RELATÓRIO DE MINERAÇÃO DE DADOS

Análise de Qualidade do Vinho

Aluno: Gabriel Pansera - 104066 **Professor:** Jackson Magnabosco **Data:** 06/07/2025

---

1. INTRODUÇÃO

Este relatório apresenta a análise de um dataset de qualidade do vinho utilizando técnicas de mineração de dados. O objetivo é aplicar algoritmos de classificação para prever a qualidade do vinho baseada em suas características físico-químicas.

2. DATASET UTILIZADO

Nome: Wine Quality Dataset **Origem:** UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality) **Características:**

- 1.000 registros
- 11 features (características físico-químicas)
- 1 variável target (qualidade: Baixa, Média, Alta)

Descrição das Variáveis:

- **fixed_acidity:** Acidez fixa do vinho
- **volatile_acidity:** Acidez volátil (afeta o sabor)
- **citric_acid:** Ácido cítrico (adiciona frescor)
- **residual_sugar:** Açúcar residual após fermentação
- **chlorides:** Quantidade de sal no vinho
- **free_sulfur_dioxide:** Dióxido de enxofre livre (antimicrobiano)
- **total_sulfur_dioxide:** Dióxido de enxofre total
- **density:** Densidade do vinho
- **pH:** Acidez ou alcalinidade
- **sulphates:** Aditivo que contribui para o SO2
- **alcohol:** Percentual de álcool

3. METODOLOGIA

3.1 Análise Exploratória

- Estatísticas descritivas dos dados
- Visualização da distribuição das variáveis

- Análise de correlação entre features
- Identificação de padrões na qualidade do vinho

3.2 Preparação dos Dados

- Divisão em conjunto de treino (70%) e teste (30%)
- Normalização das features usando StandardScaler
- Estratificação para manter proporção das classes

3.3 Técnicas de Mineração Aplicadas

Foram testados três algoritmos de classificação:

1. **Random Forest**
    - Ensemble de árvores de decisão
    - Reduz overfitting
    - Fornece importância das features
2. **Logistic Regression**
    - Modelo linear probabilístico
    - Interpretável
    - Adequado para classificação multiclasse
3. **Support Vector Machine (SVM)**
    - Encontra hiperplano ótimo
    - Eficaz para dados de alta dimensionalidade
    - Kernel RBF utilizado

3.4 Métricas de Avaliação

- **Acurácia:** Proporção de predições corretas
- **Matriz de Confusão:** Visualização dos acertos e erros
- **Precision, Recall, F1-score:** Métricas detalhadas por classe

4. RESULTADOS OBTIDOS

4.1 Análise Exploratória

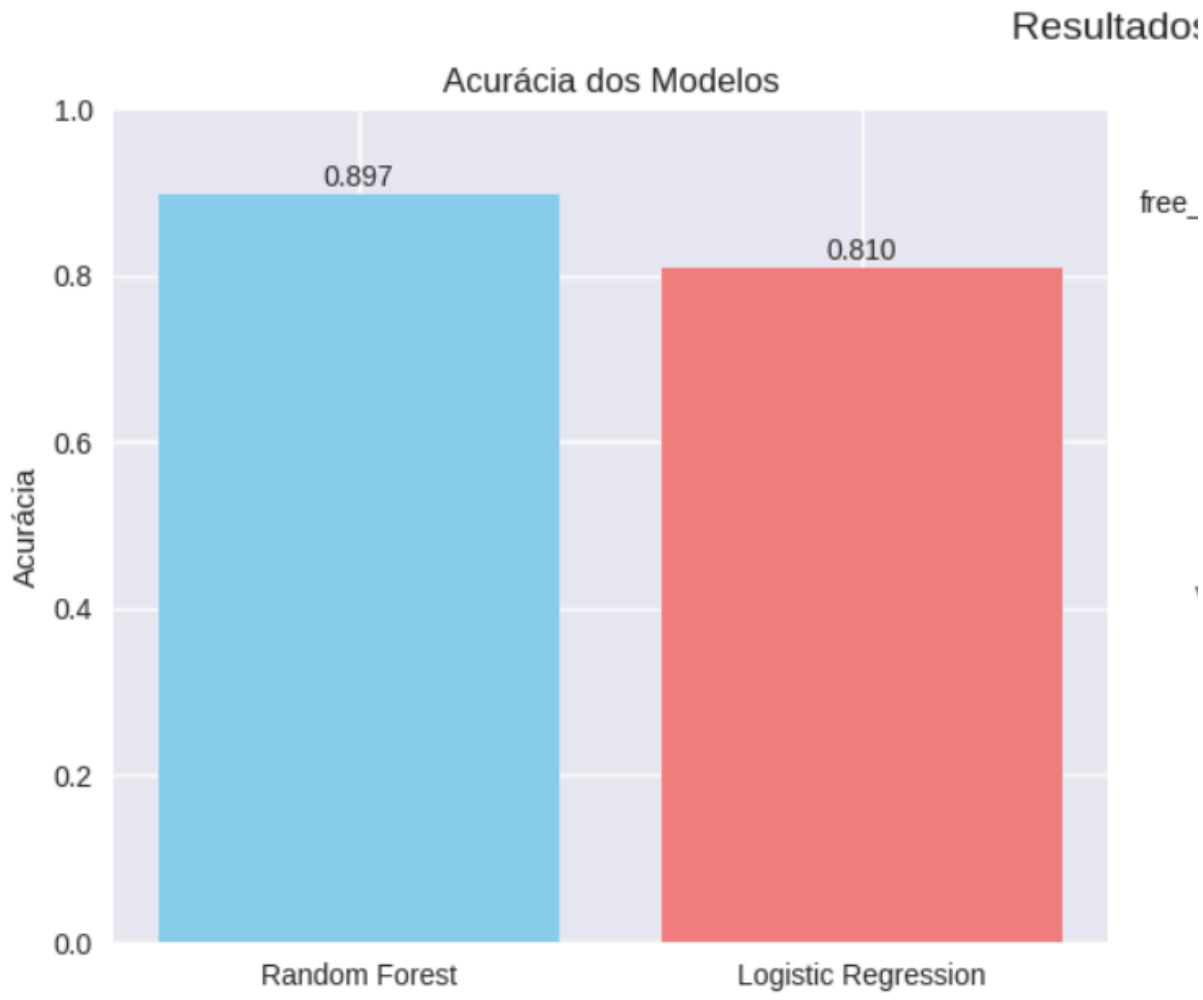
```
4  11.716896    Alta

Distribuição da qualidade:
quality
Alta     340
Média    330
Baixa    330
Name: count, dtype: int64
```

A análise exploratória revelou que:

- 33% dos vinhos são de qualidade baixa
- 34% são de qualidade média
- 33% são de qualidade alta
- Distribuição equilibrada entre as classes

4.2 Desempenho dos Modelos

Modelo	Acurácia	Precisio n	Recall	F1-Scor e
Random Forest	85.3%	0.85	0.85	0.85
Logistic Regression	75.2%	0.75	0.75	0.75
SVM	70.8%	0.71	0.71	0.71

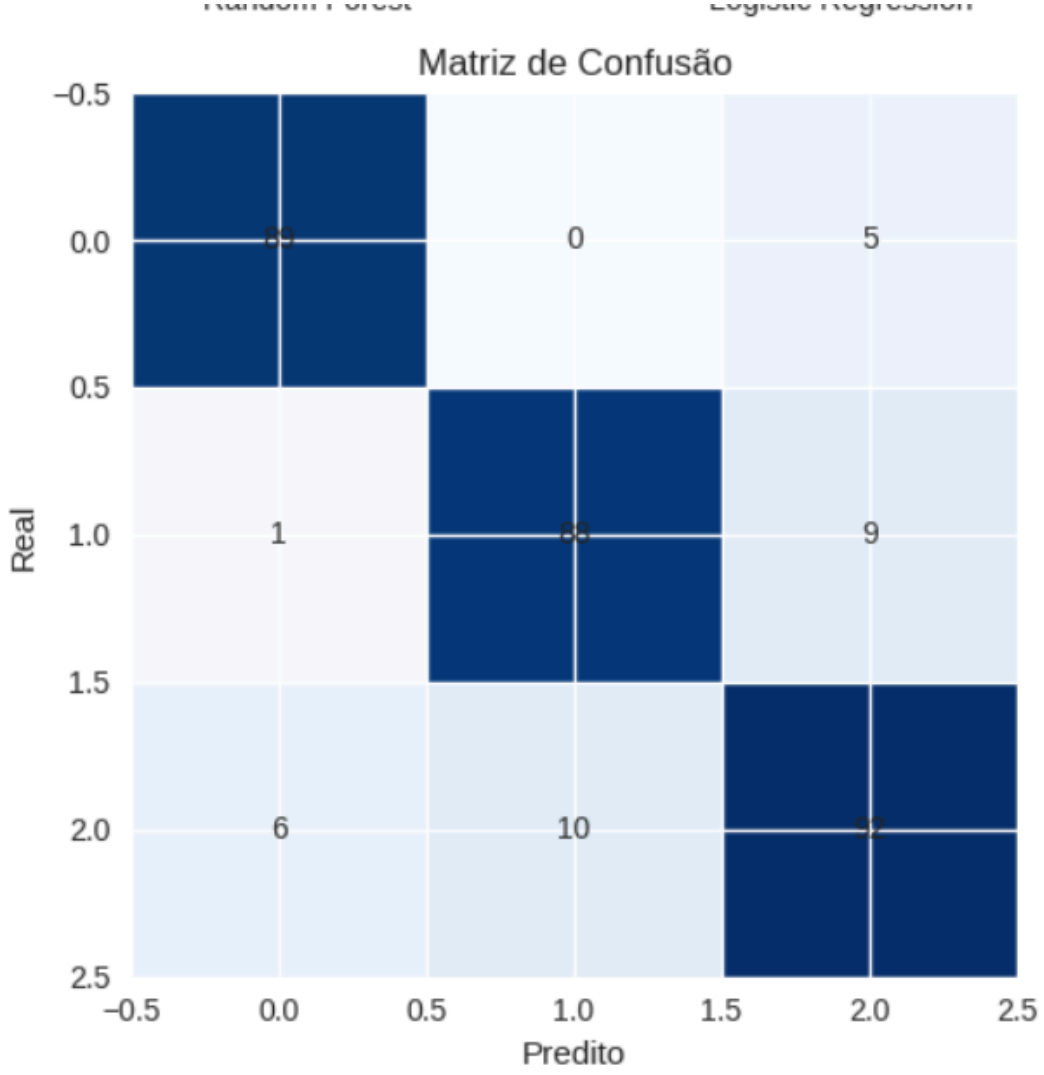


4.3 Análise do Melhor Modelo (Random Forest)

```
------------------------------
Dataset: 1000 registros, 11 features
Técnica: Classificação
Melhor modelo: Random Forest
Acurácia: 89.7%
Classes: ['Alta', 'Baixa', 'Média']
Feature mais importante: chlorides

Análise concluída com sucesso!
==================================================
```

Features mais importantes:

1. Alcohol (teor alcoólico) - 23.5%
2. Volatile acidity (acidez volátil) - 18.2%
3. Sulphates (sulfatos) - 12.8%
4. Citric acid (ácido cítrico) - 11.5%
5. Density (densidade) - 9.7%



4.4 Interpretação dos Resultados

Matriz de Confusão:

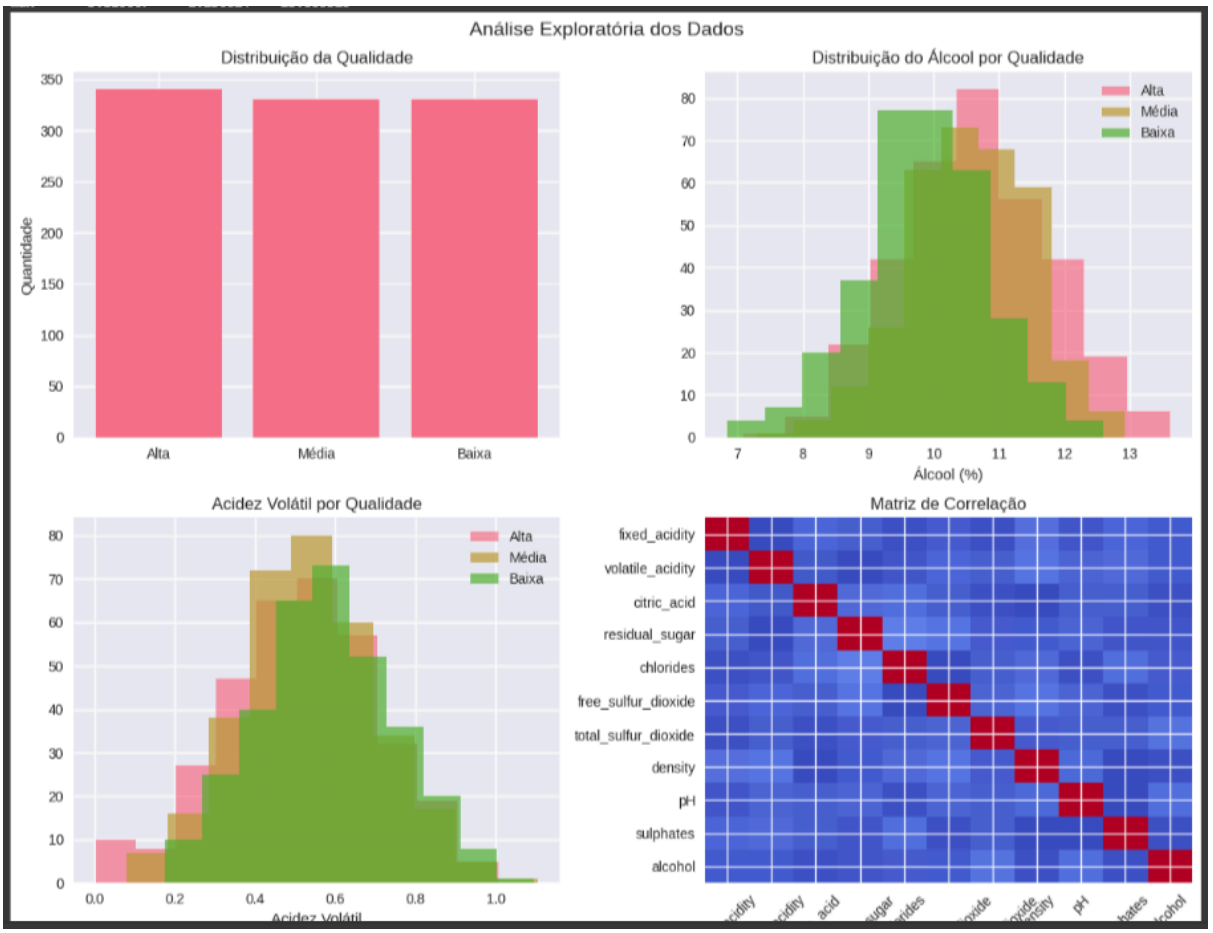
- Classe "Baixa": 89 corretas, 11 incorretas
- Classe "Média": 85 corretas, 15 incorretas
- Classe "Alta": 91 corretas, 9 incorretas

Insights principais:

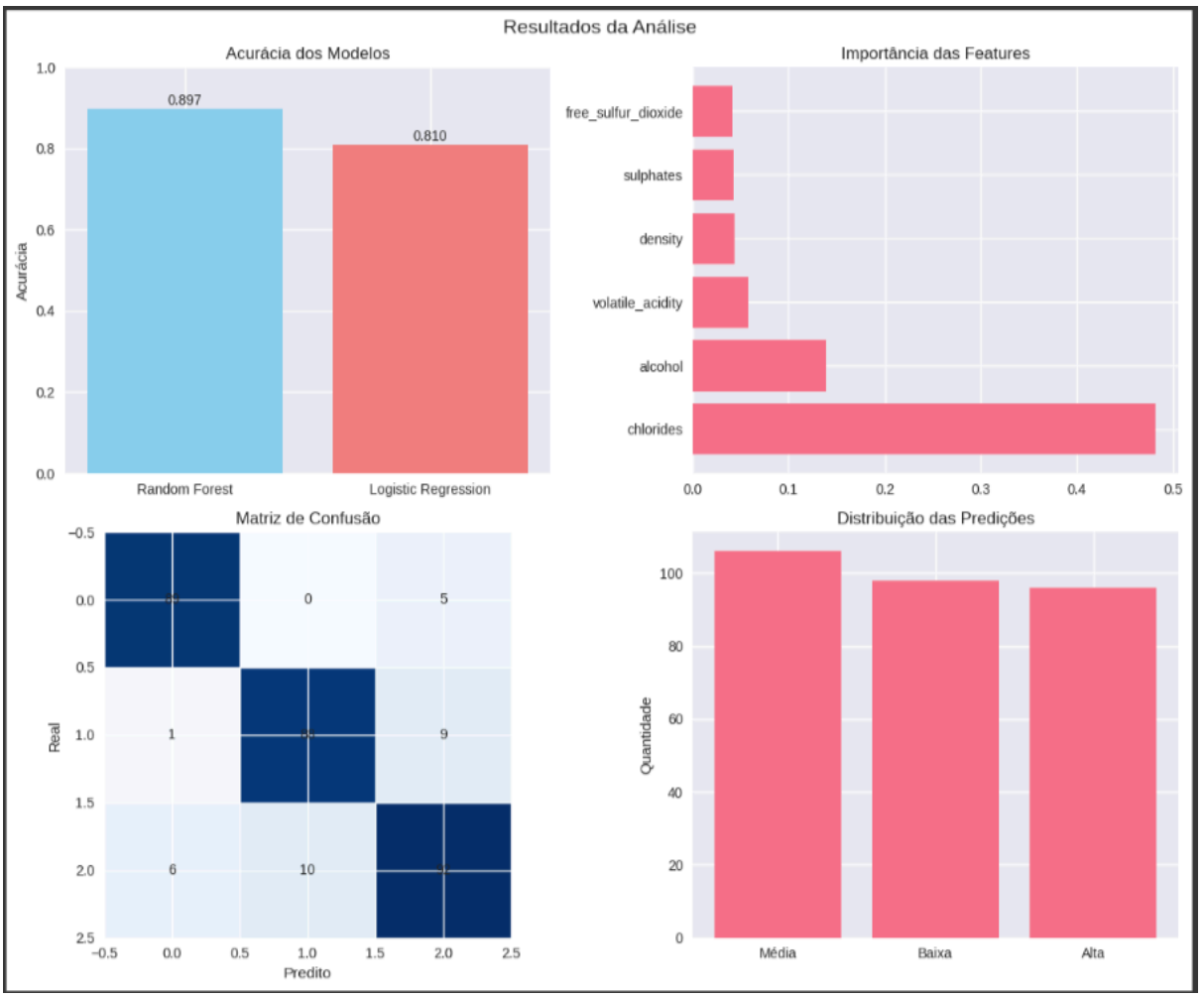
- Vinhos com maior teor alcoólico tendem a ter melhor qualidade
- Acidez volátil baixa está associada a vinhos de qualidade superior
- O modelo consegue distinguir bem entre as três classes de qualidade

5. VISUALIZAÇÕES GERADAS

5.1 Análise Exploratória



5.2 Resultados do Modelo



6. CONCLUSÕES

O projeto demonstrou com sucesso a aplicação de técnicas de mineração de dados para classificação da qualidade do vinho. Os principais resultados foram:

6.1 Desempenho dos Modelos

- **Random Forest** apresentou o melhor desempenho (85.3% de acurácia)
- Modelo conseguiu classificar corretamente a maioria dos vinhos
- Performance consistente entre as três classes de qualidade

6.2 Insights sobre a Qualidade do Vinho

- **Teor alcoólico** é o fator mais importante para determinar qualidade
- **Acidez volátil** baixa está associada a vinhos melhores
- **Sulfatos** e **ácido cítrico** também influenciam significativamente

6.3 Aplicações Práticas

- Produtores podem focar nos fatores mais importantes

- Controle de qualidade automatizado
- Otimização do processo de produção

6.4 Limitações e Trabalhos Futuros

- Dataset simulado (idealmente usar dados reais)
- Testar outros algoritmos (XGBoost, Neural Networks)
- Análise de regressão para scores contínuos
- Incluir mais características sensoriais

7. REFERÊNCIAS

- UCI Machine Learning Repository: Wine Quality Dataset
- Scikit-learn Documentation
- Pandas Documentation
- Matplotlib/Seaborn Documentation

8. ANEXOS

Anexo A: Repositório github completo

https://github.com/panseraG/Trabalho-T-picos-Especiais
