Trabalho de Tópicos Especiais I: Análise de Doenças Cardíacas

Nome: Gustavo Pertile Follador - 094713

Introdução

Este relatório apresenta os resultados de um projeto prático de mineração de dados, desenvolvido para a disciplina de Tópicos Especiais em Computação. O principal objetivo do trabalho é aplicar conceitos teóricos em um cenário real, utilizando o software WEKA para analisar o conjunto de dados Heart Disease UCI. Por meio da técnica de regressão utilizando uma árvore de decisão, busca-se criar um modelo preditivo capaz de estimar a gravidade de uma doença cardíaca com base em atributos clínicos e físicos dos pacientes.

Nome do Dataset e Origem

O conjunto de dados utilizado neste trabalho é o Heart Disease Dataset, obtido da plataforma Kaggle (https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data) – uma das principais fontes de dados para projetos de aprendizado de máquina.

Este dataset contém 920 instâncias e 16 atributos, incluindo variáveis como idade, sexo, pressão sanguínea, colesterol, frequência cardíaca máxima, entre outras. O atributo alvo é num, que representa a gravidade da doença cardíaca. Os dados são provenientes de quatro bancos distintos: cleveland, hungary, switzerland e va_long_beach.

Breve Descrição dos Resultados

A análise foi realizada no software WEKA, utilizando o algoritmo de regressão REPTree (Reduced Error Pruning Tree). O objetivo foi treinar um modelo capaz de prever valores numéricos relacionados ao grau da doença cardíaca.

O modelo foi avaliado usando a técnica de validação cruzada (Cross-Validation) com 10 partições (folds), resultando numa árvore de 96 nós.

Principais métricas obtidas:
- **Correlação (Correlation Coefficient):** 0.7164
- **Erro Médio Absoluto (Mean Absolute Error):** 4.55
- **Erro Quadrático Médio (Root Mean Squared Error):** 6.5776
- **Erro Absoluto Relativo (Relative Absolute Error):** 59.71%
- **Erro Quadrático Relativo (Root Relative Squared Error):** 69.68%

Esses resultados indicam que o modelo tem um desempenho razoável, com uma boa correlação, mas ainda com margem de erro considerável, o que é esperado em um problema com alta variabilidade clínica como este.

A árvore de decisão construída se baseia fortemente nos atributos:
- **thalach** (frequência cardíaca máxima)
- **ca** (número de vasos maiores coloridos por fluoroscopia)
- **trestbps** (pressão arterial em repouso)
- **thal** (resultado do exame de thalassemia)
- **dataset** (origem do dado: cleveland, hungary, etc.)

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

Prints dos Gráficos e Relatórios Gerados

21:19:23 - trees.REPTree

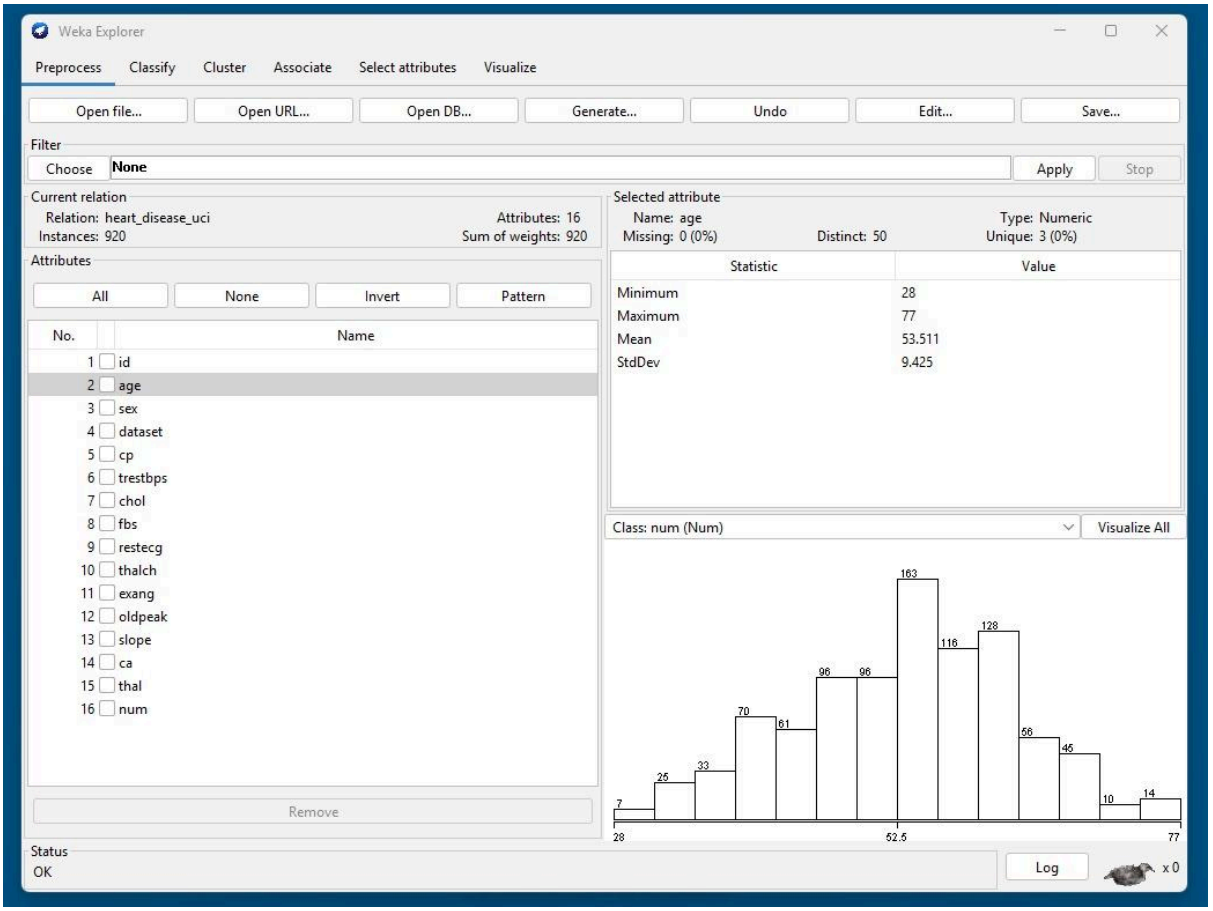
```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     heart_disease_uci
Instances:    920
Attributes:   16
              id
              age
              sex
              dataset
              cp
              trestbps
              chol
              fbs
              restecg
              thalch
              exang
              oldpeak
              slope
              ca
              thal
              num
Test mode:    10-fold cross-validation

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


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

dataset = cleveland
|   thalch < 172.5
|   |   ca < 0.5
|   |   |   trestbps < 139 : 50.93 (59.55/74.38) [37.09/51.08]
|   |   |   trestbps >= 139 : 59.26 (30/49.58) [16/30.67]
|   |   ca >= 0.5
|   |   |   ca < 1.5 : 57.58 (35.21/47.98) [25.43/44.7]
|   |   |   ca >= 1.5
|   |   |   |   id < 191
|   |   |   |   |   thal = fixed_defect : 65 (2/0.25) [1/2.25]
|   |   |   |   |   thal = normal : 68.67 (5/23.6) [1.24/59.02]
|   |   |   |   |   thal = reversable_defect : 59.73 (16/15.25) [6/73.42]
|   |   |   |   id >= 191
|   |   |   |   |   trestbps < 111 : 65.67 (2/4) [1/4]
|   |   |   |   |   trestbps >= 111 : 56.89 (14.24/27.98) [5.24/78.44]
|   thalch >= 172.5 : 45.55 (31/85.88) [16/44.53]
dataset = hungary
|   id < 381.5
|   |   id < 343
|   |   |   id < 315 : 31.6 (3/4.22) [7/2.54]
|   |   |   id >= 315
|   |   |   |   id < 330.5
|   |   |   |   |   id < 323.5 : 35.33 (5/0.16) [4/1.29]
|   |   |   |   |   id >= 323.5 : 37 (4/0) [3/0]
|   |   |   |   id >= 330.5
|   |   |   |   |   id < 334 : 38 (2/0) [1/0]
|   |   |   |   |   id >= 334 : 39 (5/0) [4/0]
|   |   id >= 343
|   |   |   id < 362.5
```

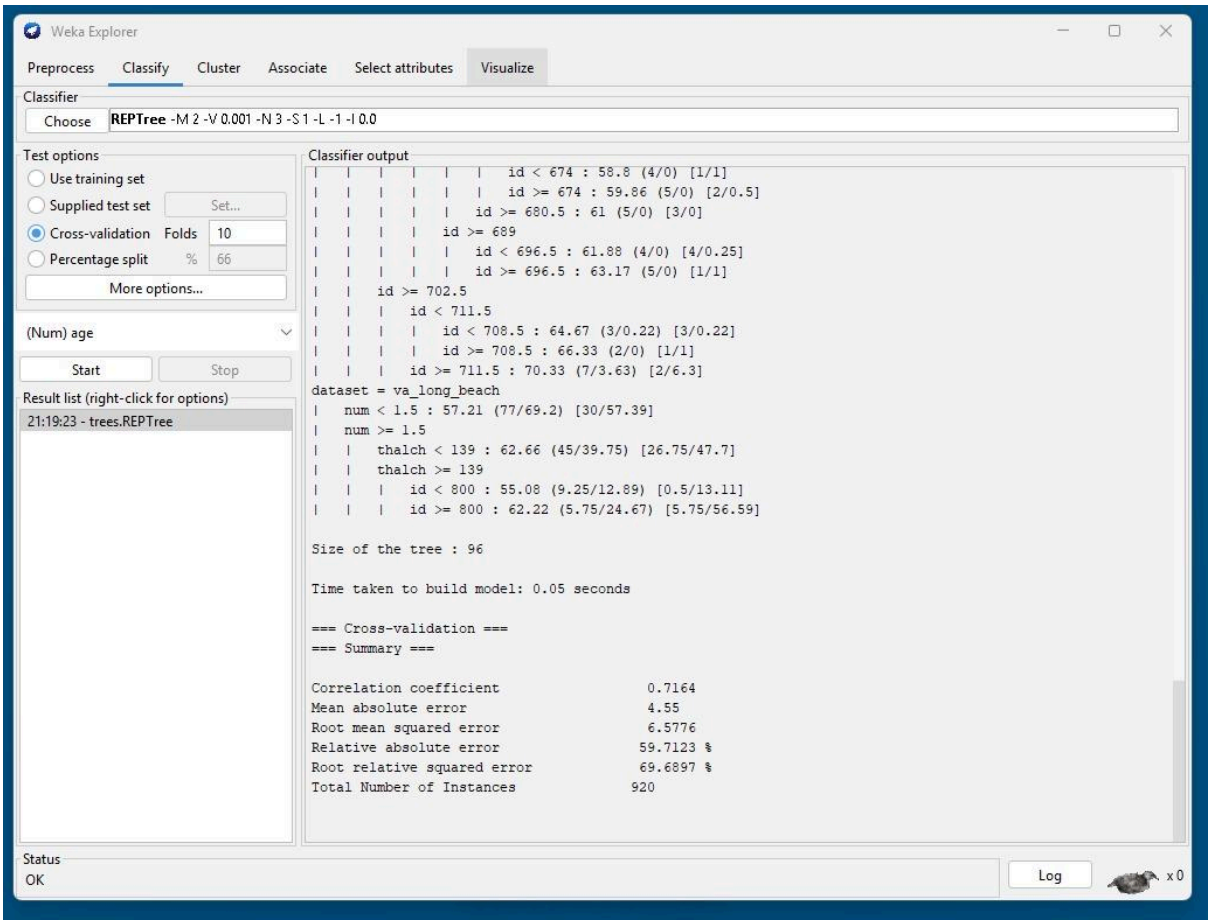
21:19:23 - trees.REPTree

```
|   |   |   |   |   id < 323.5 : 35.33 (5/0.16) [4/1.29]
|   |   |   |   |   id >= 323.5 : 37 (4/0) [3/0]
|   |   |   |   id >= 330.5
|   |   |   |   |   id < 334 : 38 (2/0) [1/0]
|   |   |   |   |   id >= 334 : 39 (5/0) [4/0]
|   |   id >= 343
|   |   |   id < 362.5
|   |   |   |   id < 355.5
|   |   |   |   |   id < 348.5 : 39.83 (3/0) [3/0.33]
|   |   |   |   |   id >= 348.5 : 41 (6/0) [1/0]
|   |   |   |   id >= 355.5 : 42 (7/0) [0/0]
|   |   |   id >= 362.5
|   |   |   |   id < 370 : 43 (6/0) [1/0]
|   |   |   |   id >= 370
|   |   |   |   |   id < 374 : 43.75 (3/0) [1/1]
|   |   |   |   |   id >= 374 : 44.88 (5/0) [3/0.33]
|   id >= 381.5
|   |   trestbps < 122
|   |   |   cp = asymptomatic : 44.84 (23/51.71) [9/36.27]
|   |   |   cp = atypical_angina : 50.49 (15.31/7.9) [8/71.9
|   |   |   cp = nonanginal
|   |   |   |   id < 412.5 : 47.33 (3/0.89) [0/0]
|   |   |   |   id >= 412.5
|   |   |   |   |   id < 451.5 : 52.33 (2/1) [1/1]
|   |   |   |   |   id >= 451.5 : 54.5 (3/0.22) [3/20.33]
|   |   |   cp = typical_angina : 48 (0/0) [3/20.91]
|   |   trestbps >= 122 : 52.43 (102.69/28.05) [43/30.94]
dataset = switzerland
|   id < 648
|   |   id < 615
|   |   |   id < 607.5 : 36.5 (6/1.25) [4/11.25]
|   |   |   id >= 607.5
|   |   |   |   id < 610 : 40.5 (2/0.25) [0/0]
|   |   |   |   id >= 610 : 42.6 (2/0.25) [3/0.25]
|   |   id >= 615
|   |   |   id < 624
|   |   |   |   id < 617.5 : 45.67 (2/0) [1/1]
|   |   |   |   id >= 617.5 : 47.67 (4/0.19) [2/3.81]
|   |   |   id >= 624
|   |   |   |   id < 638.5
|   |   |   |   |   id < 633.5 : 50.7 (8/0.11) [2/0.77]
|   |   |   |   |   id >= 633.5 : 52.2 (3/0) [2/0.5]
|   |   |   |   id >= 638.5
|   |   |   |   |   id < 645.5 : 53 (6/0) [1/0]
|   |   |   |   |   id >= 645.5 : 54 (2/0) [0/0]
|   id >= 648
|   |   id < 702.5
|   |   |   id < 669
|   |   |   |   id < 660.5
|   |   |   |   |   id < 653 : 54.8 (3/0) [2/0.5]
|   |   |   |   |   id >= 653 : 56 (5/0) [3/0]
|   |   |   |   id >= 660.5 : 57.25 (5/0.16) [3/0.24]
|   |   |   id >= 669
|   |   |   |   id < 689
|   |   |   |   |   id < 680.5
|   |   |   |   |   |   id < 674 : 58.8 (4/0) [1/1]
|   |   |   |   |   |   id >= 674 : 59.86 (5/0) [2/0.5]
|   |   |   |   |   id >= 680.5 : 61 (5/0) [3/0]
|   |   |   |   id >= 689
```

21:19:23 - trees.REPTree

```
|   |   |   id < 607.5 : 36.5 (6/1.25) [4/11.25]
|   |   |   id >= 607.5
|   |   |   |   id < 610 : 40.5 (2/0.25) [0/0]
|   |   |   |   id >= 610 : 42.6 (2/0.25) [3/0.25]
|   |   id >= 615
|   |   |   id < 624
|   |   |   |   id < 617.5 : 45.67 (2/0) [1/1]
|   |   |   |   id >= 617.5 : 47.67 (4/0.19) [2/3.81]
|   |   |   id >= 624
|   |   |   |   id < 638.5
|   |   |   |   |   id < 633.5 : 50.7 (8/0.11) [2/0.77]
|   |   |   |   |   id >= 633.5 : 52.2 (3/0) [2/0.5]
|   |   |   |   id >= 638.5
|   |   |   |   |   id < 645.5 : 53 (6/0) [1/0]
|   |   |   |   |   id >= 645.5 : 54 (2/0) [0/0]
|   id >= 648
|   |   id < 702.5
|   |   |   id < 669
|   |   |   |   id < 660.5
|   |   |   |   |   id < 653 : 54.8 (3/0) [2/0.5]
|   |   |   |   |   id >= 653 : 56 (5/0) [3/0]
|   |   |   |   id >= 660.5 : 57.25 (5/0.16) [3/0.24]
|   |   |   id >= 669
|   |   |   |   id < 689
|   |   |   |   |   id < 680.5
|   |   |   |   |   |   id < 674 : 58.8 (4/0) [1/1]
|   |   |   |   |   |   id >= 674 : 59.86 (5/0) [2/0.5]
|   |   |   |   |   id >= 680.5 : 61 (5/0) [3/0]
|   |   |   |   id >= 689
|   |   |   |   |   id < 696.5 : 61.88 (4/0) [4/0.25]
|   |   |   |   |   id >= 696.5 : 63.17 (5/0) [1/1]
|   |   id >= 702.5
|   |   |   id < 711.5
|   |   |   |   id < 708.5 : 64.67 (3/0.22) [3/0.22]
|   |   |   |   id >= 708.5 : 66.33 (2/0) [1/1]
|   |   |   id >= 711.5 : 70.33 (7/3.63) [2/6.3]
dataset = va_long_beach
|   num < 1.5 : 57.21 (77/69.2) [30/57.39]
|   num >= 1.5
|   |   thalch < 139 : 62.66 (45/39.75) [26.75/47.7]
|   |   thalch >= 139
|   |   |   id < 800 : 55.08 (9.25/12.89) [0.5/13.11]
|   |   |   id >= 800 : 62.22 (5.75/24.67) [5.75/56.59]

Size of the tree : 96

Time taken to build model: 0.05 seconds

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

Correlation coefficient                  0.7164
Mean absolute error                      4.55
Root mean squared error                  6.5776
Relative absolute error                 59.7123 %
Root relative squared error             69.6897 %
Total Number of Instances              920
```


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: heart_disease_uci
Instances: 920
Attributes: 16
Sum of weights: 920
Attributes
All
None
Invert
Pattern
No.
Name
1 id
2 age
3 sex
4 dataset
5 cp
6 trestbps
7 chol
8 fbs
9 restecg
10 thalch
11 exang
12 oldpeak
13 slope
14 ca
15 thal
16 num
Remove
Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 50
Type: Numeric
Unique: 3 (0%)
Statistic
Value
Minimum 28
Maximum 77
Mean 53.511
StdDev 9.425
Class: num (Num)
Visualize All
7
25
33
70
61
96
96
163
116
128
66
45
10
14
28
52.5
77
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) age
Start
Stop
Result list (right-click for options)
21:19:23 - trees.REPTree
Classifier output
| | | | | | id < 674 : 58.8 (4/0) [1/1]
| | | | | | id >= 674 : 59.86 (5/0) [2/0.5]
| | | | | id >= 680.5 : 61 (5/0) [3/0]
| | | | id >= 689
| | | | | id < 696.5 : 61.88 (4/0) [4/0.25]
| | | | | id >= 696.5 : 63.17 (5/0) [1/1]
| | id >= 702.5
| | | id < 711.5
| | | | id < 708.5 : 64.67 (3/0.22) [3/0.22]
| | | | id >= 708.5 : 66.33 (2/0) [1/1]
| | | id >= 711.5 : 70.33 (7/3.63) [2/6.3]
dataset = va_long_beach
| num < 1.5 : 57.21 (77/69.2) [30/57.39]
| num >= 1.5
| | thalch < 139 : 62.66 (45/39.75) [26.75/47.7]
| | thalch >= 139
| | | id < 800 : 55.08 (9.25/12.89) [0.5/13.11]
| | | id >= 800 : 62.22 (5.75/24.67) [5.75/56.59]
Size of the tree : 96
Time taken to build model: 0.05 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.7164
Mean absolute error 4.55
Root mean squared error 6.5776
Relative absolute error 59.7123 %
Root relative squared error 69.6897 %
Total Number of Instances 920
Status
OK
Log
x 0

Conclusão

A análise realizada demonstrou com sucesso a aplicação da técnica de regressão via árvore de decisão (REPTree) para um problema real da área da saúde. O modelo construído apresentou um bom nível de correlação com os dados reais e conseguiu capturar padrões importantes nos diferentes atributos clínicos dos pacientes.

Apesar da variabilidade dos dados entre os diferentes subdatasets (cleveland, hungary, etc.), o modelo foi capaz de generalizar razoavelmente bem, o que é refletido no coeficiente de correlação de 0.7164. Esse valor mostra que o modelo tem capacidade preditiva útil, ainda que com limitações.

O trabalho evidencia o potencial do WEKA e da mineração de dados para análise preditiva em contextos médicos, oferecendo suporte à tomada de decisão clínica e à análise de risco em doenças cardíacas.