UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luiza de Melo Nogueira

1. Introdução

Descrição do dataset informando suas características relevantes e informação do que se deve analisar neste neste.

2. Observações

Algum problema que aconteceu? Alguma observação? Algum imprevisto? Se não houve problemas, deixe em branco.

3. Resultados e discussão

Nesta seção deve-se descrever como foram as resoluções de cada questão. Crie sessões indicando a questão e discuta a implementação e resultados obtidos nesta. Explique o fluxograma do processo de cada questão, indicando quais processamentos são realizados nos dados. Sempre que possível, faça gráficos, mostre imagens, diagramas de blocos para que sua solução seja a mais completa possível. Discuta sempre sobre os números obtidos em busca de motivos de erros e acerto.

4. Conclusões

Os resultados esperados foram satisfeitos? Se não, qual o motivo? Qual a sua análise?

5. Próximos passos

Depois desse relatório, quais os próximos passos você sugere para este projeto?

Relatório Parte B — Projeto de Regressão com Dados Educacionais

---

Introdução

Este relatório apresenta a análise de um conjunto de dados educacionais com o objetivo de prever o desempenho final dos estudantes — medido pela nota G3. O dataset, intitulado Student Alcohol Consumption, foi obtido a partir do Kaggle e contém informações socioeconômicas, comportamentais e acadêmicas de alunos do ensino médio em Portugal.

*A tarefa central é aplicar técnicas de **regressão linear** e suas variações (Ridge e Lasso) para:*

- *Compreender quais variáveis mais influenciam a nota final (G3)*
- *Avaliar o desempenho preditivo de modelos diferentes*
- *Interpretar coeficientes e realizar seleção de atributos*

Descrição do Dataset

*O dataset contém **33 colunas** e **395 amostras**, com atributos que incluem:*

- *Informações pessoais e familiares (ex: age, sex, Medu, Fedu)*
- *Condições de estudo (studytime, failures, absences)*
- *Notas parciais: G1, G2*
- *Nota final: G3 (variável-alvo)*

Os dados foram padronizados antes da aplicação dos modelos, a fim de equilibrar escalas e melhorar o desempenho de regularizações como Lasso e Ridge.

Observações

Não houve problemas com valores ausentes. Um ajuste de caminho para carregar o arquivo .csv foi necessário devido à estrutura do KaggleHub, e o separador correto foi identificado como.

Resultados e Discussão

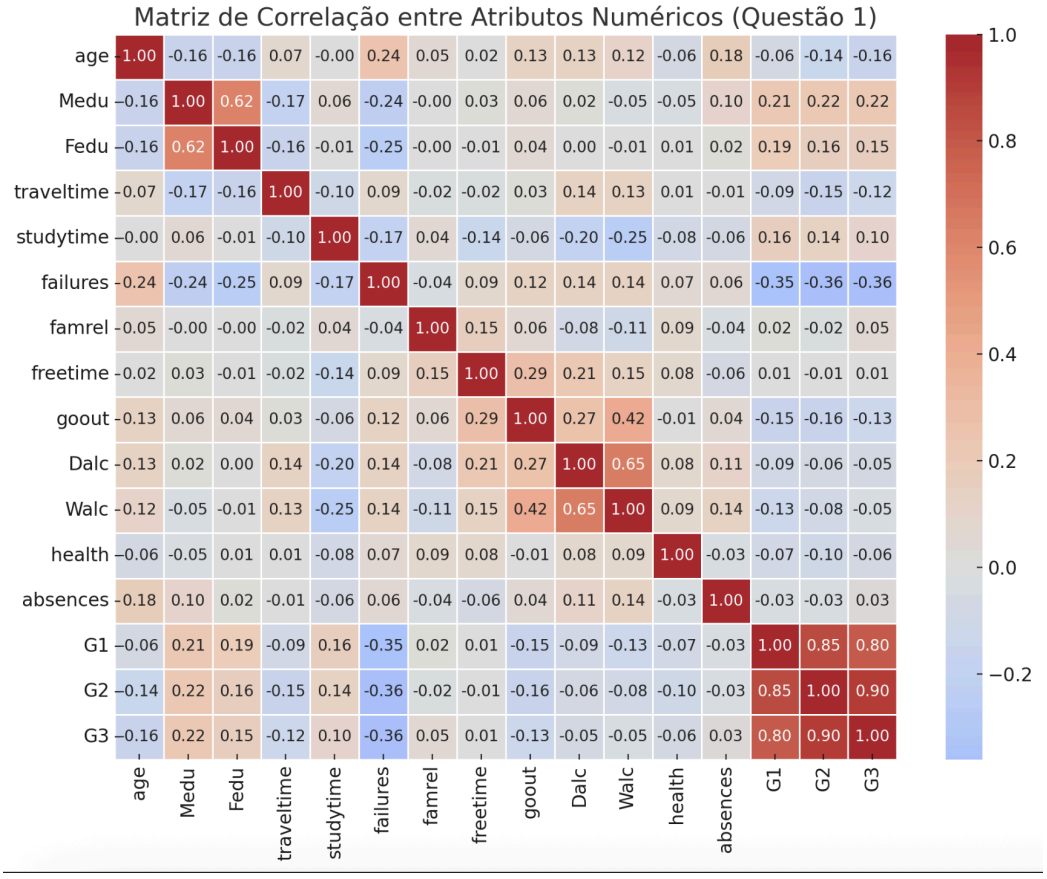
Questão 1 — Pré-processamento e Correlação

***Etapas**:*

- *Carregamento do dataset*
- *Verificação de valores ausentes (nenhum)*
- *Cálculo de correlação entre atributos e G3*
- *Padronização das colunas numéricas com StandardScaler*
- *Salvamento em regressao_ajustado.csv*

***Resultado**:*

Correlação mais alta com G3:
- G2: 0.90
- G1: 0.80



Questão 2 — Regressão Linear Simples

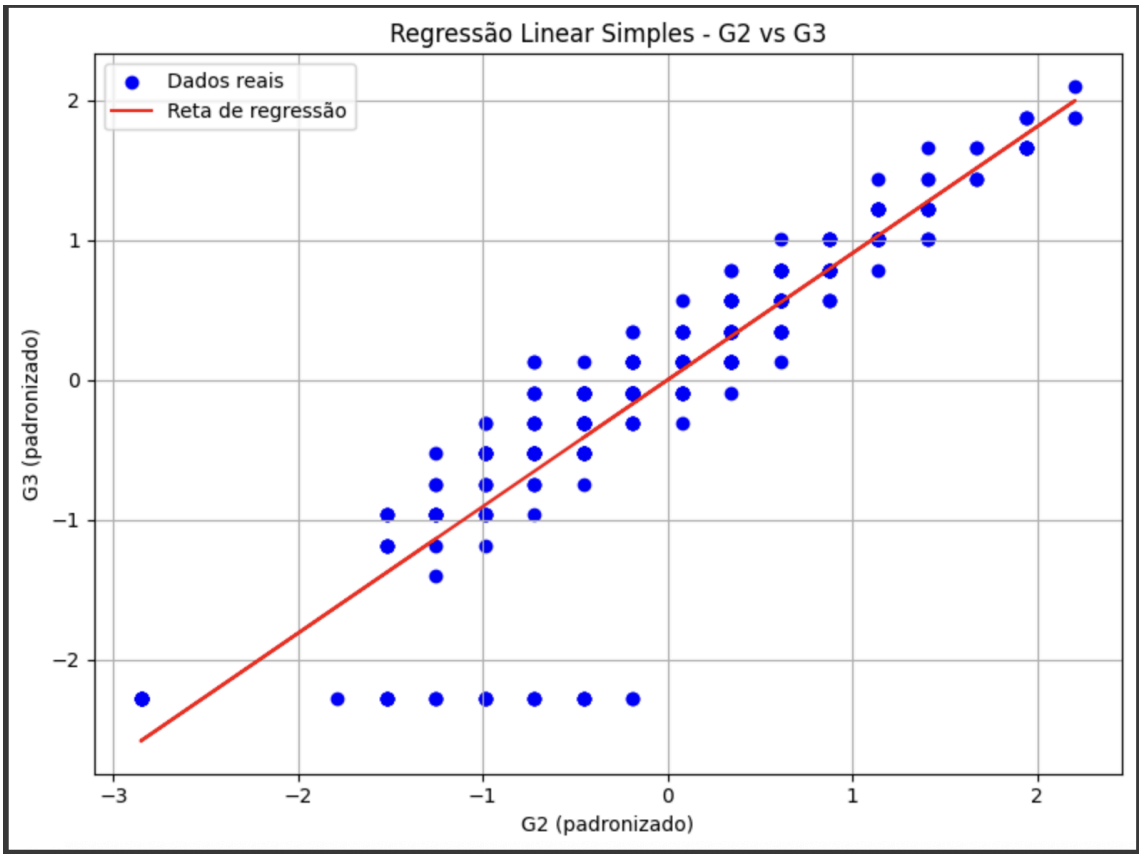
***Entrada**: apenas G2*

***Saída**: G3*

***Métricas**:*

- *Coeficiente angular: 0.9049*
- *Intercepto: ~0*
- *RMSE: 0.4257*
- *R^2: 0.8188*

***Gráfico**: A reta se ajusta muito bem aos dados. O valor de R^2 acima de 80% confirma a **força preditiva de G2 isoladamente**.*



Questão 3 — Linear vs Ridge vs Lasso

***Modelos comparados**:*

- *Regressão linear*
- *Ridge (alpha=1.0)*
- *Lasso (alpha=0.1)*

***Avaliação**: 5-fold cross-validation*

Métrica

Modelo	***RMSE médio***	***Rˆ2 médio***
Ridge	***0.4188***	***0.8077***
Regressão linear	***0.4191***	***0.8075***
Lasso	***0.4254***	***0.8034***

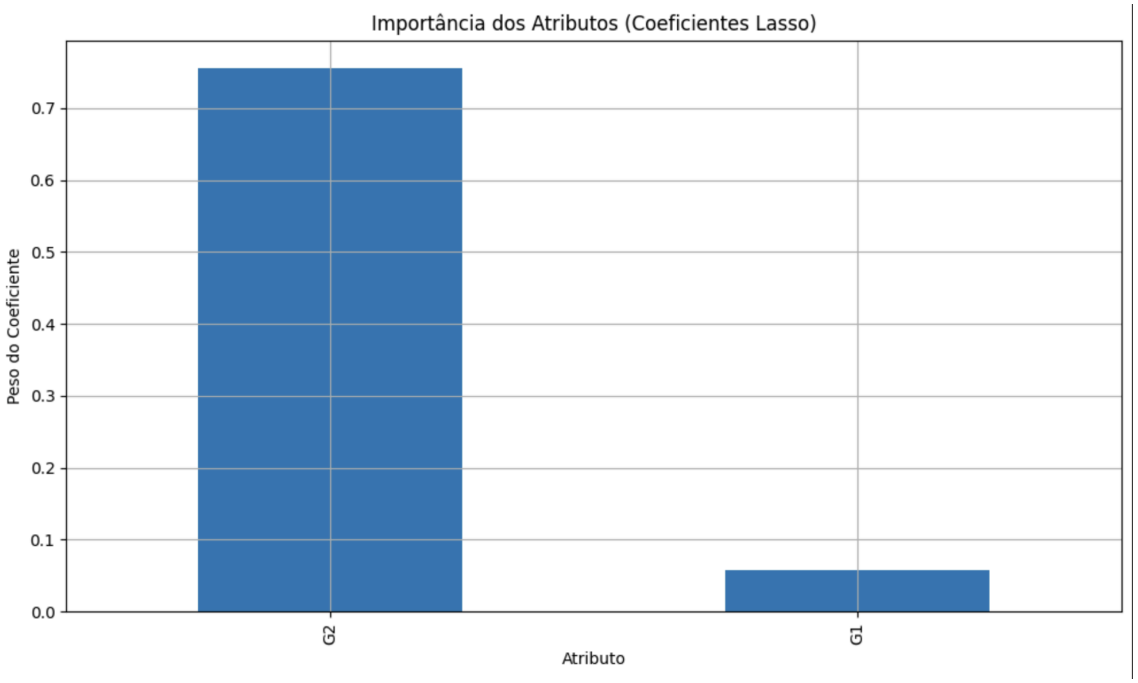
*O modelo **Ridge teve o melhor desempenho geral**, embora a diferença entre Ridge e Linear seja mínima. O Lasso apresentou desempenho levemente inferior.*

Questão 4 — Coeficientes e Seleção de Atributos com Lasso

***Coeficientes não-nulos**:*

G2 0.7561
G1 0.0572

***Gráfico**:*



O Lasso eliminou todas as variáveis exceto G2 e G1, reforçando o padrão já observado nas correlações: as notas anteriores são os melhores preditores da nota final.

Conclusões

- *Os **resultados foram satisfatórios**: os modelos identificaram corretamente que o desempenho anterior (G1, G2) é o principal fator de influência sobre a nota final G3.*
- *O **modelo Ridge** teve desempenho levemente superior em termos de robustez.*
- *O Lasso foi útil na **redução automática de atributos**, simplificando a interpretação.*

Próximos Passos

- *Explorar modelos não lineares (como Árvore de Regressão, Random Forest).*
- *Aplicar regressão polinomial para investigar curvaturas.*
- *Avaliar subgrupos de alunos (por idade, gênero, tempo de estudo).*
- *Utilizar GridSearchCV para encontrar o melhor alpha em Ridge/Lasso.*