RELATÓRIO DE ANÁLISE DE MINERAÇÃO DE DADOS: QUALIDADE DE VINHOS

105140 - EMILY L. BALESTRIN
TÓPICOS ESPECIAIS EM COMPUTAÇÃO I
UNIVERSIDADE REGIONAL INTEGRADA DO ALTO URUGUAI E DAS MISSÕES
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Link do repositório: https://github.com/EmilyBalestrin/WineQuality_Dataset

1. Nome do Dataset e Origem

O conjunto de dados utilizado neste projeto é o **"Wine Quality - Red Wine"**. Ele foi obtido do repositório de Machine Learning da UCI (Universidade da Califórnia, Irvine), uma fonte confiável e amplamente utilizada para datasets de pesquisa. O dataset contém 1.599 registros e 12 colunas, atendendo aos requisitos mínimos de 100 registros e 5 colunas estipulados no projeto.

2. Breve Descrição dos Resultados Obtidos

O objetivo do projeto foi aplicar técnicas de mineração de dados para classificar a qualidade de vinhos tintos com base em seus atributos físico-químicos. Para isso, o dataset foi submetido a um processo de análise exploratória, pré-processamento e, por fim, à aplicação de três algoritmos de classificação distintos: **Random Forest, Logistic Regression e Decision Tree**.

Os principais resultados da análise foram:

- **Análise Exploratória:** A qualidade dos vinhos no dataset foi avaliada em uma escala de 3 a 8. Para simplificar a classificação, foi criada uma nova coluna binária chamada good_quality, onde vinhos com nota maior ou igual a 7 foram considerados de "boa qualidade" (valor 1) e os demais de "qualidade ruim" (valor 0).
- **Correlação de Atributos:** O mapa de calor (heatmap) das correlações revelou que o álcool, os sulfatos, o ácido cítrico e a acidez volátil são os atributos com maior impacto na qualidade do vinho.
- **Desempenho dos Modelos:** Os três modelos de classificação foram treinados e avaliados com base em métricas como acurácia, precisão, revocação e F1-Score. O

modelo **Random Forest** apresentou o melhor desempenho geral, com uma acurácia de **90%** e um F1-Score de **0.60**. A Regressão Logística teve o menor desempenho entre os três.

A seguir, um print comparativo com os resultados de cada modelo:

Figura 1: Print comparativo entre os modelos

```
Comparativo de Desempenho dos Modelos:

               Modelo  Acurácia  Precisão  Revocação  F1-Score
        Random Forest    0.9000    0.7273     0.5106    0.6000
Logistic Regression    0.8594    0.5500     0.2340    0.3284
        Decision Tree    0.8719    0.5714     0.5106    0.5393

Process finished with exit code 0
```

Fonte: Autor, 2025

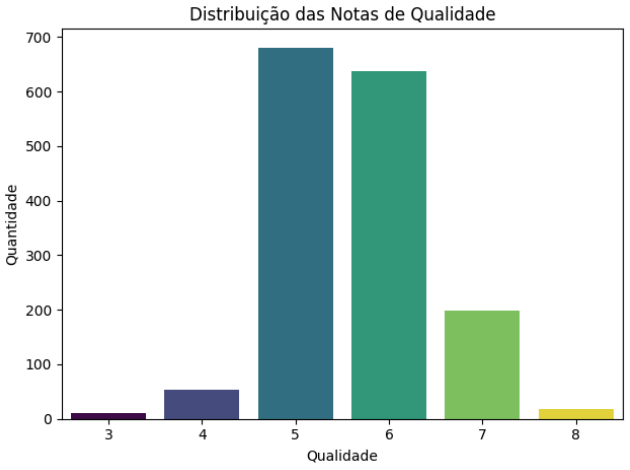
3. Gráficos e Relatórios Gerados

Conforme solicitado, seguem os principais gráficos gerados pelo software para a visualização e análise dos dados. Estes gráficos foram salvos na pasta /results do projeto.

Distribuição da Qualidade dos Vinhos

Este gráfico de contagem mostra a distribuição das notas de qualidade (de 3 a 8) no dataset. A maior parte dos vinhos se concentra nas notas 5 e 6.

Figura 2: Gráfico de Distribuição das Notas de Qualidade

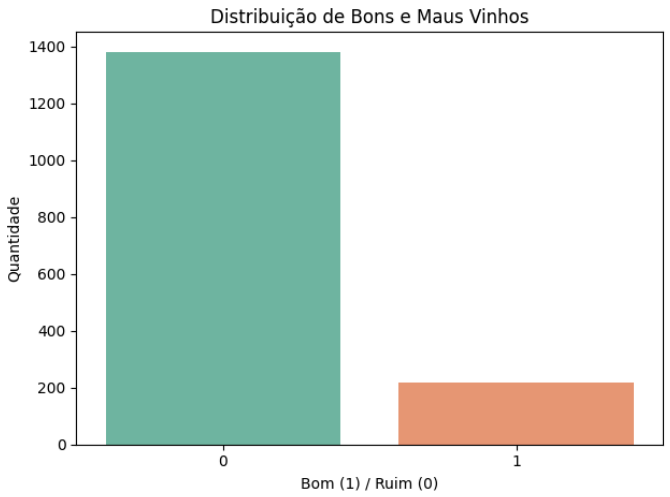


Fonte: Autor, 2025

Distribuição de Vinhos de "Boa Qualidade"

Este gráfico mostra a distribuição da variável binária good_quality. Nota-se que o dataset é desbalanceado, com uma quantidade significativamente maior de vinhos classificados como de qualidade ruim (0).

Figura 2: Gráfico de Distribuição de Bons e Maus Vinhos

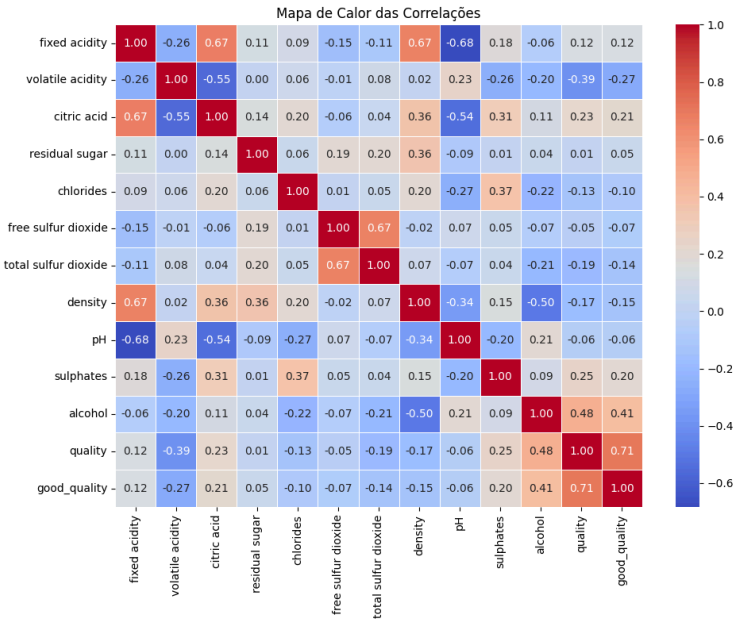


Fonte: Autor, 2025

Heatmap de Correlação

O mapa de calor ilustra a correlação entre todas as variáveis do dataset. Cores mais quentes (próximas de 1) indicam uma correlação positiva forte, enquanto cores mais frias (próximas de -1) indicam uma correlação negativa.

Figura 3: Gráfico de Mapa de Calor das Correlações

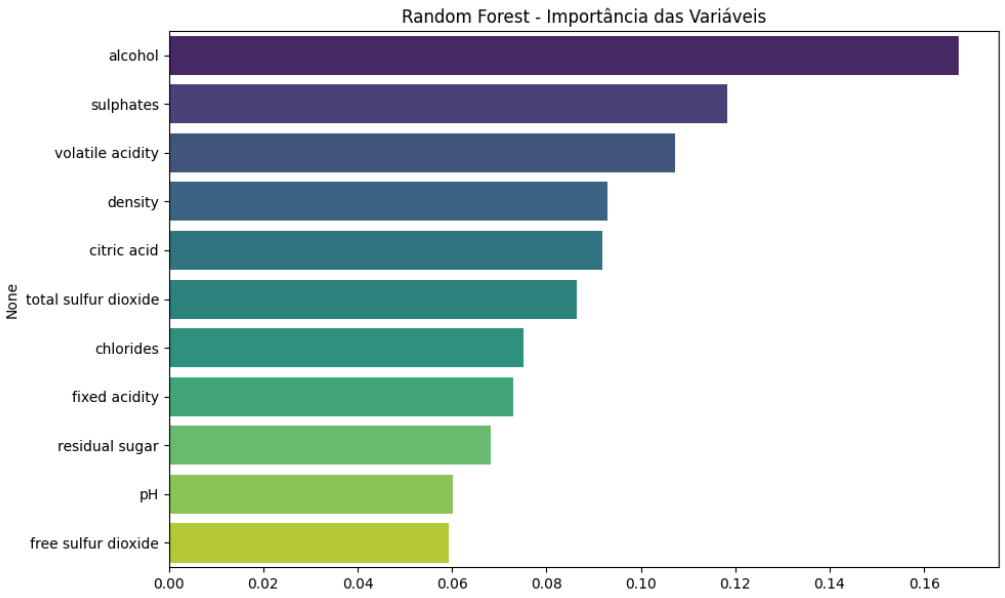


Fonte: Autor, 2025

Importância das Variáveis (Random Forest)

Este gráfico de barras mostra a importância de cada atributo na determinação da qualidade do vinho, segundo o modelo Random Forest. O álcool é, de longe, o atributo mais influente.

Figura 4: Gráfico de Importância de Variáveis

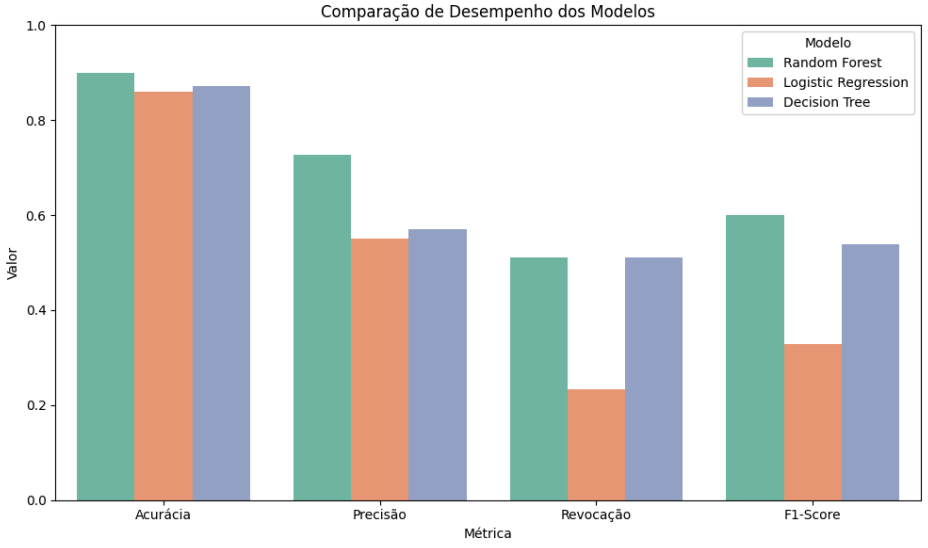


Fonte: Autor, 2025

Gráfico Comparativo de Desempenho dos Modelos

Este gráfico de barras compara o desempenho dos três modelos com base nas métricas de avaliação.

Figura 5: Gráfico de Comparação de Desempenho dos Modelos



Fonte: Autor, 2025

4. Publicação do Projeto

O projeto, incluindo o código-fonte, o dataset, os resultados e este relatório em PDF, foi disponibilizado publicamente em um repositório no GitHub, conforme os requisitos. O arquivo README.md no repositório contém as instruções detalhadas para a reprodução da análise.

5. Referências

CORTEZ, P. et al. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, v. 47, n. 4, p. 547-553, 2009.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

THE PANDAS DEVELOPMENT TEAM. pandas-dev/pandas: Pandas. Zenodo, 2020. Disponível em: https://doi.org/10.5281/zenodo.3509134. Acesso em: 28 jun. 2025.

WASKOM, M. L. et al. seaborn. Zenodo. Disponível em: https://doi.org/10.5281/zenodo.592845. Acesso em: 28 jun. 2025.