Ficha1-DC

João Nunes (A82300)
Luís Braga (A82088)

21/02/2020

Conteúdo

1 Exercício 1 **2**
1.1 a) Quantas instâncias (registos) tem este data set? . . . 2
1.2 b) Quantos atributos (colunas) tem este data set? . . . 2
1.3 c) Quantos e quais os valores possíveis para o atributo "age"? . . . 2
1.4 d) Quais os valores possíveis para o atributo "contact-lens"? . . . 2
1.5 d) Qual o atributo que tem "reduced" como um dos valores? . . . 2

2 Exercício 2 **3**
2.1 a) Quantas instâncias registos tem este data set? . . . 3
2.2 b) Quantos atributos (colunas) tem este data set? . . . 3
2.3 c) A classe "iris-setosa"tende a ter maiores ou menores valores de "sepal.length"? . . . 3
2.4 d) A classe "iris-viginica" tende a ter maiores ou menores valores de "petal.width"? . . . 3
2.5 e) Qual destes atributos, sozinho, parece dar uma melhor indicação da "class"? . . . 4

3 Exercício 3 **4**
3.1 a) Identificar quais os atributos deste data set? . . . 4
3.2 b) A utilização de um algoritmo de classificação poderá trazer conhecimento específico através dos dados apresentados. Indique um objetivo que possa ser atingido com a aplicação de algoritmos de classificação,quando o mesmo for executado em dados semelhantes, mas previamente desconhecidos. 5

4 Exercício 4 **5**
4.1 b) Observar a "Confusion Matrix"e indicar quais as maiores falhas no processo de classificação. . . . 5
4.2 c) Qual o número de "headlamps" que foram classificadas como "build wind float"? . . . 5
4.3 d) Qual o número de instâncias classificadas corretamente como "vehic wind non-float"? . . . 6
4.4 e) Qual o número de instâncias classificadas corretamente como "vehic wind float"? . . . 6
4.5 f) Na lista de resultados obtidos clicar com o botão direito e selecionar "Visualize tree". Copiar os resultados para a ficha de solução e descrever sucintamente o processo de classificação do algoritmo. . 7

5 Exercício 5 **7**
5.1 a) Correr o algoritmo de classificação J48 com os parâmetros por defeito. Indicar a percentagem de instâncias corretamente classificadas. . . . 7
5.2 b) Utilizando somente 2 casas decimais, abra a configuração do algoritmo J48 e coloque a opção "unpruned" a "True". Corra novamente a classificação e indique a percentagem de instâncias corretamente classificadas . . . 8

6 Exercício 4 continuação **9**
6.1 a) Retirar o atributo "Fe". Qual o resultado da classificação? . . . 9
6.2 b) Retirar todos excepto "Ri", "Mg". Qual o resultado da classificação? . . . 9

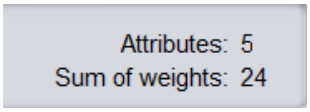
1 Exercício 1

1.1 a) Quantas instâncias (registos) tem este data set?

O dataset possui 24 registos, como se pode ver na seguinte figura.

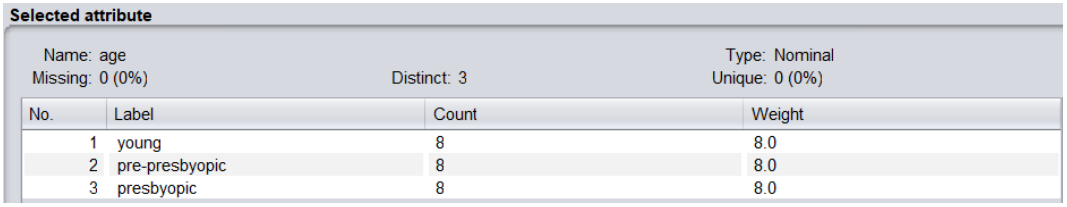
1.2 b) Quantos atributos (colunas) tem este data set?

O dataset possui cerca de 5 colunas, sendo que o mesmo pode ser comporvado por uma análise da *Current relation*.



1.3 c) Quantos e quais os valores possíveis para o atributo “age”?

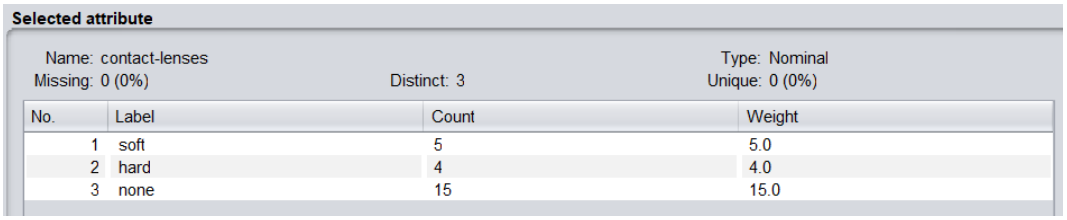
Existem três valores possíveis para o atributo age, o valor "young", "pre-presbyopic"e "presbyopic". Como se poderá verificar na seguinte figura.



No.	Label	Count	Weight
1	young	8	8.0
2	pre-presbyopic	8	8.0
3	presbyopic	8	8.0

1.4 d) Quais os valores possíveis para o atributo “contact-lens”?

O atributo "contact-lens"possui três valores possíveis, o valor "soft", "hard"ou "none".



No.	Label	Count	Weight
1	soft	5	5.0
2	hard	4	4.0
3	none	15	15.0

1.5 d) Qual o atributo que tem “reduced” como um dos valores?

O atributo que possui o valor "reduced"é o atributo "tear-prode-rate".

Selected attribute

Name: tear-prod-rate Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	reduced	12	12.0
2	normal	12	12.0

2 Exercício 2

2.1 a) Quantas instâncias registos tem este data set?

Neste caso o *data set* possui 150 registos.

Current relation

Relation: iris
Instances: 150

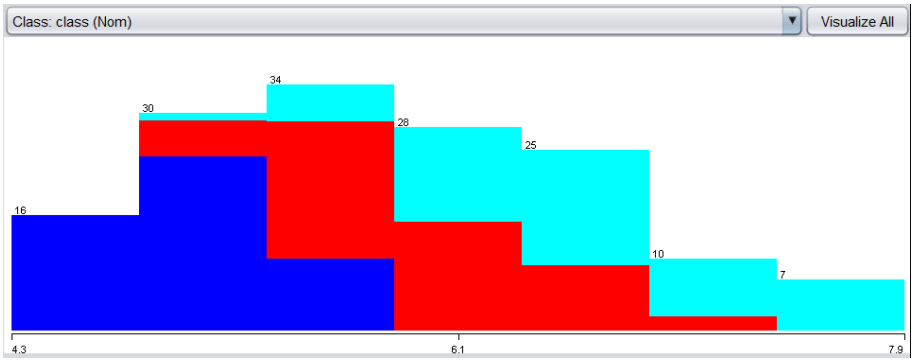
2.2 b) Quantos atributos (colunas) tem este data set?

O data set possui 5 atributos como se pode comprovar na figura abaixo.

Attributes: 5
Sum of weights: 150

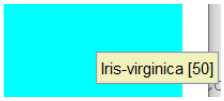
2.3 c) A classe “iris-setosa”tende a ter maiores ou menores valores de “sepal.length”?

Tendo em conta que a classe "iris-setosa"é identificada pela cor azul, no seguinte gráfico mostra-se a relação do "sepal-length"juntamente com o tipo de classe. De onde é possível verificar que a classe "iris-setosa"ao analisar o "speal-length"em média é menor que as outras classes.

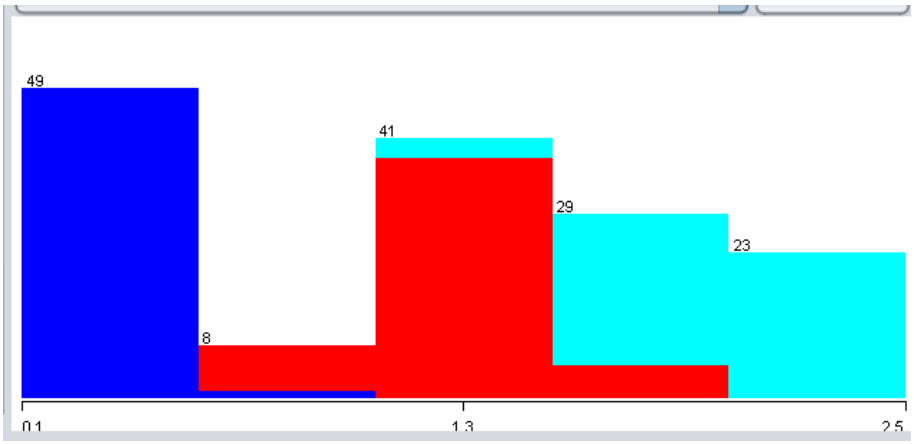


2.4 d) A classe “iris-viginica” tende a ter maiores ou menores valores de “petal.width”?

A classe "iris-virginica"representada pela cor apresentada abaixo:

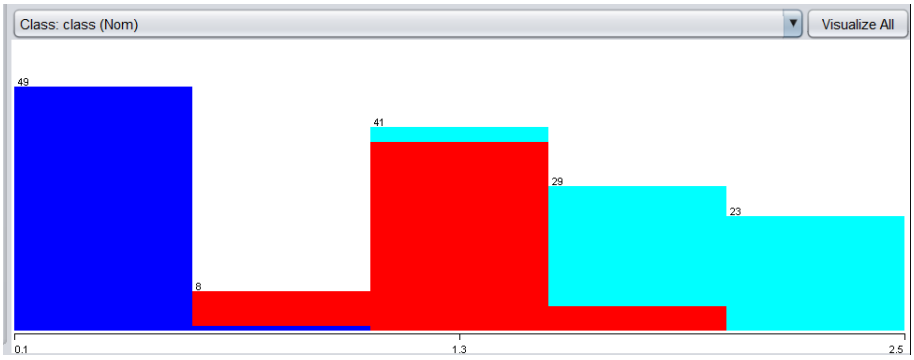


Tende a ter valores maiores de "petal.width"que as restantes classes como se pode ver na figura seguinte, em que a classe em questão domina a área com os maiores valores.



2.5 e) Qual destes atributos, sozinho, parece dar uma melhor indicação da “class”?

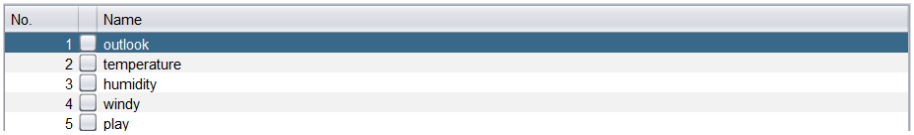
O atributo que dá uma melhor indicação da class é o atributo "petal-width"uma vez analisando este atributo cada classe encontra-se bem distribuída e facilmente identificável a partir do "petal-width"ao contrário dos outros atributos onde é mais misturada a distribuição das classes pelos valores dos atributos.



3 Exercício 3

3.1 a) Identificar quais os atributos deste data set?

Este *data set* possui cerca de 5 atributos, sendo eles o "outlook", "temperature", "humidity", "windy"e por fim o "play".



3.2 b) A utilização de um algoritmo de classificação poderá trazer conhecimento específico através dos dados apresentados. Indique um objetivo que possa ser atingido com a aplicação de algoritmos de classificação,quando o mesmo for executado em dados semelhantes, mas previamente desconhecidos.

Um algoritmo de classificação, e tendo em conta o caso apresentado com base nos outros dados ou seja a temperatura, humidade, vento e "outlook"consegue classificar e prever se pode jogar ou não.

4 Exercício 4

4.1 b) Observar a "Confusion Matrix"e indicar quais as maiores falhas no processo de classificação.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

A maior falha do algoritmo de classificação J48 concentra-se na previsão dos valore a e b, onde em 16 casos preveu como sendo b quando era verdadeiramente a, e em 15 casos preveu como sendo a quando era verdadeiramente b.

4.2 c) Qual o número de "headlamps" que foram classificadas como "build wind float"?

Foram três casos classificados como "build wind float".

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

4.3 d) Qual o número de instâncias classificadas corretamente como “vehic wind non-float”?

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Zero casos foram previstos como sendo d quando verdadeiramente eram d.

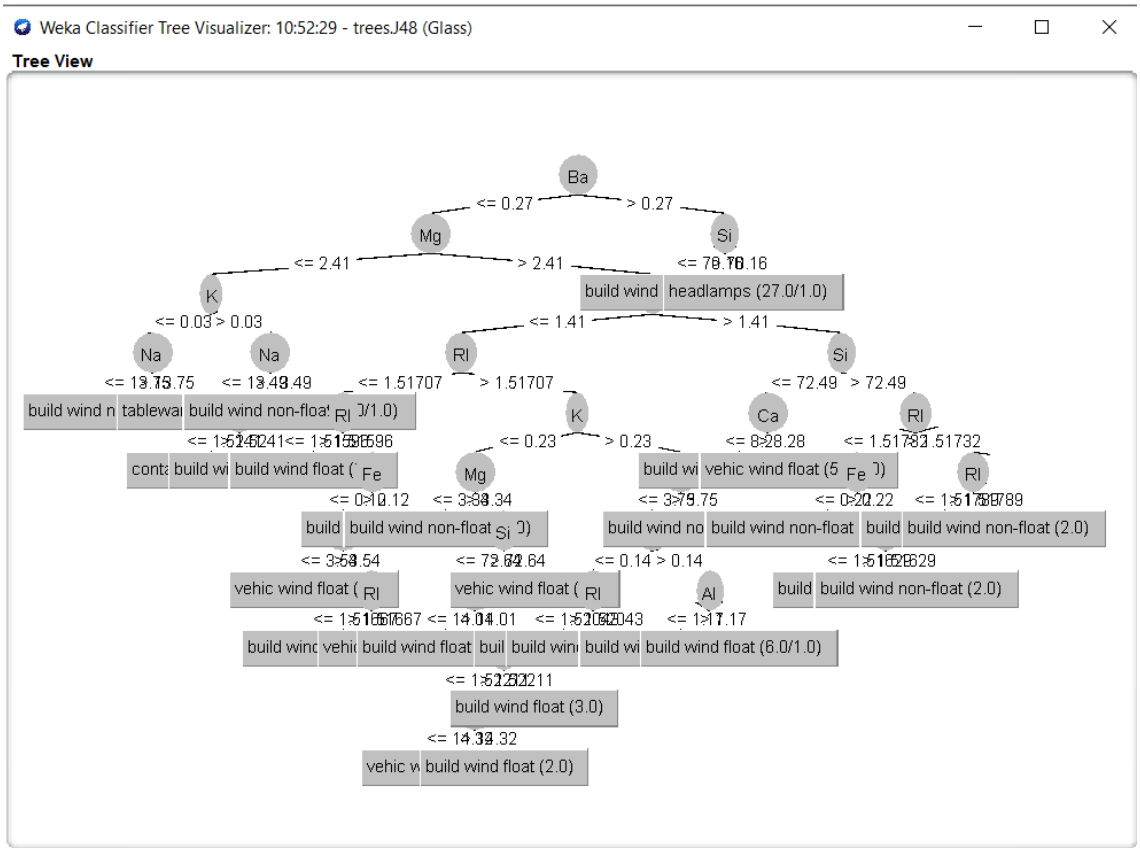
4.4 e) Qual o número de instâncias classificadas corretamente como “vehic wind float”?

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Seis casos foram previstos como sendo c quando verdadeiramente eram c.

4.5 f) Na lista de resultados obtidos clicar com o botão direito e selecionar "Visualize tree". Copiar os resultados para a ficha de solução e descrever sucintamente o processo de classificação do algoritmo.



O algoritmo de J48 é um algoritmo baseado em *decision trees* onde divide os dados de cada classe de forma adequada, onde cada novo caso que depois surge toma um dos valores dos ramos consoante o seu valor, conseguindo desta maneira fazer a previsão.

5 Exercício 5

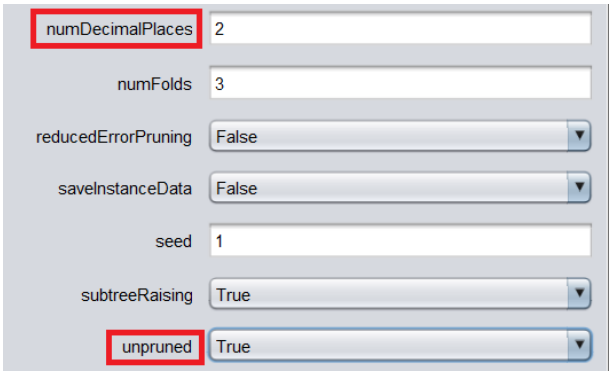
5.1 a) Correr o algoritmo de classificação J48 com os parâmetros por defeito. Indicar a percentagem de instâncias corretamente classificadas.

O algoritmo J48 com os parâmetros standard consegue prever corretamente 73.6842% dos casos.

```
=== Summary ===

Correctly Classified Instances          42               73.6842 %
Incorrectly Classified Instances        15               26.3158 %
Kappa statistic                          0.4415
Mean absolute error                      0.3192
Root mean squared error                  0.4669
Relative absolute error                 69.7715 %
Root relative squared error             97.7888 %
Total Number of Instances               57
```

5.2 b) Utilizando somente 2 casas decimais, abra a configuração do algoritmo J48 e coloque a opção "unpruned" a "True". Corra novamente a classificação e indique a percentagem de instâncias corretamente classificadas



Utilizando essas configurações o algoritmo conseguiu prever corretamente em 78.9474% dos casos.

```
=== Summary ===

Correctly Classified Instances          45               78.9474 %
Incorrectly Classified Instances        12               21.0526 %
Kappa statistic                          0.5378
Mean absolute error                      0.2677
Root mean squared error                  0.432
Relative absolute error                 58.5226 %
Root relative squared error             90.4708 %
Total Number of Instances               57
```

6 Exercício 4 continuação

6.1 a) Retirar o atributo “Fe”. Qual o resultado da classificação?

Depois de retirar o atributo "Fe", foi possível verificar que o algorimto preveu corretamente em 67.2895% dos casos e preveu incorretamente em 32.7103% dos casos.

```
=== Summary ===

Correctly Classified Instances         144               67.2897 %
Incorrectly Classified Instances        70               32.7103 %
Kappa statistic                          0.5519
Mean absolute error                      0.1029
Root mean squared error                  0.285
Relative absolute error                 48.5797 %
Root relative squared error             87.8206 %
Total Number of Instances              214
```

6.2 b) Retirar todos excepto “Ri”, “Mg”. Qual o resultado da classificação?

Removendo todos excepto esses dois atributos e o "Type"foi possível melhorar de maneira insignificante a percentagem de acerto para 68.6916% e preveu incorretamente em 31.3084%.

```
=== Summary ===

Correctly Classified Instances         147               68.6916 %
Incorrectly Classified Instances        67               31.3084 %
Kappa statistic                          0.5628
Mean absolute error                      0.1124
Root mean squared error                  0.267
Relative absolute error                 53.082  %
Root relative squared error             82.2535 %
Total Number of Instances              214
```