Descoberta de Conhecimento - FE04

Vasco Ramos, PG42852 ; Carolina Marques PG42818

```
@attribute 'age' real
@attribute 'sex' { female, male}
@attribute 'cp' { typ_angina, asympt, non_anginal, atyp_angina}
@attribute 'trestbps' real
@attribute 'chol' real
@attribute 'fbs' { t, f}
@attribute 'restecg' { left_vent_hyper, normal, st_t_wave_abnormality}
@attribute 'thalach' real
@attribute 'exang' { no, yes}
@attribute 'oldpeak' real
@attribute 'slope' { up, flat, down}
@attribute 'ca' real
@attribute 'thal' { fixed_defect, normal, reversable_defect}
@attribute 'num' { '<50', '>50_1'}
```

1.1

O que significam as mensagens pop-up que aparecem ao arrastar o rato sobre o gráfico?

- As mensagens pop-up indicam qual o valor do atributo e dentro do tuplo aparecem os valores onde a classe se encontra compreendida. (Ex: em age 63 (57.8,62.6), os valores encontram-se dentro do range de um aumento de doença cardíaca)
- Quando os atributos são nominais o pop-up diz qual o valor (número total) de instâncias que se são desta classe (Ex: *left_vent_hyper[147]*, temos que 147 instâncias pertencem a esta classe)

age

a) Numérico
b) 0%
c) Min: 29 | Max: 77 | Média: 54.366 | Desvio padrão: 9.082
d) Sim, existem 4
e) Com o aumento da idade (até um certo limite ~60 anos), existe uma probabilidade maior de aumento de doença cardíaca

sex

a) Nominal
b) 0%
c) Não aplicável
d) Não

e) Dado que existem mais Homens com doença cardíaca, estes têm uma maior probabilidade de aumento de doença cardíaca (no entanto, ressalvar que existem mais instâncias de homens do que mulheres no dataset)

cp

a) Nominal
b) 0%
c) Não aplicável
d) Não
e) Quem não tem sintomas têm uma maior probabilidade de aumento de doença cardíaca

trestbps

a) Numérico
b) 0%
c) Min: 94 | Max: 200 | Média: 131.624 | Desvio padrão: 17.538
d) Sim, existem 16
e) Quando trestbps é 71 existe uma maior probabilidade de aumento de doença cardíaca

chol

a) Numérico
b) 0%
c) Min: 126 | Max: 564 | Média: 246.264 | Desvio padrão: 51.831
d) Sim, existem 62
e) O aumento da doença cardíaca acontece quando o colesterol sérico em mg/dl se encontra compreendido entre 50 e 69.

fbs

a) Nominal
b) 0%
c) Não aplicável
d) Não
e) Quando açúcar no sangue em jejum < 120 mg/dl existe um aumento da doença cardíaca

restecg

a) Nominal
b) 0%
c) Não aplicável
d) Não
e) Os resultados eletrocardiograma de repouso mostram que quando apresenta hipertrofia ventricular esquerda provável ou definitiva apresenta um aumento da doença cardíaca

thalach

a) Numérico
b) 0%
c) Min: 71 | Max: 202 | Média: 149.647 | Desvio padrão: 22.905

d) Sim, existem 28
e) Em valores entre 50 a 69 existe um aumento do nível da doença cardíaca

exang

a) Nominal
b) 0%
c) Não aplicável
d) Não
e) Existe um aumento do nível da doença cardíaca quando apresentam angina induzida pelo exercício

oldpeak

a) Numérico
b) 0%
c) Min: 0 | Max: 6.2 | Média: 1.04 | Desvio padrão: 1.161
d) Sim, existem 10
e) Em valores entre 0 e 0.62 existe um aumento do nível da doença cardíaca

slope

a) Nominal
b) 0%
c) Não aplicável
d) Não
e) Existe um aumento do nível da doença cardíaca quando o declive do segmento ST do exercício de pico é flat

ca

a) Numérico
b) 2%
c) Min: 0 | Max: 3 | Média: 0.674 | Desvio padrão: 0.938
d) Não
e) Número de veias principais se encontra entre 0 e 0.5 existe um aumento do nivel da doença cardíaca

thal

a) Nominal
b) 1%
c) Não aplicável
d) Não
e) Quando Thalassemia é reversable_defect existe um maior aumento do nível da doença cardíaca

num

a) Nominal
b) 0%

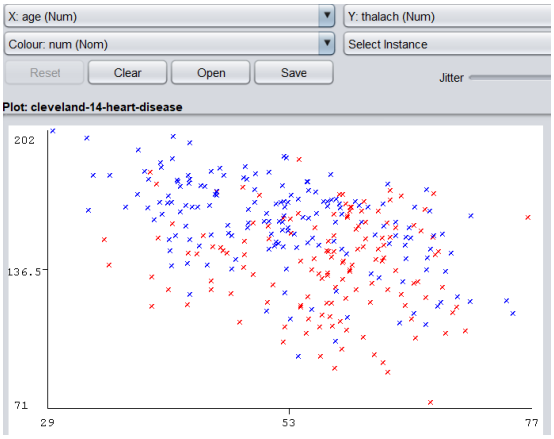
c) Não aplicável
d) Não
e) --

1.2.

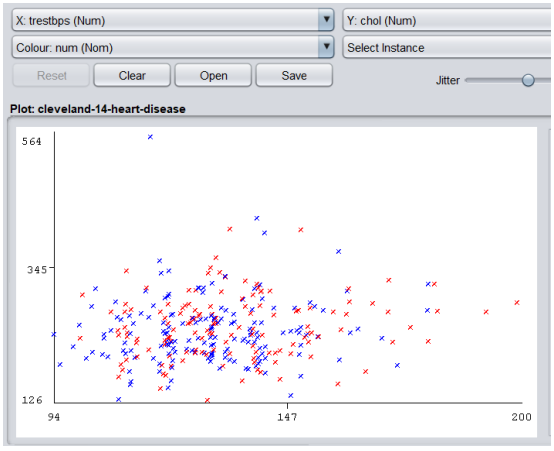
a) O thalac apesar de não aparentar grande correlação com a presença de doenças cardíacas, aparenta ter um maior número de casos não associados a doença cardíaca para thalac > 136.5, pelo que se pode dizer que, em certa medida, apresenta alguma relação. O ca, por outro lado, parece estar mais associado a doenças cardíacas.

b) Os pares de atributos que aparentam estar mais relacionados são:

- age-thalach
- thresbps-chol

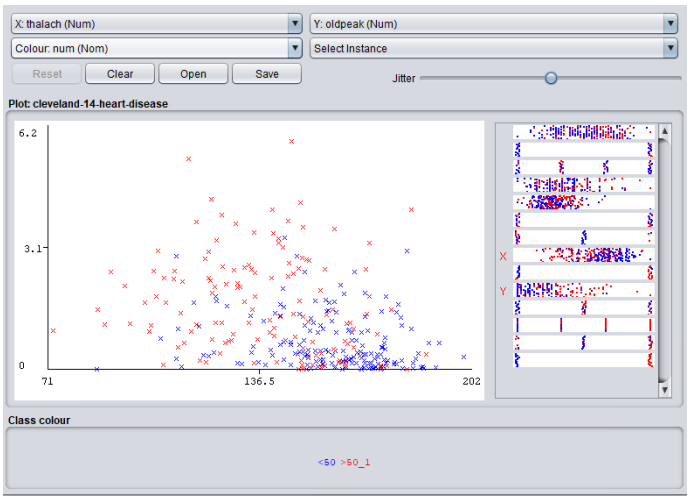


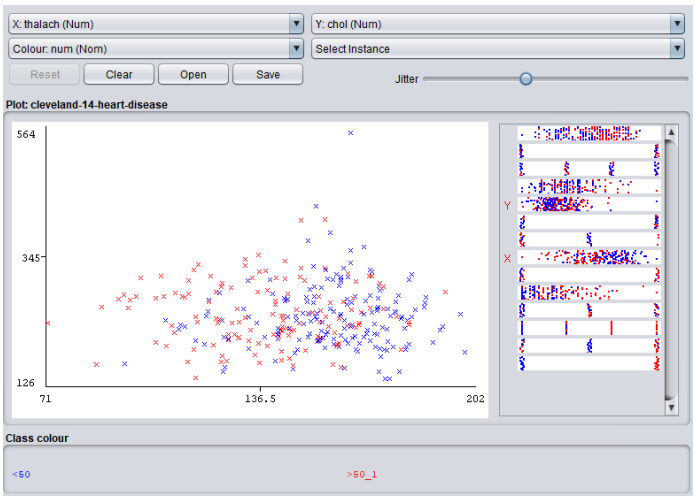
Como se pode ver nesta primeira imagem, o atributo age e o atributo thalac são inversamente proporcionais (correlação negativa), isto é, conforme a idade aumenta, o thalac tende a diminuir.



Como se pode ver nesta segunda imagem, os atributos tresbps e chol são diretamente proporcionais (correlação positiva), pois, à medida que o tresbps diminui o chol tende também a diminuir.

1.3.





Para os atributos thalach e oldpeak (primeira imagem) consegue-se observar que existe uma concentração de doenças cardíacas. Isto surge quando oldpeak está entre 0 e 3.1 e thalach entre 90 e 140.

Já para thalach e chol (segunda imagem) existe uma concentração do aumento do nível da doença cardíaca quando thalach se encontra entre 105 e 162 e chol entre 164 e 353. Como se pode ver nas imagens, em ambos é possível distinguir áreas "densas" de doenças cardíacas.

2.1. Usou-se o filtro Attribute Selection com o evaluator InfoGainAttributeEval e search Ranker. Esta filtragem dos dados que melhor representam o dataset na ótica da classe num são:

- thal
- cp
- ca
- oldpeak
- exang
- thalack
- restecg
- num

Apesar de nenhum destes ser os que identificámos na resposta à pergunta 1.1, alguns dos atributos apresentados na resposta à pergunta 1.3, o que significa que as associações multivariadas de atributos com o atributo classe (num) são de enorme importância para o processo de decisão.

2.2.

[a] Para resolver a questão de substituir missing values com o valor médio, usámos o filtro **ReplaceMissingValues** (este mecanismo usa a média para atributos numéricos e a moda para atributos nominais).

Foram eliminados os atributos nominais e de seguida foi usado um modelo de regressão linear para saber como se podiam substituir os valores em falta no atributo ca.

[b] Apesar de se ter obtido uma fórmula e substituídos os valores através desta fórmula (ver figura ao lado), há que notar que isto não é viável. Sendo que o *Root Relative Squared Error* é muito alto (com valores a rondar quase os 90%), esta substituição não trará benefícios ao dataset nem acaba por ser aplicável por dar origem a valores inadequados.

```
Linear Regression Model

ca =

      0.0331 * age +
      0.1859 * oldpeak +
     -1.3258

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0.4472
Mean absolute error                      0.6575
Root mean squared error                  0.8217
Relative absolute error                 86.8286 %
Root relative squared error             89.7349 %
Total Number of Instances               99
Ignored Class Unknown Instances                 4
```

2.3. Seguindo os passos sintetizados no guião, primeiro garantiu-se que não existia nenhum valor nulo com o **ReplaceMissingValues** (caso contrário não era possível aplicar o restante procedimento). Depois escolheu-se o filtro **InterquartileRange** com as configurações default para identificar os outliers (produzindo dois atributos extra: índices de outlier e extreme value). Utilizou-se estes dois para remover

os valores que eram outliers e extreme values através do filtro **RemoveWithValues** primeiro para o parâmetro Outlier e depois para o parâmetro Extreme Value. No fim de ter removido os valores encontrados, procedeu-se à remoção dos dois atributos extra criados do dataset (que neste momento já são inúteis) e guardou-se o dataset.

3.1.

a) Em todos os datasets (original e os resultantes do pré-tratamento), o algoritmo **OneR** escolheu o thal como o atributo mais preditivo. Nestas execuções usou-se o mecanismo 10-fold Cross Validation. Comparando com as conclusões obtidas na questão 1.1, é possível perceber que a nossa conclusão de atributos mais preditivos não estava correta e que, de acordo com este algoritmo, o atributo mais preditivo é, de facto, o thal com a atribuição possível de ver na figura.

```
thal:
        fixed_defect    -> >50_1
        normal  -> <50
        reversable_defect       -> >50_1
```

b) Em média, a precisão do algoritmo **OneR** com a estratégia de teste de 10-fold Cross Validation rondou os 72%. Por outro lado, com o mesmo algoritmo mas com a estratégia de teste de training set a precisão rondou os 77%. Isto justifica-se pelo facto de no caso do training set, este usa exatamente o mesmo conjunto de dados para treino e teste, pelo que, regra geral, tem sempre melhores resultados que os restantes mecanismos, apesar de não ser um método de avaliação tão fiável. Por outro lado, o mecanismo de Cross Validation divide o dataset em K grupos neste caso 10 e itera sobre os 10 grupos de forma a usar, em cada uma das iterações, o grupo da iteração em questão para teste e os restantes para treino, introduzindo um pouco mais de entropia no modelo, o que baixa o sucesso do teste, mas torna-a mais fiável.

3.2. Ao utilizar o algoritmo **JRip** nos vários datasets conclui-se que os resultados foram melhores quando se aplicou pruning, já que:

- Sem pruning: média de 78% de precisão
- Com pruning: média de 82% de precisão

Isto justifica-se pelo facto de que as lógicas de decisão unpruned são geralmente maiores e mais complexas. Ao usar pruning existe um passo adicional no algoritmo que olha para a decisão e procura por partes da decisão que possam ser removidas sem afetar nitidamente a performance do modelo. Isto permite que a decisão seja mais pequena e simples de entender, bem como reduz significativamente o risco de fazer overfit aos dados de treino, pelo que acaba por produzir melhores resultados aquando do teste do modelo em dados que não os de treino.

3.3. a) e b) Do que foi testado, a não utilização de pruning deu sempre resultados mais baixos, pelo que se corrobora o que foi verificado e concluído na questão anterior. Relativamente ao outro parâmetro sugerido, minNumObj, que corresponde ao número mínimo de instâncias/registos por folha conclui-se que quanto maior este número menos preciso é o modelo, já que quanto maior este número menor é a possibilidade de o algoritmo diferenciar as instâncias, resultando num maior erro de classificação. Relativamente a este último parâmetro verificou-se também que se este valor fosse 0 a percentagem

diminuiu comparativamente a se fosse 1 ou 2, o que nos leva a concluir que um valor demasiado pequeno também pode ser prejudicial, pois aumenta o risco de overfitting do dataset de treino.

3.4. Nesta pergunta, decidiu-se aplicar outros dois algoritmos: Logistic Regression e Naive Bayes. Tanto num como noutro obteve-se melhores resultados, comparando com os algoritmos dos exercícios anteriores. Deixa-se aqui as imagens dos resultados.

Logistic Regression:

```
Correctly Classified Instances         256               84.4884 %
Incorrectly Classified Instances        47               15.5116 %
Kappa statistic                          0.6852
Mean absolute error                      0.2143
Root mean squared error                  0.343
Relative absolute error                 43.1888 %
Root relative squared error             68.8783 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,891    0,210    0,835      0,891    0,862      0,687    0,909     0,918     <50
                 0,790    0,109    0,858      0,790    0,823      0,687    0,909     0,875     >50_1
Weighted Avg.    0,845    0,164    0,846      0,845    0,844      0,687    0,909     0,898
```

Naive Bayes:

```
Correctly Classified Instances         253               83.4983 %
Incorrectly Classified Instances        50               16.5017 %
Kappa statistic                          0.6661
Mean absolute error                      0.1846
Root mean squared error                  0.3634
Relative absolute error                 37.2065 %
Root relative squared error             72.9737 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,867    0,203    0,836      0,867    0,851      0,667    0,904     0,918     <50
                 0,797    0,133    0,833      0,797    0,815      0,667    0,904     0,883     >50_1
Weighted Avg.    0,835    0,171    0,835      0,835    0,835      0,667    0,904     0,902
```

4.2. K-Means construiu várias partições de objectos e depois avalia cada uma usando um critério. O algoritmo recebe um input com n objetos e um k definido pelo utilizador e retorna um conjunto com k clusters que minimiza o critério de erro quadrado.

Neste caso, o K mais adequado é quando este é igual a 2 (representado na figura em baixo)..

```
Cluster 0: 65,female,non_anginal,140,417,t,left_vent_hyper,157,no,0.8,up,1,normal
Cluster 1: 58,male,asympt,114,318,f,st_t_wave_abnormality,140,no,4.4,down,3,fixed_defect

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                Full Data              0                1
                           (303.0)          (169.0)          (134.0)
=================================================================================
age                         54.3663          52.1006          57.2239
sex                            male             male             male
cp                           asympt       non_anginal           asympt
trestbps                   131.6238          129.497          134.306
chol                       246.264          240.4024         253.6567
fbs                              f                f                f
restecg                     normal           normal   left_vent_hyper
thalach                    149.6469         159.0828         137.7463
exang                            no               no              yes
oldpeak                     1.0396           0.5805           1.6187
slope                           up               up             flat
ca                          0.6745           0.4302           0.9826
thal                        normal           normal reversable_defect



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      169 ( 56%)
1      134 ( 44%)
```

Após comparar com o clustering quando 2 > k >= 10, quando k=2 existe uma distribuição mais uniforme dos dados, sendo portanto este o mais adequado.

4.3. e 4.4. Para estudar as medidas apresentadas vai ser analisado o resultado de clustering quando k = 2.

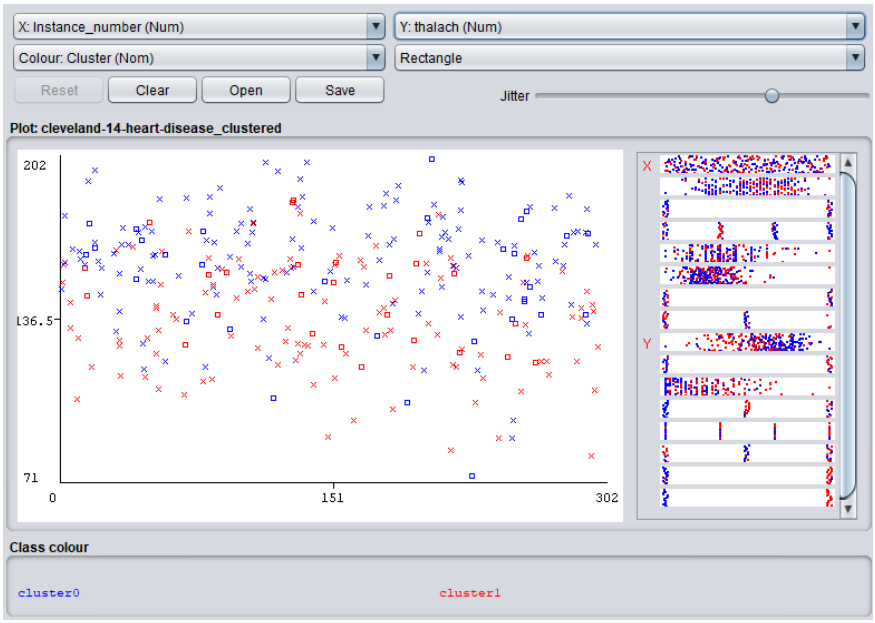
```
Final cluster centroids:
                                                Cluster#
Attribute                       Full Data             0                1
                                  (303.0)       (169.0)          (134.0)
=============================================================================
age                              54.3663       52.1006          57.2239
                                +/-9.0821     +/-9.1753        +/-8.1367

sex                                 male          male             male
  female                      96.0 ( 31%)   69.0 ( 40%)      27.0 ( 20%)
  male                       207.0 ( 68%)  100.0 ( 59%)     107.0 ( 79%)

cp                                asympt    non_anginal           asympt
  typ_angina                  23.0 (  7%)   12.0 (  7%)      11.0 (  8%)
  asympt                     143.0 ( 47%)   34.0 ( 20%)     109.0 ( 81%)
  non_anginal                 87.0 ( 28%)   76.0 ( 44%)      11.0 (  8%)
  atyp_angina                 50.0 ( 16%)   47.0 ( 27%)       3.0 (  2%)

trestbps                        131.6238      129.497          134.306
                                +/-17.5381    +/-15.9277       +/-19.1045

chol                             246.264       240.4024         253.6567
                                +/-51.8308    +/-47.5201       +/-56.1148

fbs                                    f             f                f
  t                           45.0 ( 14%)   24.0 ( 14%)      21.0 ( 15%)
  f                          258.0 ( 85%)  145.0 ( 85%)     113.0 ( 84%)

restecg                           normal        normal   left_vent_hyper
  left_vent_hyper            147.0 ( 48%)   57.0 ( 33%)      90.0 ( 67%)
  normal                     152.0 ( 50%)  111.0 ( 65%)      41.0 ( 30%)
  st_t_wave_abnormality        4.0 (  1%)    1.0 (  0%)       3.0 (  2%)
```

Para os clusters 0 e 1, podemos notar que o sexo é escolhido como masculino nos 2, cp tem 2 pontos diferentes, nomedamente non_anginal no primeiro cluster e asympt no segundo (este segundo correspondeu ao full data), etc.

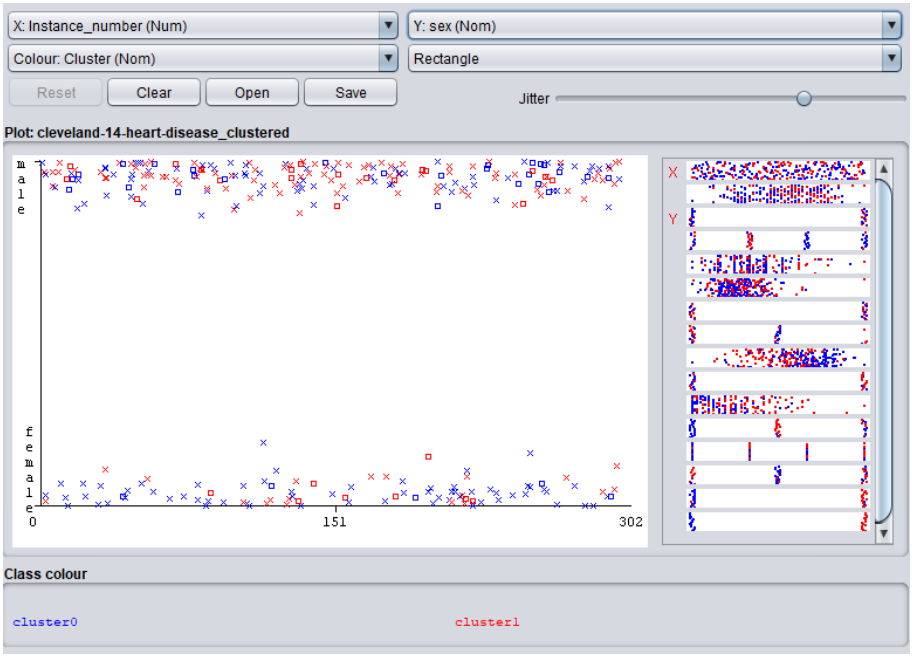
O mais importante a notar ainda é o desvio padrão. Em valores numéricos, como no caso de age, o centróide definido para o cluster 0 é de 52.1006 com um desvio padrão de 9.1753. Isto indica-nos que este cluster aglomera valores de age entre 52.1006 e a distância euclidiana de 9.1753 (mais ou menos este valor). Para os valores nominais é apresentado o nome do atributo e os valores possíveis que este pode ter. No caso do restecg, para o cluster 0 temos que normal foi o ponto escolhido para este cluster. Como pode ser observado na figura em cima normal foi a escolha deste cluster pois apresenta a maior percentagem (65% vs 33% para left_vent_hyper e 0% para st_t_wave_abnormality).

4.4.

Para certos casos (ex: thalac(imagem em baixo)), os clusters não são muito fáceis de descrever pois a dispersão apresentada torna este processo muito complicado.



No entanto, para casos como o sex conseguimos compreender a distribuição apresentada. Neste caso conseguimos observar que o sexo feminino está maioritariamente concentrado no cluster0 (notar: apesar de apresentar mais instancias no cluster0 não foi este o ponto definido para o cluster, mas sim sex = male. Isto deve-se a algo abordado anteriormente: número de instâncias do sexo masculino ser mais)



4.6.

```
Correctly Classified Instances         246               81.1881 %
Incorrectly Classified Instances        57               18.8119 %
Kappa statistic                          0.6187
Mean absolute error                      0.2317
Root mean squared error                  0.3925
Relative absolute error                 46.6963 %
Root relative squared error             78.8036 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.855    0.239    0.810      0.855    0.832      0.620    0.829     0.823     <50
                 0.761    0.145    0.814      0.761    0.787      0.620    0.829     0.742     >50_1
Weighted Avg.    0.812    0.196    0.812      0.812    0.811      0.620    0.829     0.786

=== Confusion Matrix ===

   a   b   <-- classified as
 141  24 |   a = <50
  33 105 |   b = >50_1


                J48 pruned tree
                ------------------

                cluster = cluster1
                |   ca <= 0: <50 (121.88/12.0)
                |   ca > 0
                |   |   age <= 62
                |   |   |   exang = no
                |   |   |   |   slope = up
                |   |   |   |   |   sex = female: <50 (6.0/1.0)
                |   |   |   |   |   sex = male
                |   |   |   |   |   |   fbs = t: <50 (4.0)
                |   |   |   |   |   |   fbs = f
                |   |   |   |   |   |   |   age <= 43: <50 (2.56)
                |   |   |   |   |   |   |   age > 43: >50_1 (12.28/1.28)
                |   |   |   |   slope = flat: >50_1 (9.28/2.28)
                |   |   |   |   slope = down: <50 (1.0)
                |   |   |   exang = yes: <50 (2.0)
                |   |   age > 62: <50 (10.0)
                cluster = cluster2
                |   ca <= 0
                |   |   exang = no
                |   |   |   age <= 52: >50_1 (7.0/2.0)
                |   |   |   age > 52: <50 (14.0/2.0)
                |   |   exang = yes
                |   |   |   restecg = left_vent_hyper
                |   |   |   |   thal = fixed_defect: >50_1 (0.0)
                |   |   |   |   thal = normal: <50 (5.0)
                |   |   |   |   thal = reversable_defect: >50_1 (13.42/2.0)
                |   |   |   restecg = normal: >50_1 (16.0/2.0)
                |   |   |   restecg = st_t_wave_abnormality: >50_1 (1.0)
                |   ca > 0: >50_1 (77.58/4.0)

                Number of Leaves  :     17

                Size of the tree :      30
```

Este novo classificador resultante do uso do algoritmo J48, apresenta melhores resultados em questão de desempenho comparativamente com OneR e J48. No entanto, este classificador não supera todos os anteriores, visto que fica um pouco abaixo dos valores atingidos com o algoritmo JRip.

5.1. Considerando que este é um problema de classificação, as medidas que se pensa ser mais adequada são: precisão, recall e F1 score. Deste 3 o mais importante será o F1 score já que, de certa forma, a precisão e o recall avaliam conceitos diretamente opostos e o F1 score é capaz de nível e calcular uma métrica que têm em conta as duas anteriores através de um meio termo.

5.2.

	OneR			**JRip**			**J48**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
heart-c	0.718	0.716	0.717	0.815	0.815	0.815	0.779	0.779	0.778
heart-c1	0.718	0.716	0.717	0.803	0.802	0.801	0.786	0.785	0.784
heart-c2	0.721	0.719	0.719	0.818	0.818	0.818	0.788	0.848	0.812
heart-c3	0.718	0.716	0.717	0.828	0.828	0.828	0.789	0.789	0.788
heart-c4	0.718	0.716	0.717	0.815	0.815	0.815	0.779	0.779	0.778

5.3. A primeira conclusão a retirar é que as alterações feitas nos dados que deram origem aos vários datasets teve pouco ou nenhum efeito claramente visível a nível de resultados (isto também pode ser do facto que na realidade os problemas e outliers existentes no dataset eram quase inexistentes). Para além disto, dá para concluir que tendo em conta os resultados o algoritmo com maior sucesso é o JRip. Para além disso este é também o mais consistente ao nível do F1 Score, o que significa que permite encontrar um bom meio termo entre a precisão e o recall.