
**DEPARTAMENTO DE ENGENHARIA E CIÊNCIA DA COMPUTAÇÃO**
**CURSO DE CIÊNCIA DA COMPUTAÇÃO**
**URI ERECHIM**

**Felipe Meneguzzi**

# Data Mining

**Erechim, RS**
**2025**

## Classificação de Vinhos

Data Mining baseado em um dataset de qualidade de vinhos, código simples, mas eficiente, para elencar qualidade dos vinhos.

## Nome do Dataset e Origem

Para este trabalho de análise e mineração de dados, foi utilizado o Wine Quality Dataset, disponível no repositório oficial da UCI Machine Learning Repository.

**Nome:** Wine Quality
**Origem:** UCI Machine Learning Repository
**Link:** https://archive.ics.uci.edu/ml/datasets/Wine+Quality
**Total de registros:** 6497
**Total de atributos:** 11
**Variável-alvo:** quality (nota de 0 a 10 atribuída ao vinho com base em avaliações sensoriais)

## Descrição dos Resultados Obtidos

O objetivo da análise foi aplicar uma técnica de classificação supervisionada para prever se um vinho pode ser considerado "bom" ou "ruim" com base em suas propriedades físico-químicas. A classificação foi binarizada da seguinte forma:

- Bom: nota de qualidade maior ou igual a 7
- Ruim: nota inferior a 7

A técnica aplicada foi a Random Forest, um algoritmo de classificação baseado em múltiplas árvores de decisão. Essa escolha se deve à sua robustez, capacidade de lidar com conjuntos de dados desbalanceados e facilidade de interpretação da importância das variáveis.

Além disso, o conjunto de dados foi dividido em duas partes:

- 80% para treino
- 20% para teste

## Prints dos Gráficos e Relatórios Gerados

Após o treinamento do modelo, foram obtidas as seguintes métricas de avaliação:

## Acurácia do modelo

O modelo se mostrou bastante eficaz em classificar corretamente os vinhos com nota inferior a 7. Para vinhos com nota igual ou superior a 7, a performance foi razoável, refletindo o desbalanceamento natural do dataset

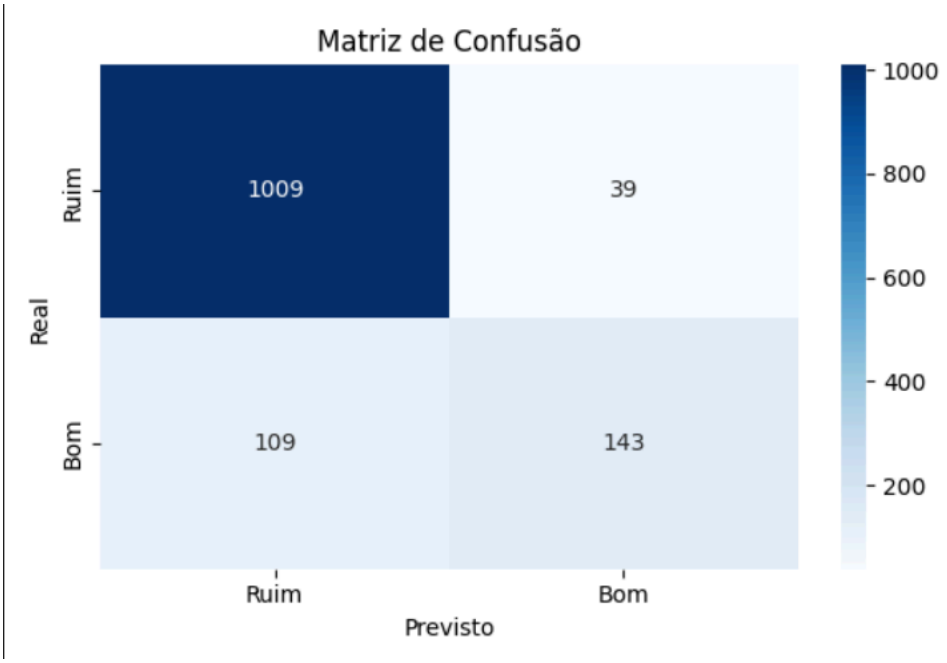
✅ Acurácia do modelo: 88.62%

📋 Relatório de Classificação:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Ruim (<7) | 0.90 | 0.96 | 0.93 | 1048 |
| Bom (>=7) | 0.79 | 0.57 | 0.66 | 252 |
| accuracy | | | 0.89 | 1300 |
| macro avg | 0.84 | 0.77 | 0.80 | 1300 |
| weighted avg | 0.88 | 0.89 | 0.88 | 1300 |

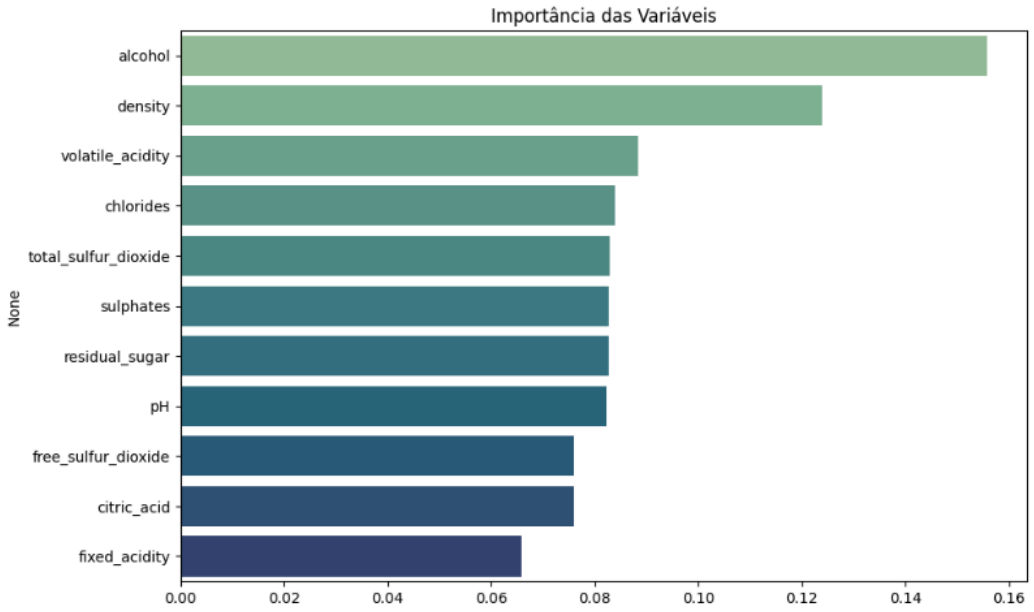
**Matriz de Confusão**

A matriz de confusão mostra a quantidade de acertos e erros do modelo em relação às duas classes previstas:



## Importância dos Atributos

O gráfico abaixo mostra as variáveis que mais contribuíram para a decisão do modelo. A variável alcohol foi a mais influente, seguida por density e volatile_acidity




## Conclusão

Através da aplicação da técnica de Random Forest, foi possível construir um modelo eficiente para prever a qualidade dos vinhos, com alta acurácia geral e boa capacidade de generalização.
A análise também permitiu identificar quais atributos físico-químicos são mais relevantes na classificação da qualidade do vinho. O projeto foi desenvolvido em Python, utilizando bibliotecas como pandas, scikit-learn, matplotlib e seaborn.

## Bibliografia

**Repositório GitHub**: https://github.com/Felipe-Meneguzz1/Data-Mining-in-Python
**wine_analysis.py**: Script com todo o código do projeto
**requirements.txt:** Dependências utilizadas
**matriz_confusao.png**: Gráfico gerado
**importancia_atributos.png:** Gráfico gerado