FICHA 6

DATA UNDERSTANDING

1. For each attribute find the following information.
   a. The attribute type, e.g. nominal, ordinal, numeric.
      R.: AGE – Numeric
      SEX – Nominal
      CP – Nominal
      TRESTBPS – Numeric
      COL – Numeric
      CHOL – Numeric
      FBS – Nominal
      RESTECG – Nominal
      THALACH – Numeric
      EXANG – Nominal
      OLDPEAK – Numeric
      SLOPE – Nominal
      CA – Numeric
      THAL – Nominal
      NUM – Nominal

   b. Percentage of missing values in the data.
      R.: 2% de missing data no atributo CA e 1% de missing data no atributo THAL.

   c. Max, min, mean, standard deviation.
      R.: AGE – Min 29, Max 77, Mean 54.366, StdDev 9.082
      SEX – Atributo Nominal
      CP – Atributo Nominal
      TRESTBPS – Min 94, Max 200, Mean 131.624, StdDev 17.538
      COL – Min 0, Max 1, Mean 0.275, StdDed 0.118
      CHOL – Min 126, Max 564, Mean 246.264, StdDev 51.831
      FBS – Atributo Nominal
      RESTECG – Atributo Nominal
      THALACH – Min 71, Max 202, Mean 149.647, StdDev 22.905
      EXANG – Atributo Nominal
      OLDPEAK – Min 0, Max 6.2, Mean 1.04, StdDev 1.161
      SLOPE – Atributo Nominal
      CA – Min 0, Max 3, Mean 0.674, StdDev 0.938
      THAL – Atributo Nominal
      NUM – Atributo Nominal

   d. Are there any records that have a value for the attribute that no other record has?
      R.: Sim: age, trestbps, col, chol, thalach, oldpeak.

e. Study the histogram at the lower right and informally describe how the attribute seems to influence the risk for heart disease. What does it mean the pop-up messages that appear when dragging the mouse over the graphic?
R.: A mensagem pop-up que aparece indica o valor/label do atributo

Análise dos histogramas por atributo:

AGE – Aparentemente, quanto maior a idade, maior o risco de doença cardíaca;

SEX – À primeira vista, parece que a probabilidade de um homem ter uma doença cardíaca é superior à das mulheres. No entanto, há mais registos de homens do que de mulheres, sendo que a comparação não é muito fiável;

CP – A dor no peito do tipo asympt tem elevada probabilidade de vir a resultar em doença cardíaca, pelos registos do dataset;

TRESTBPS – A resting blood pressure mais elevada (elevada entre 94 e 147, já que depois deste valor há poucos registos) indica um maior risco de doença;

COL – Para a zona do histograma onde existem registos ([126;345] aprox), conclui-se que o risco de doença é maior para valores de colesterol entre [180;262];

CHOL – " " "

FBS – Embora haja mais registos no dataset para nível de açúcar no sangue = false, o risco de ter doença cardíaca é semelhante, independentemente da presença de açúcar;

RESTECG – Os resultados eletrocardiográficos em repouso do tipo "left_hyper_vent" aparenta ter uma maior probabilidade de resultar em doença;

THALACH – Aparentemente, quanto maior o batimento cardíaco máximo atingido, maior a probabilidade do indivíduo vir a ter uma doença;

EXANG – Em caso positivo de angina (dor no peito) induzida por exercício, a probabilidade de ter uma doença é maior;

OLDPEAK – Em valores inferiores de depressão do segmento ST induzida pelo exercício em relação ao repouso, a probabilidade de doença cardíaca é inferior (e vice versa);

SLOPE – Em caso de "flat" na inclinação do segmento ST de pico de exercício, há maior probabilidade de doença;

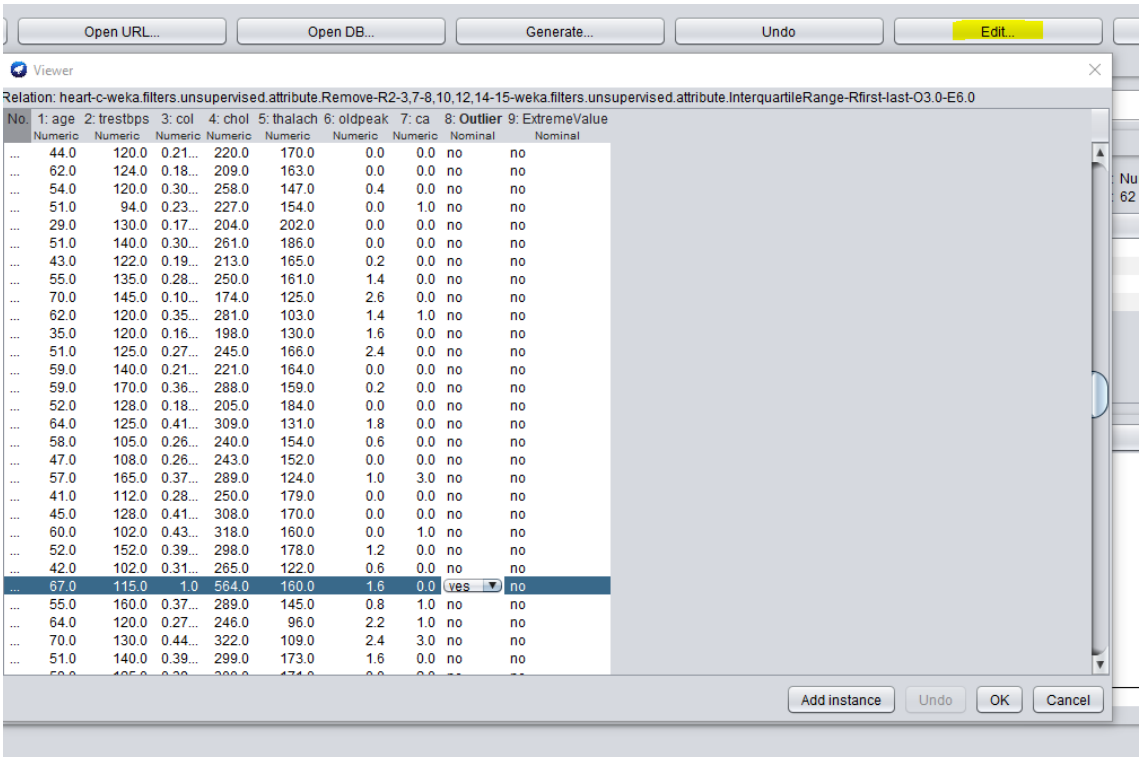
CA – No caso do valor 0 ou 1 do número de grandes vasos (0-3) (colorido por fluoroscopia), a probabilidade de doença é mais elevada;

THAL – Para o tipo "reversable_defect", há maiores casos de doença cardíaca;

NUM – é a class do dataset.

f. Are there any outliers for the attribute under consideration?

i. Investigate the possibility of using the Weka filter InterquartileRange to detect outliers 2.
R.: (1º Eliminar todos os valores nominais, deixando os numéricos)
Existe um outlier.



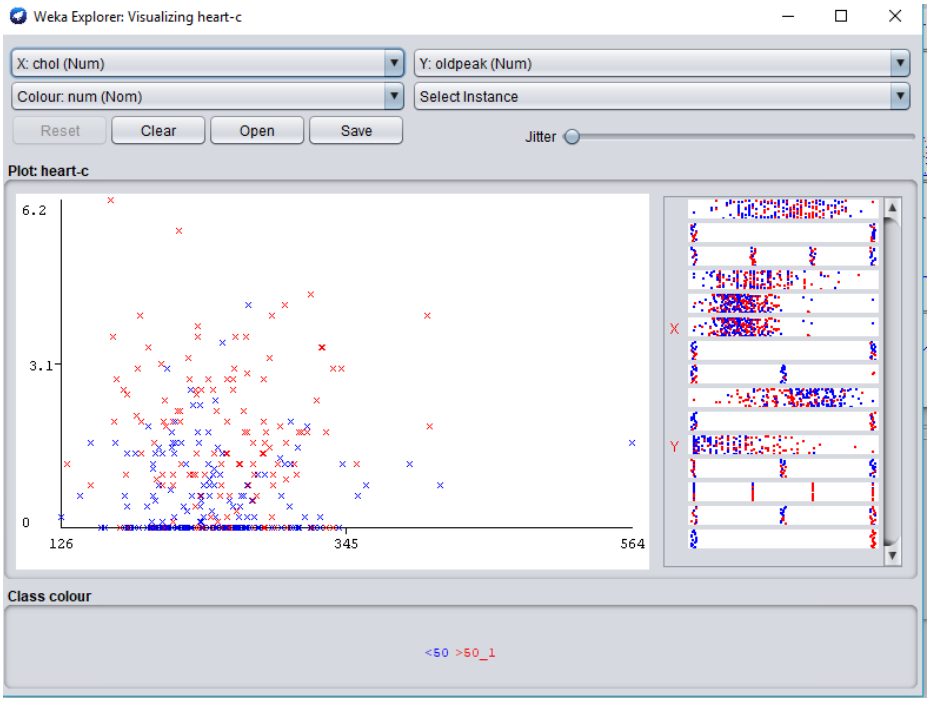
2. Switch to the Visualize tab on the upper part of the screen to visualize 2D-scatter plots for each pair of attributes.
    a. Which attributes seem to be the most/least linked to heart disease? Summarize in a table your findings concerning the predictive value of each attribute.
    R.:

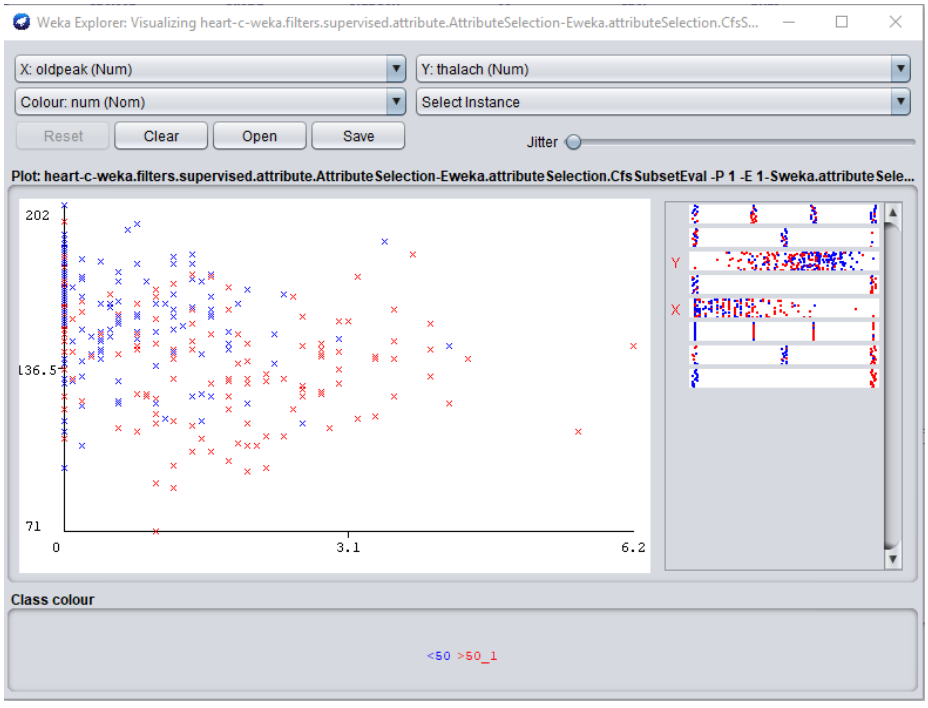
Age	Menos relacionada
Sex	Mais relacionada, homens com mais probabilidade
Cp	Mais relacionada, asympt com mais probabilidade
Trestbps	Menos relacionada
Col	Menos relacionada
Chol	Menos relacionada
Fbs	Menos relacionada
Restecg	Menos relacionada
Thalach	Mais relacionada, quanto menor o valor de Thalach maior a probabilidade
Exang	Mais relacionada, se o valor for “Yes” maior a probabilidade
Oldpeak	Mais relacionada, quanto maior o valor de Oldpeak maior a probabilidade
Slope	Menos relacionada
Ca	Mais relacionada, se o valor for 0, menor a probabilidade
Thal	Menos relacionada

b. Does any pair of attributes seem to be correlated?
R.: Chol e o Col.

3. Investigate also possible multivariate associations of attributes with the class attribute, i.e. study scatter plots of two attributes X and Y and try to identify possible ”dense” heart disease areas (if any).

a. If you find ”dense” heart disease areas in any scatter plot then quantify the heart disease rate in these areas with respect to the entire data set.
R.: Entre thalack e oldpeak é possível encontrar áreas diferenciadas de probabilidade de doença. Também entre oldpeak e chol.





DATA PREPROCESSING

1. Attribute selection.
   Investigate the possibility of using the Weka filter AttributeSelection for selecting a subset of attributes with good predicting capability. Then, describe briefly the filter(s) you used and compare the results you obtained with the conclusions you obtained in the previous section. Save the dataset with the selected attributes in the file heart-c1.arff.
   R.: O filtro "AttributeSelection" é um filtro supervisionado de atributos que pode ser usado para selecionar atributos. É bastante flexível e permite que vários métodos de avaliação e de pesquisa sejam combinados. Se este filtro for usado, as conclusões combinam com as da secção anterior, o que dá uma segurança extra para o seu uso.

2. Handling missing values. Consider the following methods for handling missing values and investigate each possibility within Weka. Note that, as rule of thumb, if an attribute has more than 5% missing values then the records should not be deleted and it is advisable to impute values where data is missing, using a suitable method.
   R.: Não existem atributos com mais de 5% de valores em falta, sendo que não foram apagados quaisquer registos.

   a. Replace the missing values by the attribute mean, if the attribute is numeric. Otherwise, replace missing values by attribute mode (if the attribute is categorical). Save the dataset you obtained without missing values in the file heart-c2.arff.
      R.: (Foi utilizado o filtro "ReplaceMissingValues")

   b. Investigate the possibility of using (linear) regression to estimate the missing values for each attribute. Save the dataset you obtained without missing values in the file heart-c3.arff.
      R.: Depois de selecionar apenas os atributos numéricos e aplicar o classificador "LinearRegression", foi obtida a fórmula:

```
ca =
        0.0331 * age +
        0.1859 * oldpeak +
        -1.3258
```

      Na próxima secção encontram-se os resultados comparativos de cada algoritmo com os dois datasets (dataset com valores substituídos pela formula da regressão linear e dataset da média).

3. Eliminating outliers.
   Eliminate the outlier records and save the dataset you obtained without outliers in the file heart-c34.arff.

MINING THE DATA

(usar dataset heart-c_MEAN.arff (substituição dos valores em falta pela média) e

heart-c_REG_LIN_FIXED.arff (substituição dos valores com ajuda de regressão linear))

1. Start with OneR classifier.
   a. What can you conclude? Compare your conclusions with your previous conclusions obtained in section 1.1.
   R.: O algoritmo OneR gera apenas uma regra para cada preditor nos dados, e, de seguida, seleciona a regra com o menor erro total como "regra única". Assim sendo, a regra obtida foi a seguinte:

   Dataset heart-c_MEAN.arff

   thal:

   fixed_defect -> >50_1
   normal -> <50
   reversable_defect -> >50_1

   Dataset heart-c_REG_LIN_FIXED

   thal:

   fixed_defect -> >50_1
   normal -> <50
   reversable_defect -> >50_1
   ? -> <50

   Podemos concluir que o tal é o atributo que mais contribui para doença cardíaca (de acordo com este algoritmo), se estiver fora dos valores normais. Assim, se não for do tipo "normal", i.e., se for "fixed_defect" ou "reversable_defect" existe o risco de doença cardíaca. Esta conclusão é semelhante à obtida na secção anterior, mas com mais precisão (conclusão anterior: para tal = "reversable_defect", há maiores casos de doença cardíaca)

   Conclui-se também que não há qualquer diferença, para já, entre usar regressão linear ou a média para substituir os valores em falta.

   b. Compare the accuracy of the classifier on the training set with the accuracy estimation obtained through 10 fold-cross validation. How do you explain the difference (if any)?
   R.: Accuracy com cross-validation 10 folds: 71.9472%
   Accuracy com training set: 76.5677%

   (Estes valores são os mesmos para os dois datasets.)
   É normal que a accuracy seja superior com o uso de training set para tipo de treino, porque há overfitting.

2. Use JRip classifier, i.e. the Weka version of the rule classifier RIPPER.
   a. Build a classifier with and without rule pruning. Which one is preferable? Motivate your answer.
   R.: É preferível o uso de pruning.

   Dataset heart-c_MEAN.arff [WITH PRUNNING] 80.8581% acertos
   [WITHOUT PRUNNING] 77.2277% acertos

   Dataset heart-c_REG_LIN_FIXED [WITH PRUNNING] 81.1881%
   [WITHOUT PRUNNING] 77.8878%

   b. Describe the patterns you obtained and compare with your previous conclusions.
   R.: Não há diferenças muito significativas entre os dois dataset, uma vez que a accuracy de ambos sob o mesmo ambiente é semelhante.

3. Use J48 classifier, i.e. the Weka version of the decision tree classifier C4.5.
   a. Investigate the use of different J48's parameters such as pruning and minimum number of records in the leaves.
   R.:

   Dataset heart-c_MEAN.arff
   [WITH PRUNNING & MIN OBJ = 2] 78.8779% acertos
   [WITHOUT PRUNNING & MIN OBJ = 2] 77.5578% acertos
   [WITH PRUNNING & MIN OBJ = 3] 79.8680% acertos
   [WITHOUT PRUNNING & MIN OBJ = 3] 77.8878% acertos

   AQUI
   Dataset heart-c_REG_LIN_FIXED
   [WITH PRUNNING & MIN OBJ = 2] 79.2079% acertos
   [WITHOUT PRUNNING & MIN OBJ = 2] 78.2178% acertos
   [WITH PRUNNING & MIN OBJ = 3] 77.5578% acertos
   [WITHOUT PRUNNING & MIN OBJ = 3] 77.2277% acertos

   b. Describe the patterns you obtained and compare with your previous conclusions.
   R.: Não há diferenças muito significativas entre os dois dataset, uma vez que a accuracy de ambos sob o mesmo ambiente é semelhante.

CLUSTERING TENDENCY

Investigate whether there is a clustering tendency in the dataset. You may start by clustering the data with SimpleKMeans algorithm, for some $2 \leq k \leq 10$.

1. Do not use the class attribute, num, for clustering.
   R.:

   - COM K=2: consegue-se verificar a existência de algum clustering em cp (anon_anginal e asympt), restecg (left_vent_hyper e normal), exang (no e yes), slope (up e flat) e, por fim, thal (normal e reversable_defect). As distribuição de clustered instances é 56% e 44%. No entanto, apesar de se poder verificar a existência de alguns clusters, estes não têm grande valor pois a distribuição é quase 50/50.

   - COM K=5: as distribuições continuam muito próximas de 100/num_clusters = 20%, ou seja, os resultados são pouco expressivos.

   - COM K=6: é possível começar a verificar algumas tendências. Por exemplo, o cluster 5 apresenta apenas 7% das instâncias. O squared error é de 486.

   - COM K=7: as tendências aumentam a sua expressividade. Desta vez, dois clusters possuem % de instâncias menor, a saber, o cluster 1 (3%) e o 5 (5%). Também existem clusters com maior % de instâncias do que se verificaria numa distribuição linear, a saber, cluster 4 (24%) e 6 (26%). O squared error é de 487.

   - COM K=8: existem na mesma dois clusters com % de instâncias inferiores ao expectável e outros dois com superiores. De qualquer modo, o squared error é de 464, ou seja, diminuiu consideravelmente.

   - COM K=9 e K=10: as mudanças são menores e começa a verificar-se uma tendência a aproximar da distribuição linear, ou seja, não é possível verificar clusters com tanta precisão. A divisão é demasiado grande.

2. Find a suitable value for k, i.e. the number of clusters you are going to build. Justify your choice of k.
   R.: Tendo em conta os resultados da questão anterior, pode-se verificar que os melhores K são K=7 ou K=8. Isto acontece pois, para K pequeno não há diferenciação suficiente, enquanto que para um K maior que 8 existe uma dispersão inadvertida dos dados. K=7 tem 4 valores relativamente distantes do linear, enquanto que K=8 tem 5. K=8 acaba por ser a escolha a fazer.

3. Use class to cluster evaluation and make sure that standard deviations are also computed for numerical attributes.
   R.: Representa parte dos dados obtidos.

Final cluster centroids:

Attribute	Full Data (303.0)	Cluster# 0 (41.0)	1 (9.0)	2 (32.0)	3 (44.0)	4 (71.0)
age	54.3663 +/-9.0821	58.6829 +/-8.5013	60.1111 +/-3.1402	51.7813 +/-10.5608	57.6364 +/-7.5053	47.3662 +/-7.5767
sex	male	female	female	male	male	male
male	207.0 (68%)	1.0 (2%)	2.0 (22%)	30.0 (93%)	40.0 (90%)	54.0 (76%)
female	96.0 (31%)	40.0 (97%)	7.0 (77%)	2.0 (6%)	4.0 (9%)	17.0 (23%)
cp	asympt	non_anginal	asympt	asympt	asympt	non_anginal
typ_angina	23.0 (7%)	3.0 (7%)	1.0 (11%)	4.0 (12%)	6.0 (13%)	1.0 (1%)
asympt	143.0 (47%)	4.0 (9%)	8.0 (88%)	22.0 (68%)	26.0 (59%)	9.0 (12%)
non_anginal	87.0 (28%)	25.0 (60%)	0.0 (0%)	3.0 (9%)	9.0 (20%)	39.0 (54%)
atyp_angina	50.0 (16%)	9.0 (21%)	0.0 (0%)	3.0 (9%)	3.0 (6%)	22.0 (30%)
trestbps	131.6238 +/-17.5381	133.5366 +/-14.9434	152.8889 +/-24.902	124.875 +/-14.738	129.7273 +/-18.1715	126.169 +/-14.7619
col	0.2746 +/-0.1183	0.329 +/-0.1615	0.3737 +/-0.1817	0.2602 +/-0.0914	0.2818 +/-0.1008	0.2351 +/-0.0955
chol	246.264 +/-51.8308	270.0976 +/-70.7346	289.6667 +/-79.566	239.9688 +/-40.0278	249.4318 +/-44.1717	228.9718 +/-41.8173
fbs	f	f	f	f	f	f
t	45.0 (14%)	5.0 (12%)	1.0 (11%)	2.0 (6%)	5.0 (11%)	6.0 (8%)
f	258.0 (85%)	36.0 (87%)	8.0 (88%)	30.0 (93%)	39.0 (88%)	65.0 (91%)
restecg	normal	left_vent_hyper	left_vent_hyper	left_vent_hyper	left_vent_hyper	normal
left_vent_hyper	147.0 (48%)	26.0 (63%)	8.0 (88%)	26.0 (81%)	40.0 (90%)	3.0 (4%)
normal	152.0 (50%)	14.0 (34%)	0.0 (0%)	6.0 (18%)	4.0 (9%)	68.0 (95%)

4. Study the numerical measures displayed by Weka for each cluster. What can you conclude?
   R.: Pode-se verificar que os clusters existentes estão preenchidos, maioritariamente, por elementos do tipo que referem. Os desvios existentes também não são demasiado altos, nunca atingindo os 10% em relação ao valor de referência, comprovando a qualidade do clustering.

5. Use Visualize cluster assignments and try to discover a description for each cluster.
   R.:
6. Investigate the possibility of building a classifier for finding rules describing the clusters. Compare the results with your previous conclusions.
   R.:
7. Investigate the possibility of using the cluster information to build a classifier for num. Compare your results with what you obtained in section 1.3. Do you get a better classifier?
   R.:

PERFORMANCE PREDICTING

1. Weka outputs several performance measures. Choose some of the performance measures and motivate your choice.
   R.:
2. Summarize in a table the performance measures for each classifier and each dataset.
   R.:
3. What can you conclude?
   R.:

CONCLUSIONS

1. Describe your final conclusions and indicate which risk factors for heart disease have you found in the data.
   R.: