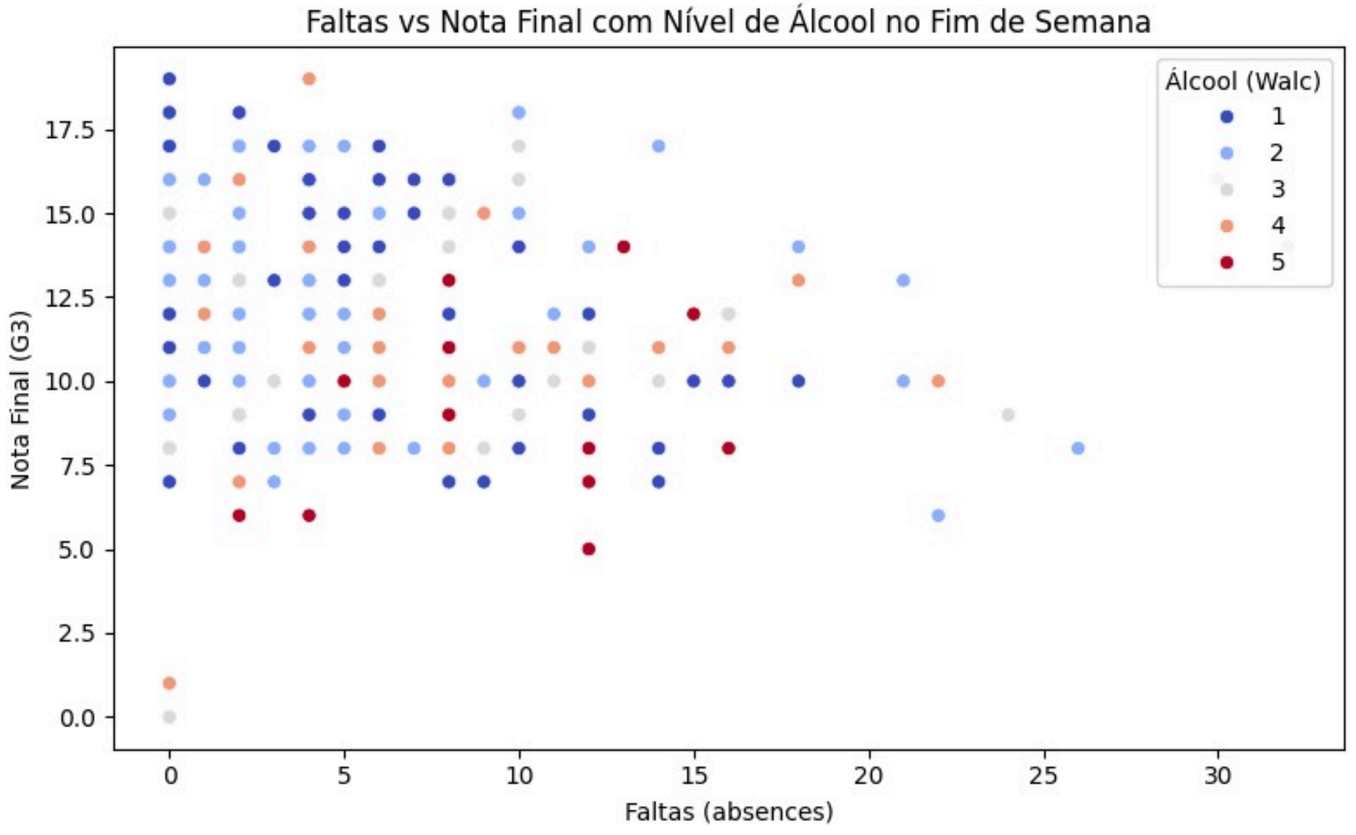
O que aprendemos com os modelos: G1 e G2 são os melhores preditores do sucesso final. Fatores como faltas (absences), tempo de estudo (studytime) e apoio familiar têm impacto relevante.

02

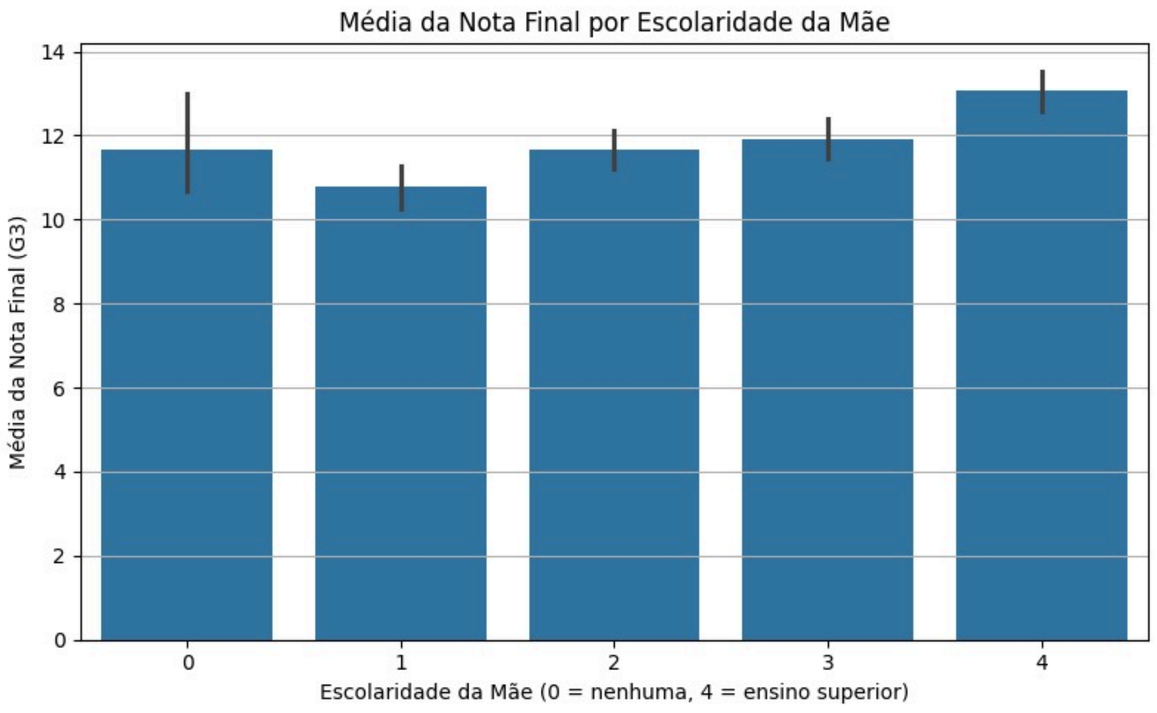
Qual modelo é mais adequado? Regressão Linear: Melhor para prever a nota exata (G3). Regressão Logística: Ideal para prever aprovação (decisão binária).

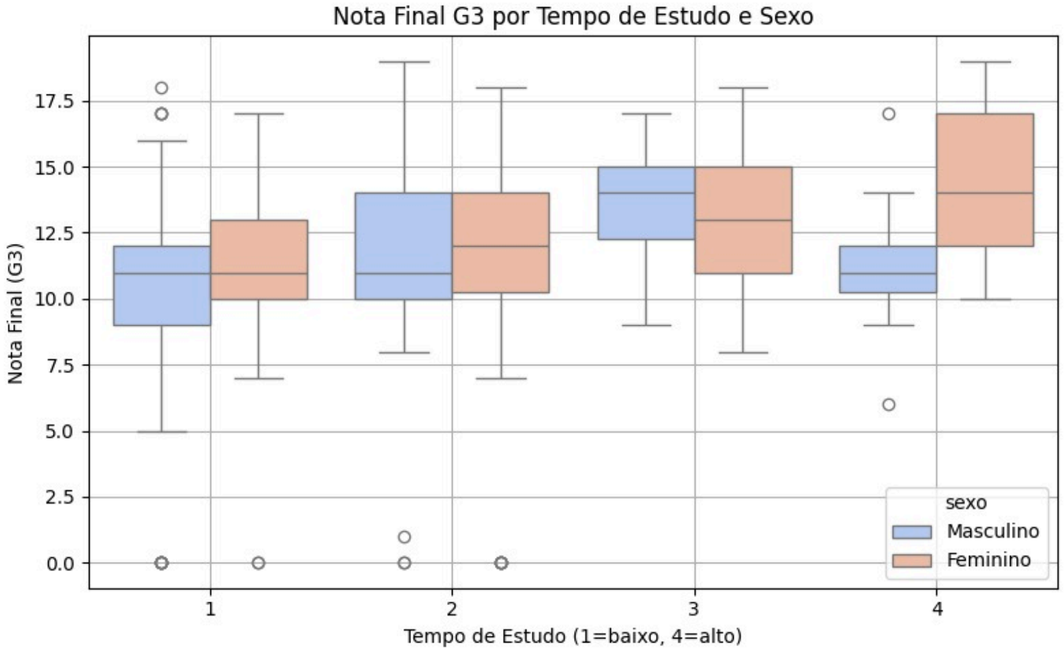
03

Limitações: Modelos lineares assumem relações simples e podem não capturar efeitos complexos.
Dados qualitativos como motivação, ambiente familiar ou saúde não estão representados.
Overfitting com muitas variáveis categóricas (one-hot)



Relações Interessantes





Conclusão

Resultados Princpais

- Foram aplicados modelos de regressão linear e logística com sucesso.
- A regressão linear previu notas com boa precisão (MSE ≈ %).
- A regressão logística classificou eficazmente a aprovação (acurácia ≈ 84,26%).
- As variáveis G1, G2, e fatores como faltas, reprovações e apoio educativo foram cruciais.

Limitações

- Dados não capturam aspetos qualitativos (ex: motivação).
- Risco de overfitting (33 features pós pré-processamento)

Melhorias

- Testar modelos não-lineares (Random Forest, SVM, Redes Neuronais).
- Aplicar regularização (Lasso/Ridge) para evitar overfitting.
- Incluir mais dados qualitativos (motivação, suporte psicológico).
- Análise de sensibilidade e técnicas de feature selection.

Thank you very much!