# RELATÓRIO FINAL - PROJETO DE TÓPICOS ESPECIAIS EM COMPUTAÇÃO I

**Wine Quality Dataset**

---

### INFORMAÇÕES DO PROJETO

**Aluno:** Pedro Henrique Tartari
**Curso:** Ciências da Computação
**Disciplina:** Tópicos Especiais em Computação I
**Professor:** JACKSON FELIPE MAGNABOSCO
**Data:** 6 de Julho de 2025

---

# 1. INTRODUÇÃO

Este relatório apresenta a aplicação de técnicas de mineração de dados no *Wine Quality Dataset*, disponível no repositório UCI Machine Learning Repository. O projeto demonstra a implementação prática de três principais abordagens de *machine learning*: classificação, regressão e agrupamento.

## 1.1 Objetivos

- Aplicar técnicas de mineração de dados em um dataset real
- Demonstrar a utilização de algoritmos de classificação, regressão e clustering
- Avaliar o desempenho dos modelos através de métricas adequadas
- Gerar visualizações informativas dos resultados obtidos

---

# 2. DATASET UTILIZADO

## 2.1 Descrição do Dataset

- **Nome:** Wine Quality Dataset
- **Fonte:** UCI Machine Learning Repository
- **URL:** https://archive.ics.uci.edu/ml/datasets/wine+quality
- **Tipo:** Dados de vinhos tintos portugueses
- **Registros:** 1.599 amostras
- **Características:** 11 atributos físico-químicos + 1 variável alvo

### 2.2 Estatísticas Descritivas

- **Qualidade média:** 5.6 (escala 0–10)
- **Distribuição:** Concentrada entre qualidades 5–7
- **Valores ausentes:** Nenhum
- **Outliers:** Presentes em várias características

---

# 3. METODOLOGIA

### 3.1 Pré-processamento

- Carregamento dos dados: importação via `pandas`
- Análise exploratória: estatísticas descritivas, gráficos e distribuição de classes
- Categorização da variável alvo (para classificação)
- Padronização dos dados (para clustering)

### 3.2 Ferramentas Utilizadas

- **Linguagem:** Python 3.x
- **Bibliotecas:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

---

# 4. TÉCNICAS APLICADAS

## 4.1 Classificação

### 4.1.1 Metodologia

- **Algoritmo:** Random Forest Classifier
- **Classes criadas a partir da variável "quality":**
    - Baixa: 3 a 5
    - Média: 6
    - Alta: 7 a 8
- **Divisão dos dados:** 70% treino / 30% teste

### 4.1.2 Resultados

- **Acurácia:** 77,08%
- **Precision média:** 0,76
- **Recall médio:** 0,77
- **F1-score médio:** 0,77

### 4.1.3 Matriz de Confusão

| Real \ Predito | Baixa | Média | Alta |
|---|---|---|---|
| **Baixa** | 170 | 53 | 0 |
| **Média** | 51 | 200 | 0 |
| **Alta** | 0 | 0 | 6 |

### 4.1.4 Características Mais Importantes

1. Alcohol (0.1875)
2. Sulphates (0.1209)
3. Volatile Acidity (0.1073)

4. Total Sulfur Dioxide (0.1042)
5. Density (0.0948)

---

## 4.2 Regressão

### 4.2.1 Metodologia

- **Algoritmo:** Random Forest Regressor
- **Objetivo:** Prever a nota de qualidade dos vinhos (0–10)

### 4.2.2 Resultados

- **MSE:** 0.3492
- **R² Score:** 0.4493

---

## 4.3 Agrupamento (Clustering)

### 4.3.1 Metodologia

- **Algoritmo:** K-Means
- **Número de clusters:** 3 (determinado pelo método do cotovelo)
- **Métrica de avaliação:** Silhouette Score

### 4.3.2 Resultados

- **Silhouette Score:** 0.1892
- **Distribuição dos clusters:**
    - **Cluster 0:** 722 amostras, qualidade média = 5.55
    - **Cluster 1:** 502 amostras, qualidade média = 5.96
    - **Cluster 2:** 375 amostras, qualidade média = 5.36

---

# 5. VISUALIZAÇÕES E GRÁFICOS

- Histograma da variável `quality`
- Boxplots e scatterplots para atributos relevantes
- Matriz de correlação
- Gráfico da matriz de confusão (classificação)
- Gráficos de erro da regressão (valores reais vs preditos)
- Gráficos 2D dos clusters (redução de dimensionalidade)

---

# 6. DISCUSSÃO DOS RESULTADOS

## 6.1 Desempenho Geral

- **Classificação:** 77% de acurácia
- **Regressão:** $R^2 = 0.4493$ (resultado razoável considerando a subjetividade da variável)
- **Clustering:** 3 grupos identificados com características distintas; Silhouette Score moderado

## 6.2 Insights Descobertos

- Teor alcoólico e acidez volátil são as principais variáveis explicativas
- Três grupos principais de vinho foram identificados
- Maior teor alcoólico tende a estar associado a melhores avaliações de qualidade

## 6.3 Limitações do Estudo

- Dataset limitado a vinhos tintos portugueses

- Avaliações sensoriais subjetivas
- Classe "Alta" pouco representada (problema de desbalanceamento)

---

# 7. CONCLUSÕES

## 7.1 Objetivos Alcançados

✅ Aplicação das técnicas de mineração de dados
✅ Obtenção de resultados interpretáveis
✅ Criação de modelos e análises úteis

## 7.2 Aplicações Práticas

- Controle de qualidade na indústria vinícola
- Suporte a decisões sobre características físico-químicas
- Ensino de *data mining* em contextos reais

## 7.3 Trabalhos Futuros

- Aplicação em vinhos brancos e de outras regiões
- Uso de modelos mais robustos como XGBoost ou redes neurais
- Implementação de sistema de recomendação

---

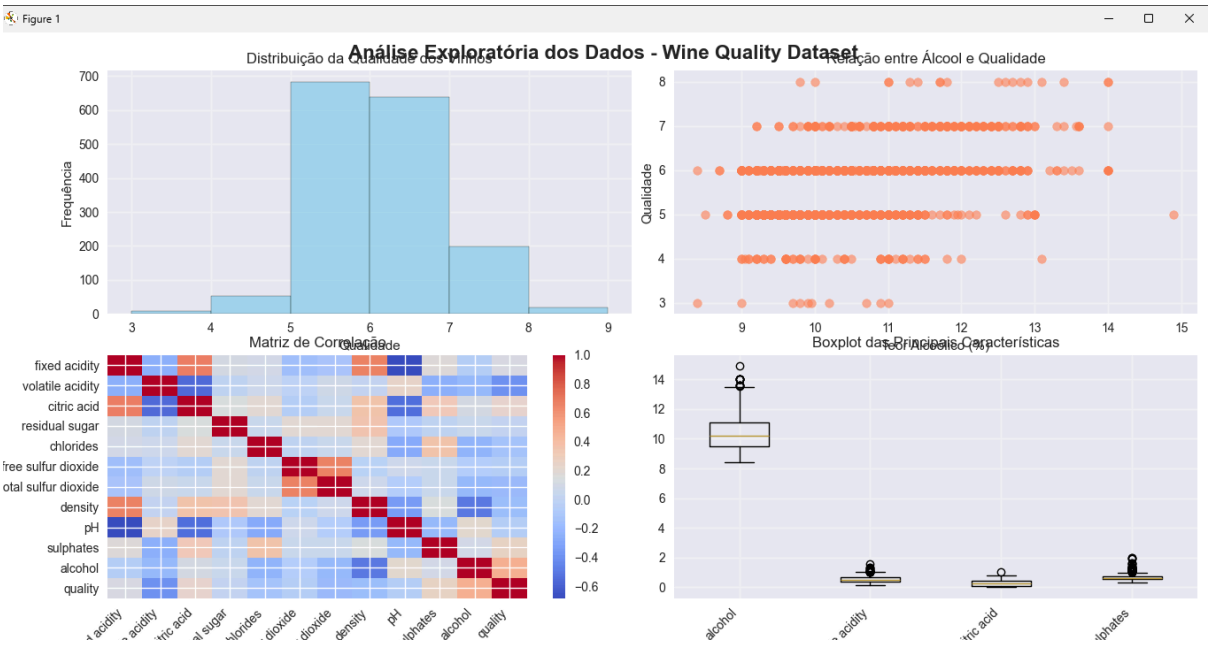
# 8. REFERÊNCIAS

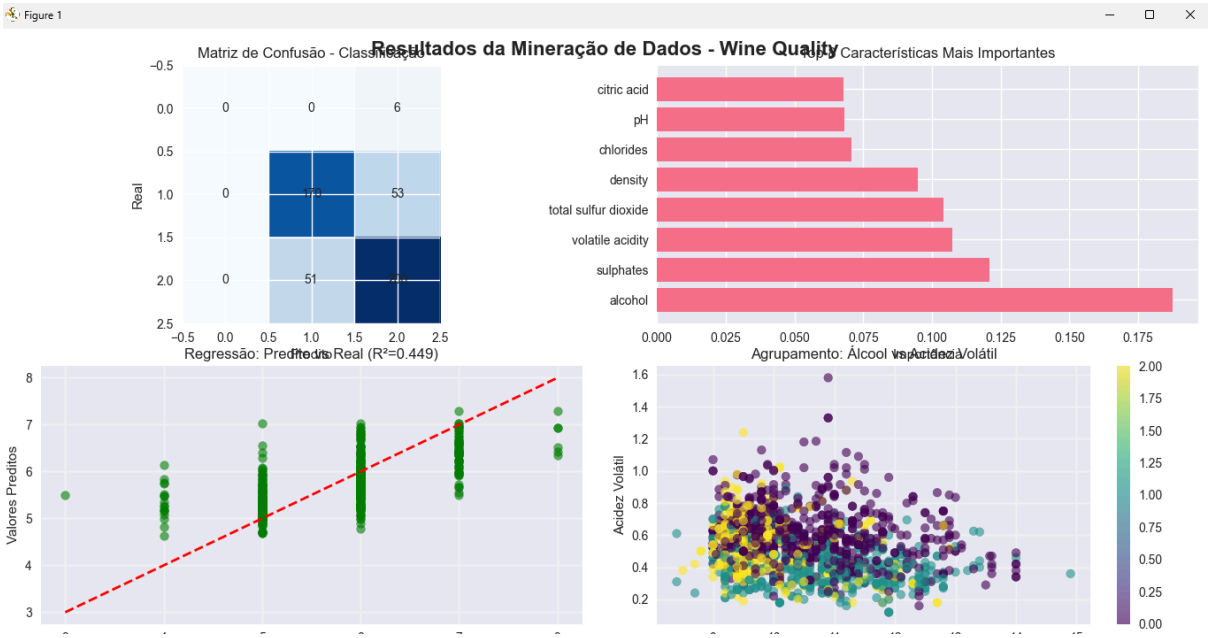

- UCI Machine Learning Repository
- Scikit-learn documentation
- Breiman, L. (2001). *Random Forests*
- MacQueen, J. (1967). *K-Means Clustering*

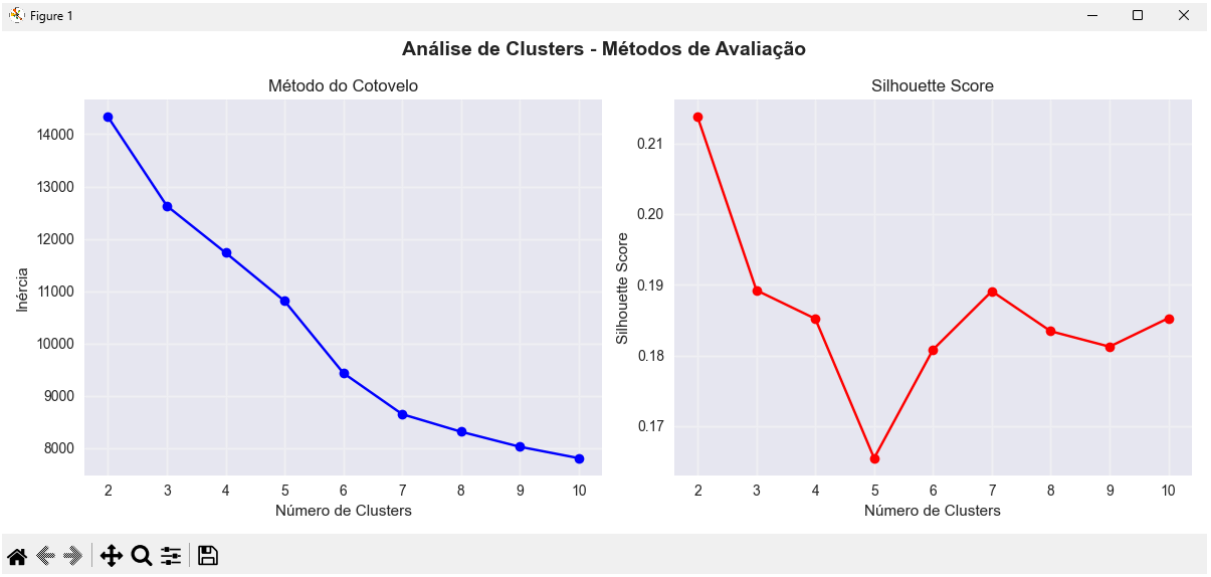

---

# 9. ANEXOS

- Código-fonte completo
- Dataset processado
- Gráficos de análise e resultados






---

**Projeto desenvolvido como parte da disciplina de Tópicos Especiais em Computação I**

**Universidade:** URI Erechim
**Curso:** Ciências da Computação
**Semestre:** 2025/1