Trabalho de Tópicos Especiais I: Análise de Qualidade de Vinhos

Nome: Leonardo Pertile Follador - 066477 **Turma:** 2023

Introdução

Este relatório apresenta os resultados de um projeto prático de mineração de dados, desenvolvido para a disciplina de Mineração de Dados. O principal objetivo do trabalho é aplicar conceitos teóricos em um cenário real, utilizando o software WEKA para analisar o conjunto de dados "Wine Quality". Através da técnica de regressão com uma árvore de decisão, busca-se criar um modelo preditivo capaz de estimar o valor numérico da qualidade de um vinho com base em suas características físico-químicas.

1. Nome do Dataset e Origem

O conjunto de dados utilizado neste trabalho é o "Wine Quality", especificamente a variante de vinhos tintos. Ele foi obtido do UCI Machine Learning Repository, uma fonte confiável e amplamente utilizada para datasets de aprendizado de máquina.

O dataset consiste em 1.599 registros e 12 colunas, que representam 11 atributos físico-químicos (como acidez, açúcar, álcool) e 1 atributo alvo, que é a nota de qualidade do vinho, variando de 0 a 10.

2. Breve Descrição dos Resultados

A análise foi realizada no software WEKA, aplicando a técnica de mineração de dados de regressão. O objetivo era treinar um modelo capaz de prever o valor numérico da nota de qualidade (quality) de um vinho.

O algoritmo escolhido foi o REPTree, que, neste contexto, constrói uma árvore de regressão. O modelo foi avaliado usando a técnica de validação cruzada (Cross-validation) com 10 partições (folds).

Os principais resultados foram:

- **Coeficiente de Correlação (Correlation coefficient): 0.5665**. Este valor indica uma correlação positiva moderada entre os valores previstos pelo modelo e os valores reais. Quanto mais perto de 1, melhor a correlação.
- **Erro Médio Absoluto (Mean absolute error): 0.514**. Em média, a previsão do modelo errou em 0.514 pontos para cima ou para baixo, em uma escala de 0 a 10. Um erro menor é melhor.

A árvore de regressão gerada revelou que os atributos mais importantes para o modelo estimar a qualidade de um vinho foram, em ordem de importância:

1. **alcohol** (teor alcoólico)
2. **sulphates** (sulfatos)
3. **volatile acidity** (acidez volátil)
4. **total sulfur dioxide** (dióxido de enxofre total)

Essa estrutura visual da árvore permite entender como o modelo chega a uma previsão numérica, seguindo as regras em cada nó.

3. Prints dos Gráficos e Relatórios Gerados

20:17:21 - trees.REPTree

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     wine
Instances:    1599
Attributes:   12
              fixed acidity
              volatile acidity
              citric acid
              residual sugar
              chlorides
              free sulfur dioxide
              total sulfur dioxide
              density
              pH
              sulphates
              alcohol
              quality
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

alcohol < 11.45
|   sulphates < 0.58
|   |   volatile acidity < 0.34
|   |   |   alcohol < 10.2 : 5.25 (10/0.21) [10/0.37]
|   |   |   alcohol >= 10.2 : 6.17 (9/0.22) [3/0.67]
|   |   volatile acidity >= 0.34
|   |   |   volatile acidity < 0.79 : 5.23 (276/0.3) [148/0.34]
|   |   |   volatile acidity >= 0.79 : 4.79 (50/0.49) [20/0.86]
|   sulphates >= 0.58
|   |   volatile acidity < 0.41
|   |   |   total sulfur dioxide < 57.5
|   |   |   |   sulphates < 0.71
|   |   |   |   |   alcohol < 10.75 : 5.7 (30/0.37) [17/0.53]
|   |   |   |   |   alcohol >= 10.75
|   |   |   |   |   |   citric acid < 0.37 : 6.8 (4/0.5) [1/1]
|   |   |   |   |   |   citric acid >= 0.37
|   |   |   |   |   |   |   fixed acidity < 10.55
|   |   |   |   |   |   |   |   volatile acidity < 0.18 : 5.5 (2/0.25) [0/0]
|   |   |   |   |   |   |   |   volatile acidity >= 0.18 : 6.1 (8/0) [2/0.5]
|   |   |   |   |   |   |   fixed acidity >= 10.55
|   |   |   |   |   |   |   |   pH < 3.15 : 6.14 (5/0) [2/0.5]
|   |   |   |   |   |   |   |   pH >= 3.15 : 7.67 (2/0.25) [1/0.25]
|   |   |   |   sulphates >= 0.71
|   |   |   |   |   alcohol < 9.75
|   |   |   |   |   |   fixed acidity < 14.65
|   |   |   |   |   |   |   citric acid < 0.56 : 5.89 (11/0) [7/0.57]
|   |   |   |   |   |   |   citric acid >= 0.56 : 5.2 (4/0.19) [1/0.06]
|   |   |   |   |   |   fixed acidity >= 14.65 : 7 (2/0) [0/0]
|   |   |   |   |   alcohol >= 9.75
|   |   |   |   |   |   volatile acidity < 0.29 : 6.39 (21/0.2) [10/0.34]
|   |   |   |   |   |   volatile acidity >= 0.29
|   |   |   |   |   |   |   fixed acidity < 7.4 : 7.2 (6/0.22) [4/0.11]
|   |   |   |   |   |   |   fixed acidity >= 7.4
```

20:17:21 - trees.REPTree

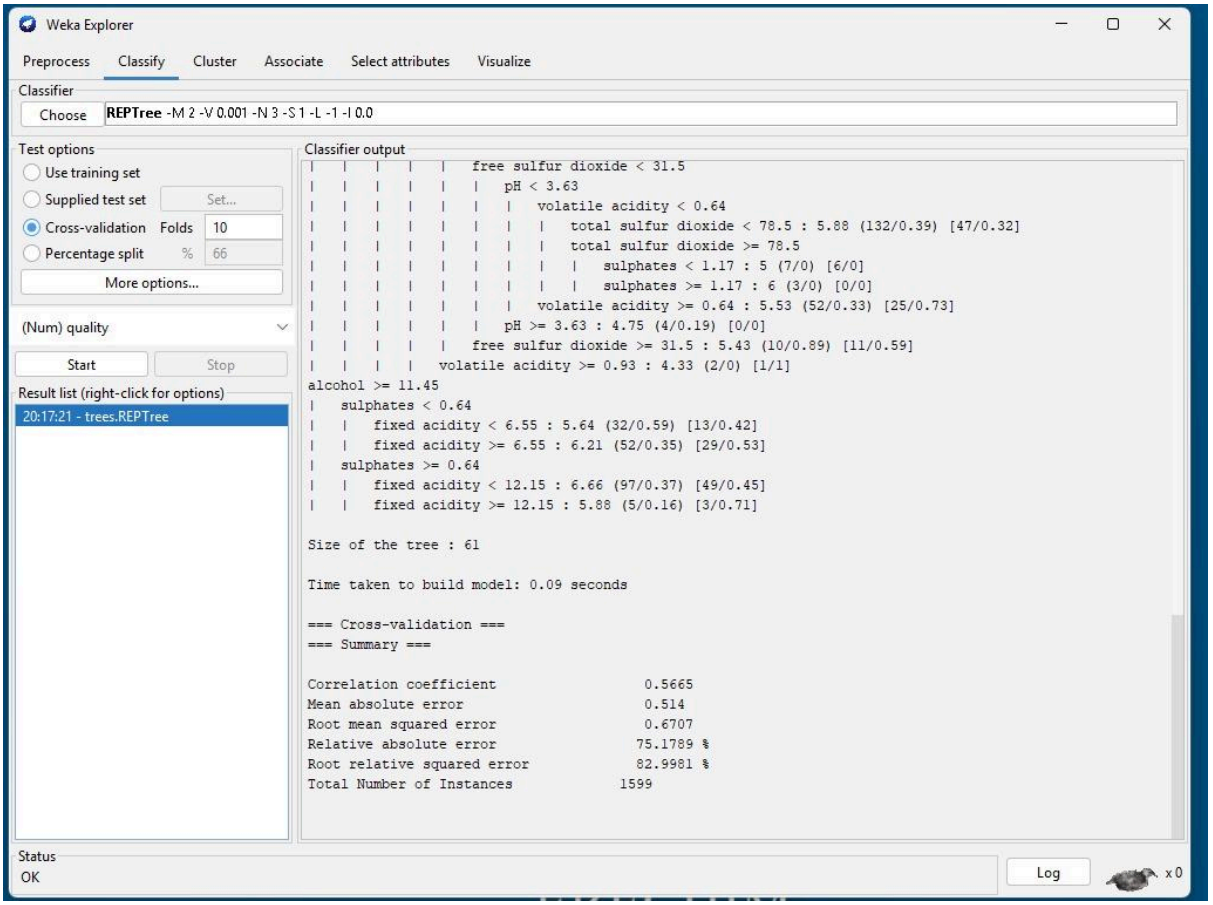
```
|   |   |   |   |   |   |   |   volatile acidity < 0.18 : 5.5 (2/0.25) [0/0]
|   |   |   |   |   |   |   |   volatile acidity >= 0.18 : 6.1 (8/0) [2/0.5]
|   |   |   |   |   |   |   fixed acidity >= 10.55
|   |   |   |   |   |   |   |   pH < 3.15 : 6.14 (5/0) [2/0.5]
|   |   |   |   |   |   |   |   pH >= 3.15 : 7.67 (2/0.25) [1/0.25]
|   |   |   |   sulphates >= 0.71
|   |   |   |   |   alcohol < 9.75
|   |   |   |   |   |   fixed acidity < 14.65
|   |   |   |   |   |   |   citric acid < 0.56 : 5.89 (11/0) [7/0.57]
|   |   |   |   |   |   |   citric acid >= 0.56 : 5.2 (4/0.19) [1/0.06]
|   |   |   |   |   |   fixed acidity >= 14.65 : 7 (2/0) [0/0]
|   |   |   |   |   alcohol >= 9.75
|   |   |   |   |   |   volatile acidity < 0.29 : 6.39 (21/0.2) [10/0.34]
|   |   |   |   |   |   volatile acidity >= 0.29
|   |   |   |   |   |   |   fixed acidity < 7.4 : 7.2 (6/0.22) [4/0.11]
|   |   |   |   |   |   |   fixed acidity >= 7.4
|   |   |   |   |   |   |   |   density < 1 : 6.29 (16/0.23) [19/0.4]
|   |   |   |   |   |   |   |   density >= 1 : 6.79 (12/0.24) [2/0.84]
|   |   |   total sulfur dioxide >= 57.5
|   |   |   |   total sulfur dioxide < 82.5 : 5.76 (24/0.39) [17/0.39]
|   |   |   |   total sulfur dioxide >= 82.5 : 5.15 (11/0.18) [2/2]
|   |   volatile acidity >= 0.41
|   |   |   alcohol < 9.85 : 5.28 (167/0.29) [83/0.3]
|   |   |   alcohol >= 9.85
|   |   |   |   volatile acidity < 0.93
|   |   |   |   |   free sulfur dioxide < 31.5
|   |   |   |   |   |   pH < 3.63
|   |   |   |   |   |   |   volatile acidity < 0.64
|   |   |   |   |   |   |   |   total sulfur dioxide < 78.5 : 5.88 (132/0.39) [47/0.32]
|   |   |   |   |   |   |   |   total sulfur dioxide >= 78.5
|   |   |   |   |   |   |   |   |   sulphates < 1.17 : 5 (7/0) [6/0]
|   |   |   |   |   |   |   |   |   sulphates >= 1.17 : 6 (3/0) [0/0]
|   |   |   |   |   |   |   volatile acidity >= 0.64 : 5.53 (52/0.33) [25/0.73]
|   |   |   |   |   |   pH >= 3.63 : 4.75 (4/0.19) [0/0]
|   |   |   |   |   free sulfur dioxide >= 31.5 : 5.43 (10/0.89) [11/0.59]
|   |   |   |   volatile acidity >= 0.93 : 4.33 (2/0) [1/1]
alcohol >= 11.45
|   sulphates < 0.64
|   |   fixed acidity < 6.55 : 5.64 (32/0.59) [13/0.42]
|   |   fixed acidity >= 6.55 : 6.21 (52/0.35) [29/0.53]
|   sulphates >= 0.64
|   |   fixed acidity < 12.15 : 6.66 (97/0.37) [49/0.45]
|   |   fixed acidity >= 12.15 : 5.88 (5/0.16) [3/0.71]

Size of the tree : 61

Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5665
Mean absolute error                      0.514
Root mean squared error                  0.6707
Relative absolute error                 75.1789 %
Root relative squared error             82.9981 %
Total Number of Instances             1599
```



Conclusão

A análise realizada demonstrou com sucesso a aplicação da técnica de regressão para um problema do mundo real. O modelo REPTree conseguiu estabelecer uma correlação moderada com os dados reais, e o erro médio absoluto de 0.514 é um resultado razoável, considerando a complexidade e subjetividade da avaliação de vinhos. A análise da árvore permitiu identificar que o teor alcoólico e os sulfatos são os fatores mais influentes para estimar a nota de um vinho. O trabalho cumpre seu objetivo ao executar o ciclo completo de um projeto de mineração de dados, destacando o potencial do WEKA para criar e avaliar modelos preditivos numéricos.