RELATÓRIO FINAL - PROJETO DE Tópicos Especiais em Computação I

RELATÓRIO FINAL

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

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

RELATÓRIO FINAL - PROJETO DE Tópicos Especiais em Computação I

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 (classificação)
- Padronização para clustering

3.2 Ferramentas Utilizadas

- Python 3.x
- 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

	Predito		
Real	Baixa	Média	Alta
Baixa	170	53	0
Média	51	200	0
Alta	0	0	6

4.1.4 Características Mais Importantes

- Alcohol (0.1875)
- Sulphates (0.1209)
- Volatile Acidity (0.1073)
- Total Sulfur Dioxide (0.1042)
- 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

R^2 Score: 0.4493

4.3 AGRUPAMENTO (CLUSTERING)

4.3.1 Metodologia

Algoritmo: K-Means

Número de clusters: 3 (baseado no método do cotovelo)

Métrica de avaliação: Silhouette Score

RELATÓRIO FINAL - PROJETO DE Tópicos Especiais em Computação I

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 (real vs predito)
- Gráficos 2D dos clusters

6. DISCUSSÃO DOS RESULTADOS

6.1 Desempenho Geral

- Classificação: 77% de acurácia
- Regressão: Modelo com R^2 de 0.4493, com bom desempenho considerando a subjetividade da variável alvo
- Clustering: Três grupos 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 identificados
- Maior teor alcoólico tende a resultar em melhor qualidade

6.3 Limitações do Estudo

- Dataset limitado a vinhos tintos portugueses
- Avaliações sensoriais subjetivas
- Classe Alta pouco representada (desbalanceamento)

RELATÓRIO FINAL - PROJETO DE Tópicos Especiais em Computação I

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 Mineração de Dados

Universidade: URI Erechim

Curso: Ciências da Computação

Semestre: 2025/1