Universidade do Minho
Escola de Engenharia

Extração de Conhecimento de Dados Estruturados

Trabalho Prático

Mestrado Integrado em Engenharia Informática
Aprendizagem e Extração de Conhecimento

1º Semestre
2017-2018
Grupo 1

A74219 - Hugo Alves Carvalho
A70676 - Marcos Morais Luís
A74260 - Luís Miguel da Cunha Lima

10 de Dezembro de 2017
Braga

Resumo

Este documento relata o trabalho prático desenvolvido no âmbito da unidade curricular de **Aprendizagem e Extração de Conhecimento**, do perfil de especialização de **Sistemas Inteligentes**, tendo como tema principal a Extração de Conhecimento em conjuntos de dados organizados.

Conteúdo

Lista de Figuras

1 Introdução

Com o passar dos anos, a quantidade de dados que são coletados e armazenados, cresce a um ritmo explosivo, sendo que grande parte destes correspondem a dados não estruturados. Neste sentido, os métodos de análise e gestão de dados, realizador por ação humana, tornam-se demasiado dispendiosos e, em alguns casos, praticamente impossíveis. Assim, é fundamental a existência de técnicas que permitam extrair as informações mais importantes a partir de um conjunto de dados.

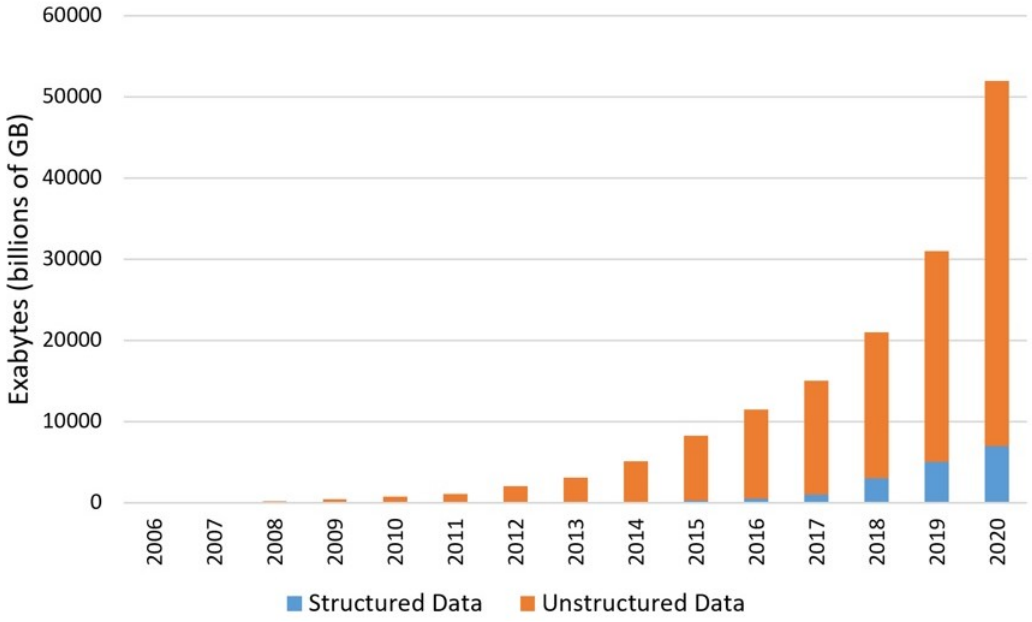

Figura 1: Crescimento dos dados estruturados vs não estruturados

Extração de Conhecimento corresponde a todo o processo de descoberta de conhecimento útil através de fontes de dados, sejam estas estruturadas ou não. Através de processos computacionais, o resultado deste processo deve construir informação legível e de fácil compreensão para o utilizador.

Com a realização deste trabalho, foram analisados três problemas distintos, com o objetivo de demonstrar e aprofundar os conhecimentos adquiridos ao longo do semestre. Para isso, foi utilizada a ferramenta *Weka*, fornecida pela universidade de Watako. Esta permite aplicar técnicas de extração de conhecimento sobre um conjunto de dados (*Dataset*) tais como associação, classificação ou segmentação, permitindo inferir conclusões sobre esse domínio.

2 Energy Use

2.1 Contextualização

Hoje em dia, vários dados são monitorizados de forma a automatizar, prevenir e ajustar diferentes parâmetros em várias situações reais. Existem cada vez mais casas a implementar sensores que conseguem monitorizar vários dados como a temperatura, humidade, luminosidade, entre outros. Desta forma, podem ser controladas informações como custos associados ou previsão de gastos.

Neste sentido, o primeiro conjunto de dados a ser interpretado, analisado e estudado, foi escolhido previamente pelos Docentes da Unidade Curricular. Os dados deste *dataset* são provenientes de várias fontes diferentes, analisando diversos fatores climatéricos e sendo estes registados periodicamente (10 em 10 minutos), durante 4 meses e meio.

Os primeiros dados registados são derivados das respostas dadas por uma rede de sensores denominada "ZigBee", onde cada nodo se encontra numa divisão diferente da casa e regista dois parâmetros: temperatura e humidade. Para além destes, são também reportados os valores de energia utilizada na casa.

Como segunda fonte de dados, é considerado um multisensor estacionado numa estação meteorológica em Chievres. Estes são incorporados nos dados anteriores, fazendo "match"através dos atributos data e tempo.

Por último, temos a adição de duas variáveis *random* que servem para testar modelos de regressão e filtrar atributos não predicativos.

É importante realçar que o *dataset* é constituído por 29 atributos e 19735 instâncias.

2.2 Objetivos

Depois de efetuada a leitura do artigo e analisado o*dataset*, foram estabelecidos os seguintes objetivos:

- Previsão do custo energético da habitação;
- Previsão da temperatura registada na estação Chiviers;
- Previsão da temperatura na área exterior da casa, na zona norte;
- Previsão do ponto de orvalho;

Apesar destes serem os objetivos iniciais e enumerados, outros parâmetros podem também ser analisados e estudados, inferindo as suas conclusões.

2.3 Descrição dos Atributos

Em seguida apresentamos a lista dos atributos acompanhados de uma breve descrição sobre eles, o seu tipo e valores.

Atributo	Descrição	Tipo	Valores
Date	Data e Tempo	Valores do tipo Data	Continuous
Appliances	Energia em Uso	Valor Numérico	Continuous
Lights	Energia Usada em luzes na casa	Valor Numerico	Continuous
T1	Temperatura na Área da Cozinha	Valor Numérico	Continuous
RH_1	Humidade na Área da Cozinha	Valor Numérico	Continuous
T2	Temperatura na Área da Sala de Estar	Valor Numérico	Continuous
RH_2	Humidade na Área da Sala de Estar	Valor Numérico	Continuous
T3	Temperatura na Área da Lavandaria	Valor Numérico	Continuous
RH_3	Humidade na Área da Lavandaria	Valor Numérico	Continuous
T4	Temperatura na Área do Escritório	Valor Numérico	Continuous
RH_4	Humidade na Área do Escritório	Valor Numérico	Continuous
T5	Temperatura na Área da Casa-de-Banho	Valor Numérico	Continuous
RH_5	Humidade na Área da Casa-de-Banho	Valor Numérico	Continuous
T6	Temperatura na Área Fora de Casa (Lado Norte)	Valor Numérico	Continuous
RH_6	Humidade na Área Fora de Casa (Lado Norte)	Valor Numérico	Continuous
T7	Temperatura na Área de "Ironing"	Valor Numérico	Continuous
RH_7	Humidade na Área de "Ironing"(Lado Norte)	Valor Numérico	Continuous
T8	Temperatura na Área do Quarto dos Filhos	Valor Numérico	Continuous
RH_8	Humidade na Área do Quarto dos Filhos	Valor Numérico	Continuous
T9	Temperatura na Área do Quarto dos Pais	Valor Numérico	Continuous
RH_9	Humidade na Área do Quarto dos Pais	Valor Numérico	Continuous
T_out	Temperatura no Exterior (Chievres weather station)	Valor Numérico	Continuous
$Press_mm$	Pressão(Chievres weather station)	Valor Numérico	Continuous
RH_out	Humidade no Exterior (Chievres weather station)	Valor Numérico	Continuous
Windspeed	Velocidade do Vento(Chievres weather station	Valor Numérico	Continuous
Visibility	Visibilidade(Chievres weather station)	Valor Numérico	Continuous
Tdewpoint	Tdewpoint(Chievres weather station)	Valor Numérico	Continuous
rv1	Variável Random	Valor Numérico	Continuous
rv2	Variável Random	Valor Numérico	Continuous

2.4 Preparação de Dados

Antes de iniciar qualquer tipo de análise, foi tomada a decisão de dividir o atributo *date*, devido à sensibilidade que este apresenta. Deste modo, dividiu-se o atributo em quatro novos atributos:

Atributo	Descrição	Tipo	Valores
Day	Dia da amostra	Valor Numérico	Continuous
Month	Mês da amostra	Valor Numérico	Continuous
Year	Ano da amostra	Valor Numérico	Continuous
Time	Horas da amostra	Valor Numérico	Continuous

Nota: O valor do atributo *Time* será composto por 6 algarismos sendo os primeiros dois são relativos à hora, os dois seguintes relativos aos minutos e os dois últimos referentes aos segundos. Com este ajuste, o *dataset* passou a ter um total de 32 atributos.

De seguida, foi verificado que não existe nenhuma ocorrência de valores em falta em qualquer um dos atributos. Como já referido na contextualização, existem várias fontes de dados, no entanto estas já se encontram no mesmo ficheiro de dados, não sendo assim necessário nenhum tipo de integração.

2.5 Análise Inicial

Após o tratamento de dados foi realizado o *upload* do *dataset* na ferramenta Weka. Inicialmente, cada um dos atributos foi analisado individualmente, verificando-se que todos se encontravam no tipo numérico.

Os quatro gráficos relativos aos atributos sobre a periodicidade das amostras encontram-se equilibrados, podendo assim concluir-se que as amostras foram retiradas de forma constante, validando a veracidade da descrição do *dataset*.

Os atributos relativos aos gastos de energia total e energia em luzes, apesar da sua uma gama alargada de valores, tendem a ter uma maior concentração de valores junto ao limite inferior do intervalo.

Nos atributos temperatura e humidade em cada uma das divisões, é possível verificar alguma correspondência, ou seja, existe uma volumetria idêntica de amostras com a temperatura baixa e a humidade alta em certas regiões da casa.

Relativamente aos dados recolhidos a partir da Estação de Chievres, é possível retirar a informação que a luminosidade é praticamente constante.

Os restantes aspetos vão variando os seus valores, mostrando uma maior concentração na numa gama intermédia do intervalo. Estas variações ocorrem devido às diferentes fases do dia, mas também por ser um período de 4,5 meses, abrangendo diferentes estações do ano. As duas variáveis *random* mantêm, ao longo da recolha, uma proporção constante de diferentes valores numa gama cingida entre os valores 0 e 50.

2.6 Relevância dos Atributos

Procedeu-se ao cálculo da relevância de cada um dos atributos do *dataset* em relação ao modelo de previsão. Para tal, foi usado o filtro de *attribute selection*, de maneira a chegar a resultados conclusivos. Para estes resultados serem suportados com uma veracidade sólida, foram corridos vários algoritmos de avaliação e de procura.

Algoritmos de Seleção

- CfsSubsetEval
- CorrelationAttributeEval
- GainRatioAttributeEval
- WrapperSubsetEval

Algoritmos de Procura

- BestFirst
- Ranker

Após vários testes realizados, chegou-se a uma tabela que faz a correspondência entre o atributo a prever e os atributos importantes na previsão do mesmo.

Atributo a prever	**Atributos Importantes**
Appliances	Time , Lights , RH_1 , T6 , RH_6 , T_out , Press_mm_hg , RH_out , WindSpeed
T_out	Day , T3 , T6 , Rh_out , WindSpeed , Tdewpoint
T6	T2, RH_6 , T_out
Tdewpoint	RH_2 , T6, RH_7 , T9, T_out

2.7 Classificação

2.7.1 Regressão

Um dos objetivos definidos pelo grupo para o estudo deste *dataset*, passa por descobrir funções que permitiram fazer uma previsão de certos valores. O custo em termos energéticos de uma casa é influenciado pelas as condições climatéricas interiores e exteriores da habitação, sendo assim um valor a prever consoante todos os outros atributos do *dataset*. Outros parâmetros como por exemplo características exteriores analisadas pela estação de Chievres, podem também ser previstos.

Deste modo, foi utilizada a metodologia de regressão linear. Este método consiste numa equação para se estimar a condicional (valor esperado) de uma variável y, dados os valores de algumas outras variáveis x.

```
Tdewpoint =

  0.00096278 * Day +
 -0.10510989 * Month +
  0.0000009  * Time +
  0.00006786 * Appliances +
 -0.00078339 * lights +
 -0.05517394 * T1 +
  0.02032387 * RH_1 +
  0.04013858 * T2 +
  0.063058   * RH_2 +
  0.00976781 * T3 +
 -0.03989211 * RH_3 +
  0.02928575 * T4 +
  0.02513284 * RH_4 +
 -0.11567713 * T5 +
  0.00176665 * RH_5 +
 -0.02769979 * T6 +
 -0.0125098  * RH_6 +
 -0.05969901 * T7 +
  0.02476544 * RH_7 +
  0.02835033 * T8 +
 -0.00738259 * RH_8 +
  0.13466336 * T9 +
 -0.00821831 * RH_9 +
  0.91480045 * T_out +
  0.00413361 * Press_mm_hg +
  0.20006735 * RH_out +
  0.04413636 * Windspeed +
 -0.0014369  * Visibility +
-24.42897129

Time taken to build model: 0.1 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correlation coefficient                  0.9943
Mean absolute error                      0.3106
Root mean squared error                  0.4471
Relative absolute error                  9.2644 %
Root relative squared error             10.6583 %
```

Figura 2: Função de previsão do atributo Tdewpoint

```
T_out =

  0.26245468 * Month +
  0.00000013 * Time +
 -0.00016623 * Appliances +
  0.15356069 * T1 +
 -0.02144836 * RH_1 +
 -0.123466   * T2 +
 -0.05351279 * RH_2 +
 -0.02038305 * T3 +
  0.03783677 * RH_3 +
 -0.00583596 * T4 +
 -0.00349593 * RH_4 +
  0.09939925 * T5 +
 -0.00277936 * RH_5 +
  0.19550493 * T6 +
  0.01680605 * RH_6 +
  0.06211114 * T7 +
 -0.0055546  * RH_7 +
 -0.01146844 * T8 +
  0.0217832  * RH_8 +
 -0.13990452 * T9 +
  0.01227059 * RH_9 +
 -0.18642664 * RH_out +
 -0.02485418 * Windspeed +
  0.00188011 * Visibility +
  0.83094283 * Tdewpoint +
 16.00562812

Time taken to build model: 0.16 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correlation coefficient                  0.9968
Mean absolute error                      0.2962
Root mean squared error                  0.426
Relative absolute error                  7.0705 %
Root relative squared error              8.0124 %
Total Number of Instances            19735
```

Figura 3: Função de previsão da temperatura na estação de Chievres

```
T6 =

  0.00396406 * Day +
 -0.32182724 * Month +
 -0.00000235 * Time +
  0.00086188 * Appliances +
  0.00583163 * lights +
 -0.71928645 * T1 +
 -0.02302804 * RH_1 +
  0.80980607 * T2 +
  0.10677615 * RH_2 +
  0.06937835 * T3 +
 -0.12881145 * RH_3 +
 -0.08964323 * T4 +
  0.03997851 * RH_4 +
 -0.19449289 * T5 +
 -0.03823163 * RH_6 +
 -0.17077703 * T7 +
  0.00762544 * RH_7 +
 -0.09530825 * T8 +
 -0.06582021 * RH_8 +
  0.33778213 * T9 +
  1.15375624 * T_out +
 -0.01826235 * Press_mm_hg +
  0.07480304 * RH_out +
 -0.0188699  * Windspeed +
 -0.14826658 * Tdewpoint +
 15.92007673

Time taken to build model: 0.21 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correlation coefficient                  0.9855
Mean absolute error                      0.7861
Root mean squared error                  1.0349
Relative absolute error                 16.5126 %
Root relative squared error             16.9934 %
Total Number of Instances           19735
```

Figura 4: Função de previsão a temperatura na área fora de casa

```
Linear Regression Model

Appliances =

-10.7714717  * Month +
  0.00010075 * Time +
  1.91278098 * lights +
 -3.12425198 * T1 +
 14.5355931  * RH_1 +
-17.97133829 * T2 +
-13.27022445 * RH_2 +
 26.68164379 * T3 +
  4.14498358 * RH_3 +
 -2.80775499 * T4 +
 -1.32972728 * RH_4 +
  6.88759579 * T6 +
  0.0657257  * RH_6 +
 -1.50721356 * RH_7 +
  7.7587445  * T8 +
 -3.81983199 * RH_8 +
 -9.71739099 * T9 +
 -7.77092456 * T_out +
 -0.36447279 * RH_out +
  1.35208921 * Windspeed +
  0.14040767 * Visibility +
  2.9402785  * Tdewpoint +
 92.61420743

Time taken to build model: 0.14 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correlation coefficient                  0.4108
Mean absolute error                     52.8375
Root mean squared error                 93.4705
Relative absolute error                 87.2956 %
Root relative squared error             91.1709 %
Total Number of Instances           19735
```

Figura 5: Função de previsão da Energia em Uso

Após os vários testes realizados conseguimos chegar a algumas conclusões interessantes. É possível verificar que o valor de erro da equação é influenciado consoante o número de atributos importantes em relação ao atributo que se está a prever o resultado. Por outro lado, observa-se por exemplo na equação da Temperatura na estação de Chievres que não existe o atributo dia, ou seja, conclui-se que o mês e o momento do dia influenciam o valor da temperatura, mas o número do dia não.

Na função de previsão da energia em uso é interessante notar o valor 26.68 na temperatura da lavandaria enquanto que como multiplicadores nas áreas comuns temos valores negativos. Isto justifica-se que enquanto que nas áreas comuns, se a temperatura for baixa, o seu multiplicador baixará pouco os custos de energia associados. Na lavandaria isto não acontece devido aos custos da maquinaria.

2.7.2 Árvores de Decisão

Como segundo algoritmo de treino, o grupo decidiu utilizar o J48. Este algoritmo consiste numa árvore em que os nodos interiores são os diferentes atributos, os ramos correspondem as decisões a tomar sobre esse mesmo atributo, e as folhas são as previsões do atributo que temos como objetivo descobrir o seu valor.

Atributo "Appliances"

Numa primeira instância utilizou-se o algoritmo sobre o atributo-alvo "Appliances", tendo alterado quer os parâmetros do algoritmo, quer os atributos utilizados na árvore de decisão. Recorreu-se a dois

conjuntos de dados: em primeiro lugar utilizou-se todo o *dataset*; em segundo lugar apenas os atributos importantes referenciados na secção 2.6.

Seguidamente, apresentam-se os dois outputs conjuntamente com uma interpretação dos dados.

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.08 seconds

=== Summary ===

Correctly Classified Instances       18583               94.1627 %
Incorrectly Classified Instances      1152                5.8373 %
Kappa statistic                          0.7985
Mean absolute error                      0.0187
Root mean squared error                  0.0968
Relative absolute error                 29.9361 %
Root relative squared error             54.7399 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,994    0,193    0,960      0,994    0,977      0,862    0,956     0,986     '(-inf-117]'
                 0,779    0,012    0,861      0,779    0,818      0,803    0,962     0,868     '(117-224]'
                 0,763    0,007    0,825      0,763    0,793      0,784    0,974     0,843     '(224-331]'
                 0,523    0,003    0,809      0,523    0,635      0,644    0,962     0,709     '(331-438]'
                 0,519    0,002    0,753      0,519    0,615      0,622    0,985     0,709     '(438-545]'
                 0,446    0,001    0,763      0,446    0,563      0,582    0,985     0,666     '(545-652]'
                 0,406    0,000    0,778      0,406    0,533      0,561    0,987     0,649     '(652-759]'
                 0,125    0,000    0,750      0,125    0,214      0,306    0,993     0,463     '(759-866]'
                 0,200    0,000    0,500      0,200    0,286      0,316    1,000     0,340     '(866-973]'
                 0,000    0,000    0,000      0,000    0,000      0,000    1,000     0,450     '(973-inf)'
Weighted Avg.    0,942    0,160    0,937      0,942    0,937      0,842    0,958     0,956

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j   <-- classified as
 16143    65    20     7     1     1     1     0     0     0 |     a = '(-inf-117]'
   343  1342    24    11     2     0     1     0     0     0 |     b = '(117-224]'
   132    51   640     9     4     2     1     0     0     0 |     c = '(224-331]'
   126    48    48   258     8     2     2     1     0     0 |     d = '(331-438]'
    32    27    22    19   110     2     0     0     0     0 |     e = '(438-545]'
    26    15    12     7    12    58     0     0     0     0 |     f = '(545-652]'
    10     5     7     7     4     8    28     0     0     0 |     g = '(652-759]'
     4     4     2     1     5     2     3     3     0     0 |     h = '(759-866]'
     2     1     1     0     0     0     0     0     1     0 |     i = '(866-973]'
     0     0     0     0     0     1     0     0     1     0 |     j = '(973-inf)'
```

Figura 6: Resultados do algoritmo J48 com todo o *dataset* como input

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.07 seconds

=== Summary ===

Correctly Classified Instances       18379               93.129  %
Incorrectly Classified Instances      1356                6.871  %
Kappa statistic                          0.7608
Mean absolute error                      0.0216
Root mean squared error                  0.1039
Relative absolute error                 34.4989 %
Root relative squared error             58.7637 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,991    0,225    0,953      0,991    0,972      0,832    0,964     0,989     '(-inf-117]'
               0,714    0,014    0,832      0,714    0,769      0,751    0,962     0,831     '(117-224]'
               0,727    0,008    0,794      0,727    0,759      0,750    0,978     0,820     '(224-331]'
               0,507    0,004    0,762      0,507    0,609      0,614    0,965     0,674     '(331-438]'
               0,514    0,002    0,703      0,514    0,594      0,598    0,987     0,665     '(438-545]'
               0,431    0,001    0,659      0,431    0,521      0,530    0,992     0,620     '(545-652]'
               0,261    0,000    0,750      0,261    0,387      0,441    0,996     0,567     '(652-759]'
               0,250    0,000    0,750      0,250    0,375      0,433    0,999     0,550     '(759-866]'
               0,400    0,000    0,667      0,400    0,500      0,516    1,000     0,517     '(866-973]'
               0,000    0,000    0,000      0,000    0,000      0,000    1,000     0,333     '(973-inf)'
Weighted Avg.  0,931    0,187    0,925      0,931    0,926      0,810    0,965     0,952

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j   <-- classified as
 16098    98    25    12     1     3     1     0     0     0 |     a = '(-inf-117]'
   439  1230    34    12     6     2     0     0     0     0 |     b = '(117-224]'
   144    58   610    15     9     2     1     0     0     0 |     c = '(224-331]'
   133    42    54   250     8     5     1     0     0     0 |     d = '(331-438]'
    36    25    22    15   109     3     1     1     0     0 |     e = '(438-545]'
    22    13    14    14    10    56     1     0     0     0 |     f = '(545-652]'
    12     7     7     4     7    12    18     1     1     0 |     g = '(652-759]'
     1     4     2     4     5     1     1     6     0     0 |     h = '(759-866]'
     1     1     0     1     0     0     0     0     2     0 |     i = '(866-973]'
     0     0     0     1     0     1     0     0     0     0 |     j = '(973-inf)'
```

Figura 7: Resultados do algoritmo J48 com top 9 atributos como input

Através deste método conseguimos tirar novas conclusões relativamente ao atributo Appliances , em primeiro lugar consegue-se ver apenas a descida em 1% na percentagem de instâncias classificadas corretamente, com isto podemos concluir que a avaliação dos atributos importantes para previsão de valores do atributo Appliances está acertada. Já o erro encontra-se nos 34%.

Atributo "Tdewpoint"

De seguida procedeu-se ao mesmo método para a previsão de valores do atributo Tdewpoint. Para dificultar a tarefa, decidiu-se alterar o número de intervalos na discretização do atributo alvo, passando assim para 20 segmentos em vez de 10.

Devido à dimensão do output, optou-se por apenas resumir os resultados. Os valores de instâncias corretamente classificadas encontram-se, mais uma vez, acima dos 90%, diferenciando-se muito pouco entre ter o *dataset* total e apenas os atributos relevantes. Neste atributo, o grupo testou também o inverso, isto é, manter um conjunto de atributos irrelevantes para a previsão do "Tdewpoint", concluindo-se que a percentagem de casos corretos apenas se ficou pelos 30%.

Atributo T_out

Em terceiro lugar procedeu-se ao estudo, com o mesmo método, mas desta vez com o foco no atributo da temperatura exterior na estação de Chievres. Para tal, procedeu-se à passagem do atributo de valor numérico para valor nominal, criando 20 intervalos.

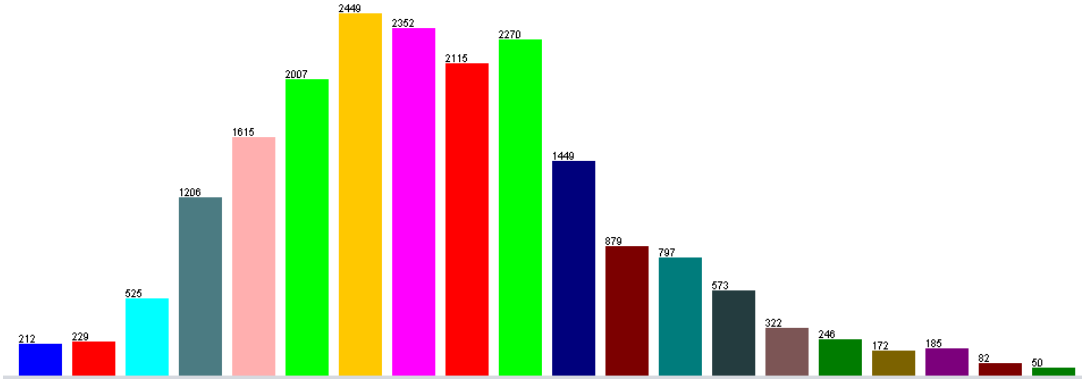

Figura 8: Atributo T_out no tipo nominal

Conseguiu-se obter novamente bons resultados com intervalos entre 90%-100% de casos de sucesso, como se pode observar na matriz de output calculada. Os valores que se encontram na diagonal da matriz são as instâncias corretas e, por sua vez, os que se encontram noutras células que não a diagonal são valores mal calculados. É possível verificar que mesmo os mal calculados não se desviam mais do que uma célula do seu correto local.

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k    l    m    n    o    p    q    r    s    t   <-- classified as
 211    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 |    a = '(-inf--3.445]'
   2  223    4    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 |    b = '(-3.445--1.89]'
   0    0  521    4    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 |    c = '(-1.89--0.335]'
   0    0    5 1195    6    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0 |    d = '(-0.335-1.22]'
   0    0    1    7 1591   16    0    0    0    0    0    0    0    0    0    0    0    0    0    0 |    e = '(1.22-2.775]'
   0    0    0    1   12 1972   22    0    0    0    0    0    0    0    0    0    0    0    0    0 |    f = '(2.775-4.33]'
   0    0    0    0    0   12 2413   21    3    0    0    0    0    0    0    0    0    0    0    0 |    g = '(4.33-5.885]'
   0    0    0    0    0    0   30 2294   26    2    0    0    0    0    0    0    0    0    0    0 |    h = '(5.885-7.44]'
   0    0    0    0    0    0    1   22 2068   24    0    0    0    0    0    0    0    0    0    0 |    i = '(7.44-8.995]'
   0    0    0    0    0    0    0    0   24 2232   14    0    0    0    0    0    0    0    0    0 |    j = '(8.995-10.55]'
   0    0    0    0    0    0    0    0    3   19 1420    7    0    0    0    0    0    0    0    0 |    k = '(10.55-12.105]'
   0    0    0    0    0    0    0    0    0    2    9  865    2    1    0    0    0    0    0    0 |    l = '(12.105-13.66]'
   0    0    0    0    0    0    0    0    0    1    1   19  760   14    0    1    1    0    0    0 |    m = '(13.66-15.215]'
   0    0    0    0    0    0    0    0    0    0    0    2    4  560    7    0    0    0    0    0 |    n = '(15.215-16.77]'
   0    0    0    0    0    0    0    0    0    0    0    0    0    7  311    3    1    0    0    0 |    o = '(16.77-18.325]'
   0    0    0    0    0    0    0    0    0    0    0    0    1    0    4  240    1    0    0    0 |    p = '(18.325-19.88]'
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0  171    1    0    0 |    q = '(19.88-21.435]'
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    2  181    1    0 |    r = '(21.435-22.99]'
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2   80    0 |    s = '(22.99-24.545]'
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2   48 |    t = '(24.545-inf)'
```

Figura 9: Confunsion Matrix - T_out

Atributo T_6

Como último atributo a ser previsto, considerou-se a temperatura na parte exterior, na área da parte norte da casa. Como atributos de relevância para a previsão deste temos a humidade na mesma área, a temperatura registada na estação, e a humidade na área da sala de estar. Nesta previsão foi conseguido um erro absoluto de apenas 4%, mantendo a elevada percentagem de casos de sucesso (superior a 96%).

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,992	0,000	1,000	0,992	0,996	0,996	1,000	1,000	'(-inf--4.352]'
	1,000	0,000	0,975	1,000	0,987	0,987	1,000	0,999	'(-4.352--2.634]'
	0,977	0,000	0,991	0,977	0,984	0,984	1,000	0,997	'(-2.634--0.916]'
	0,983	0,001	0,976	0,983	0,979	0,978	1,000	0,996	'(-0.916-0.802]'
	0,977	0,002	0,981	0,977	0,979	0,978	1,000	0,996	'(0.802-2.52]'
	0,977	0,003	0,977	0,977	0,977	0,974	0,999	0,995	'(2.52-4.238]'
	0,977	0,004	0,972	0,977	0,974	0,971	0,999	0,994	'(4.238-5.956]'
	0,975	0,004	0,975	0,975	0,975	0,972	0,999	0,994	'(5.956-7.674]'
	0,973	0,003	0,974	0,973	0,973	0,970	0,999	0,993	'(7.674-9.392]'
	0,978	0,003	0,974	0,978	0,976	0,973	0,999	0,994	'(9.392-11.11]'
	0,960	0,002	0,978	0,960	0,969	0,967	0,999	0,990	'(11.11-12.828]'
	0,977	0,003	0,954	0,977	0,965	0,963	0,999	0,985	'(12.828-14.546]'
	0,954	0,001	0,976	0,954	0,965	0,963	1,000	0,990	'(14.546-16.264]'
	0,965	0,001	0,960	0,965	0,963	0,962	1,000	0,989	'(16.264-17.982]'
	0,954	0,001	0,965	0,954	0,959	0,959	1,000	0,988	'(17.982-19.7]'
	0,973	0,000	0,973	0,973	0,973	0,973	1,000	0,995	'(19.7-21.418]'
	0,942	0,000	0,942	0,942	0,942	0,942	1,000	0,971	'(21.418-23.136]'
	0,976	0,000	0,969	0,976	0,972	0,972	1,000	0,992	'(23.136-24.854]'
	0,986	0,000	0,993	0,986	0,990	0,990	1,000	0,998	'(24.854-26.572]'
	0,987	0,000	1,000	0,987	0,994	0,994	1,000	0,997	'(26.572-inf)'
Weighted Avg.	0,974	0,002	0,974	0,974	0,974	0,972	0,999	0,993	

Figura 10: Detalhes por classe - T_6

Como podemos ver na tabela apresentada em cima, os valores da coluna TP Rate encontram-se altos (perto de 1), enquanto que na coluna FP Rate encontram-se nulos ou aproximadamente 0. Estas duas colunas significam o rácio de sucesso e insucesso respetivamente.

A coluna de precisão contém todos os valores acima de 0.95, o que significa que a proporção do número de instâncias que pertencem mesmo a esta classe, face ao número de instâncias previstos, é alto.

2.8 Associação

Com o objetivo de encontrar padrões frequentes e possíveis associações entre atributos, utilizou-se o algoritmo **Apriori** que realiza a pesquisa de regras de associação baseadas em aproximações booleanas. Por outro lado, utilizou-se também o **FilteredAssociator** que apesar de realizar menos ciclos e gerar menores *itemsets*, é uma forma de corroborar os resultados apresentados pelo primeiro algoritmo.

Foram realizados diferentes testes, alterando os parâmetros passados como argumentos.

```
Best rules found:

 1. Time='(-inf-23500]' 2192 ==> Appliances='(-inf-117]' 2184    <conf:(1)> lift:(1.21) lev:(0.02) [380] conv:(43.16)
 2. Time='(211500-inf)' 2192 ==> Appliances='(-inf-117]' 2074    <conf:(0.95)> lift:(1.15) lev:(0.01) [270] conv:(3.26)
 3. lights='(-inf-7]' T6='(0.802-4.238]' 2864 ==> Appliances='(-inf-117]' 2662    <conf:(0.93)> lift:(1.13) lev:(0.02) [305] conv:(2.5)
 4. lights='(-inf-7]' RH_out='(92.4-inf)' 3291 ==> Appliances='(-inf-117]' 3058    <conf:(0.93)> lift:(1.13) lev:(0.02) [350] conv:(2.49)
 5. lights='(-inf-7]' T2='(17.476-18.852]' 3122 ==> Appliances='(-inf-117]' 2900    <conf:(0.93)> lift:(1.13) lev:(0.02) [331] conv:(2.48)
 6. lights='(-inf-7]' T6='(0.802-4.238]' T_out='(1.22-4.33]' 2154 ==> Appliances='(-inf-117]' 1987    <conf:(0.92)> lift:(1.12) lev:(0.01) [214] conv:(2.27)
 7. lights='(-inf-7]' RH_1='(37.922-41.556]' RH_4='(37.032-39.375]' 2655 ==> Appliances='(-inf-117]' 2431    <conf:(0.92)> lift:(1.11) lev:(0.01) [246] conv:(2.09)
 8. lights='(-inf-7]' RH_out='(84.8-92.4]' 3862 ==> Appliances='(-inf-117]' 3532    <conf:(0.91)> lift:(1.11) lev:(0.02) [354] conv:(2.07)
 9. lights='(-inf-7]' T3='(19.608-20.812]' 2983 ==> Appliances='(-inf-117]' 2725    <conf:(0.91)> lift:(1.11) lev:(0.01) [270] conv:(2.04)
10. RH_6='(50.45-60.34]' 2456 ==> Appliances='(-inf-117]' 2236    <conf:(0.91)> lift:(1.11) lev:(0.01) [215] conv:(1.97)
```

Figura 11: Resultados do Algoritmo Apriori com confiança maior ou igual a 90% e suporte mínimo de 0.1

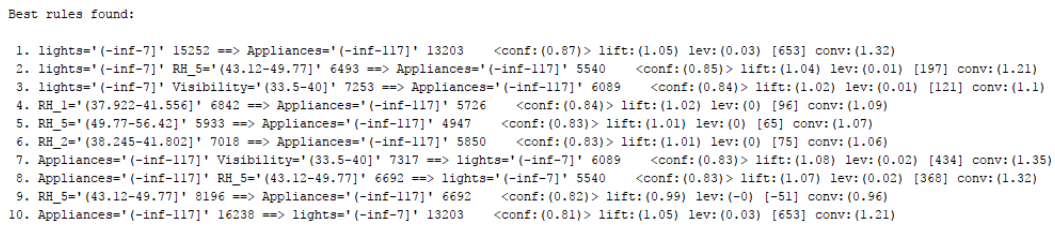

Figura 12: Resultados do Algoritmo Apriori com confiança maior ou igual a 60% e suporte mínimo de 0.2

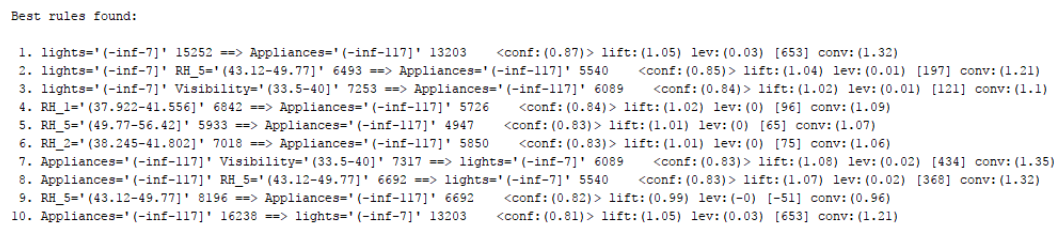

Figura 13: Resultados do Algoritmo Apriori com confiança maior ou igual a 70% e suporte mínimo de 0.15

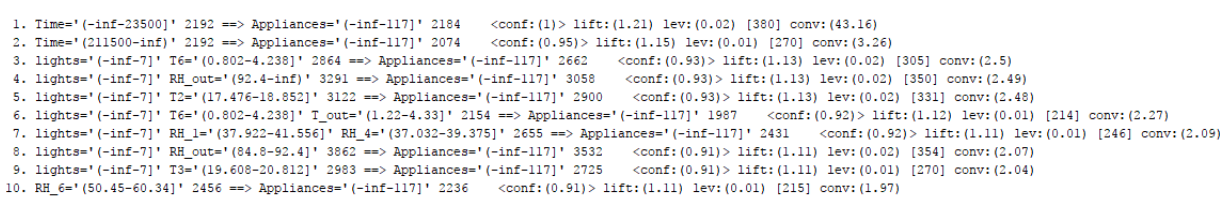

Figura 14: Resultados do Algoritmo FilteredAssociator

Após estes primeiros resultados, concluiu-se que a maior parte das regras descritas associavam o atributo "appliances"ou com o atributo "lights"ou com o "time". Apesar das boas associações, é de notar que o intervalo referente ao "appliances"é sempre o [inf-117] devido a ser o intervalo com maior numero de instâncias (cerca de 80%). Para contrariar estes resultados, o grupo alterou o filtro de discretização alterando de ***equal-width binning*** para ***equal-frequency binning***, obtendo assim uma melhor distribuição das instâncias do *dataset*.

```
Best rules found:

 1. RH_6='(-inf-5.28]' T_out='(14.575-inf)' 1098 ==> T6='(16.095-inf)' 1069    <conf:(0.97)> lift:(9.78) lev:(0.05) [959] conv:(32.96)
 2. Time='(21500-43500]' 1918 ==> lights='(-inf-5]' 1866    <conf:(0.97)> lift:(1.26) lev:(0.02) [383] conv:(8.22)
 3. Appliances='(-inf-35]' 1075 ==> lights='(-inf-5]' 992    <conf:(0.92)> lift:(1.19) lev:(0.01) [161] conv:(2.91)
 4. Time='(43500-70500]' 2055 ==> lights='(-inf-5]' 1885    <conf:(0.92)> lift:(1.19) lev:(0.02) [296] conv:(2.73)
 5. Appliances='(45-55]' 4368 ==> lights='(-inf-5]' 3995    <conf:(0.91)> lift:(1.18) lev:(0.03) [619] conv:(2.65)
 6. RH_6='(-inf-5.28]' T_out='(14.575-inf)' 1098 ==> lights='(-inf-5]' 1000    <conf:(0.91)> lift:(1.18) lev:(0.01) [151] conv:(2.52)
 7. T6='(16.095-inf)' RH_6='(-inf-5.28]' 1261 ==> lights='(-inf-5]' 1148    <conf:(0.91)> lift:(1.18) lev:(0.01) [173] conv:(2.51)
 8. Appliances='(35-45]' 2019 ==> lights='(-inf-5]' 1814    <conf:(0.9)> lift:(1.16) lev:(0.01) [253] conv:(2.23)
 9. T6='(16.095-inf)' T_out='(14.575-inf)' 1621 ==> lights='(-inf-5]' 1451    <conf:(0.9)> lift:(1.16) lev:(0.01) [198] conv:(2.15)
10. T_out='(14.575-inf)' 1967 ==> lights='(-inf-5]' 1745    <conf:(0.89)> lift:(1.15) lev:(0.01) [224] conv:(2)
11. RH_6='(-inf-5.28]' 1973 ==> lights='(-inf-5]' 1747    <conf:(0.89)> lift:(1.15) lev:(0.01) [222] conv:(1.97)
12. Time='(-inf-21500]' 1918 ==> lights='(-inf-5]' 1693    <conf:(0.88)> lift:(1.14) lev:(0.01) [210] conv:(1.93)
13. T6='(16.095-inf)' 1965 ==> lights='(-inf-5]' 1730    <conf:(0.88)> lift:(1.14) lev:(0.01) [211] conv:(1.89)
14. RH_6='(-inf-5.28]' RH_out='(-inf-57.415]' 1356 ==> lights='(-inf-5]' 1193    <conf:(0.88)> lift:(1.14) lev:(0.01) [145] conv:(1.88)
15. RH_out='(-inf-57.415]' 1978 ==> lights='(-inf-5]' 1692    <conf:(0.86)> lift:(1.11) lev:(0.01) [163] conv:(1.57)
16. T6='(16.095-inf)' RH_6='(-inf-5.28]' 1261 ==> T_out='(14.575-inf)' 1069    <conf:(0.85)> lift:(8.51) lev:(0.05) [943] conv:(5.88)
17. T_out='(-inf-0.975]' 1974 ==> T6='(-inf-0.84]' 1673    <conf:(0.85)> lift:(8.46) lev:(0.07) [1475] conv:(5.88)
18. lights='(-inf-5]' T6='(-inf-0.84]' 1481 ==> T_out='(-inf-0.975]' 1254    <conf:(0.85)> lift:(8.47) lev:(0.06) [1105] conv:(5.85)
19. T6='(-inf-0.84]' 1978 ==> T_out='(-inf-0.975]' 1673    <conf:(0.85)> lift:(8.46) lev:(0.07) [1475] conv:(5.82)
20. lights='(-inf-5]' T_out='(-inf-0.975]' 1495 ==> T6='(-inf-0.84]' 1254    <conf:(0.84)> lift:(8.37) lev:(0.06) [1104] conv:(5.56)
21. lights='(-inf-5]' T6='(16.095-inf)' 1730 ==> T_out='(14.575-inf)' 1451    <conf:(0.84)> lift:(8.41) lev:(0.06) [1278] conv:(5.56)
22. RH_6='(48.18-55.28]' 1971 ==> lights='(-inf-5]' 1643    <conf:(0.83)> lift:(1.08) lev:(0.01) [119] conv:(1.36)
23. lights='(-inf-5]' T_out='(14.575-inf)' 1745 ==> T6='(16.095-inf)' 1451    <conf:(0.83)> lift:(8.35) lev:(0.06) [1277] conv:(5.33)
24. Press_mm_hg='(749.625-751.79]' 1953 ==> lights='(-inf-5]' 1622    <conf:(0.83)> lift:(1.07) lev:(0.01) [112] conv:(1.34)
25. RH_1='(35.475-36.895]' 1964 ==> lights='(-inf-5]' 1630    <conf:(0.83)> lift:(1.07) lev:(0.01) [112] conv:(1.33)
26. Time='(141500-164500]' 2055 ==> lights='(-inf-5]' 1701    <conf:(0.83)> lift:(1.07) lev:(0.01) [112] conv:(1.31)
27. RH_1='(38.785-39.635]' 1995 ==> lights='(-inf-5]' 1651    <conf:(0.83)> lift:(1.07) lev:(0.01) [109] conv:(1.31)
28. RH_1='(-inf-35.475]' 1976 ==> lights='(-inf-5]' 1631    <conf:(0.83)> lift:(1.07) lev:(0.01) [103] conv:(1.3)
29. T6='(16.095-inf)' 1965 ==> T_out='(14.575-inf)' 1621    <conf:(0.82)> lift:(8.28) lev:(0.07) [1425] conv:(5.13)
30. T_out='(14.575-inf)' 1967 ==> T6='(16.095-inf)' 1621    <conf:(0.82)> lift:(8.28) lev:(0.07) [1425] conv:(5.1)
```

Figura 15: Resultados com equal-frequency binning

2.9 Análise de Resultados

Com a análise deste *dataset* conseguimos tirar uma variedade de conclusões interessantes. Deste modo, iremos descrever nesta secção as conclusões a que se chegou, relativamente a cada tópico descrito na secção *Objetivos*.

Relativamente ao atributo "Appliance", ou seja, a energia utilizada, conseguimos perceber que é o parâmetro mais volátil a alterações por consequência dos outros, tornando assim mais difícil arranjar uma função de regressão com um erro mínimo. Para além disso, notou-se a sua forte dependência relativamente às horas do dia e ao uso energético em luzes da casa.

A temperatura exterior tem associada a ela os outros fatores climatéricos exteriores, tendo sido possível intepretar associações entre eles e fazer modelos de previsão com uma percentagem de quase 100% na sua fiabilidade. De notar também a correspondência direta entre a temperatura exterior na estação e a temperatura exterior na área da zona Norte da casa.

Quanto ao atributo T6, ou seja, a temperatura exterior na área da zona Norte da casa, para além da associação inversa já falada no parágrafo em cima, também se conseguiu estabelecer associações com a humidade e temperatura da área da sala de estar da casa. O ponto de orvalho está associado aos outros fatores climatéricos registados pela estação, mas também aos níveis de humidade registado em algumas áreas da casa.

Para além destas previsões, foi também flagrante a associação entre as horas dos dias e os custos de energia, tanto gerais como relativamente às luzes. No período entre as [21.50h - 7.50h], os níveis de energia em uso estavam sempre no limite mínimo do intervalo de energia utilizada do total da amostra. Por outro lado, quando a temperatura aumenta na área exterior a casa e a humidade diminui, ou seja, pressupõe-se que seja já no horário diurno, o custo energético em luzes diminui. Para finalizar, a análise de resultados de associações diretas também são verificadas, no entanto o grupo assumiu que eram de pouca relevância, por exemplo entre a humidade e a temperatura numa área interior da casa.

3 Iris

No sentido de dar continuidade ao estudo das técnicas de extração de conhecimento lecionadas, o grupo escolheu um segundo conjunto de dados, denominado de "Iris". Este *dataset* consiste num conjunto de dados multivariados, introduzido pelo estatístico e biólogo britânico Ronald Ficher no seu artigo "*The use of multiple measurements in taxonomic problems* (1936)", contendo um total de 4 atributos e 150 instâncias, sem valores em falta.

3.1 Objetivo

Com o estudo deste *dataset* pretende-se inferir a relação existente entre as características das plantas e a sua respetiva classe.

Assim, encontrando os vários padrões existentes, deve ser possível identificar a classe da Iris a partir dos dados das suas caraterísticas. Existem três classes de plantas, cada uma contendo um total de 50 instâncias:

- *Iris Setosa*
- *Iris Versicolor*
- *Iris Virginica*

3.2 Descrição dos Atributos

Em seguida apresentamos a lista dos atributos, bem como uma breve descrição de cada um deles:

Atributo	**Descrição**	**Tipo**	**Valores**
Sepal Length	Comprimento da Sépala	Valor Numérico	Continuous
Sepal Width	Largura da Sépala	Valor Numérico	Continuous
Petal Length	Comprimento da Pétala	Valor Numérico	Continuous
Petal Width	Largura da Pétala	Valor Numérico	Continuous

Tabela 1: Iris *Dataset* - Descrição dos Atributos

3.3 Preparação de Dados

Com o objetivo de preparar corretamente os dados e assim obter uma extração de conhecimento mais eficaz, procedeu-se ao uso de algumas técnicas lecionadas na unidade curricular.

Em primeiro lugar, verificou-se que não existiam atributos redundantes, logo não sendo necessário reduzir o seu valor. Em seguida, constatou-se não seria preciso realizar a integração de dados uma vez que existe apenas uma fonte de dados. No mesmo sentido, não foi necessário fazer limpeza dos dados pois estes já se encontravam devidamente tratados.

Por outro lado, ao analisar cada atributo, verificou-se a existência de um número elevado de valores distintos para cada uma das características das plantas, como podemos verificar por exemplo para os valores de *Sepal Length*:

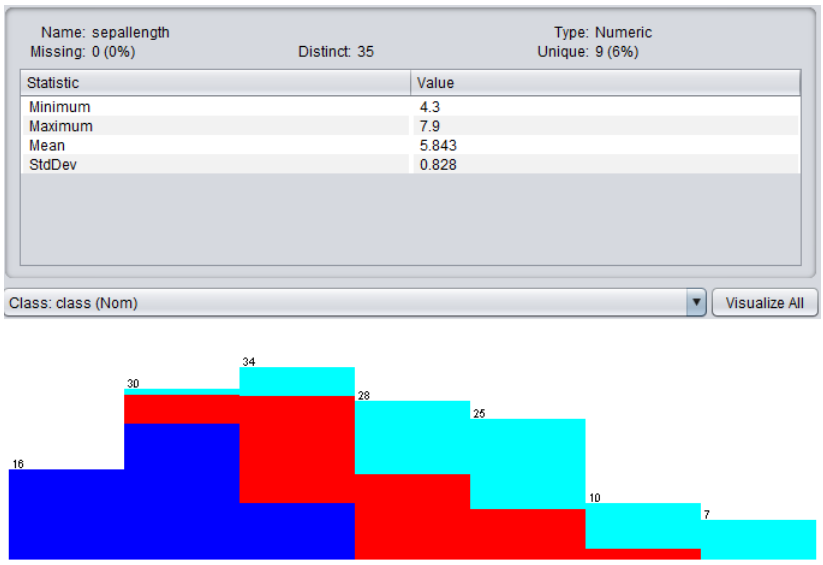

Figura 16: Iris - Análise inicial ao atributo *Sepal Length*

Deste modo, para se poder utilizar os algoritmos de associação, decidiu-se realizar um processo de discretização dos dados, reduzindo o número de valores para estes atributos. Assim, estes dados passaram a ser do tipo nominal e a estar agrupados em dez intervalos. Em seguida podemos verificar a distribuição de valores para o atributo *Sepal Length* após este processo.

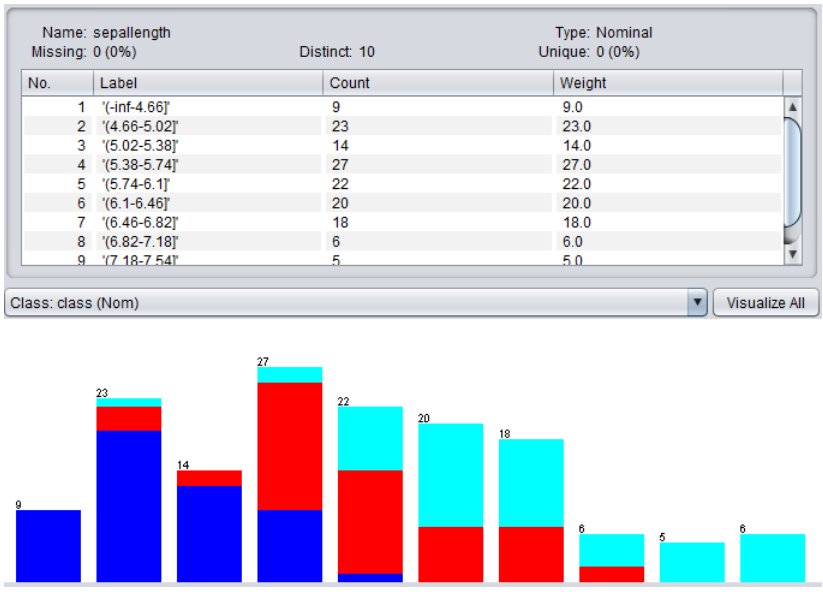

Figura 17: Iris - Análise ao atributo *Sepal Length* após discretização

3.4 Data Mining

3.4.1 Associação

Com o objetivo de encontrar padrões frequentes e possíveis associações entre atributos, utilizou-se o algoritmo **Apriori** que realiza a pesquisa de regras de associação baseadas em aproximações booleanas.

```
Best rules found:

 1. petalwidth='(-inf-0.34]' 41 ==> class=Iris-setosa 41    <conf:(1)> lift:(3) lev:(0.18) [27] conv:(27.33)
 2. petallength='(-inf-1.59]' 37 ==> class=Iris-setosa 37    <conf:(1)> lift:(3) lev:(0.16) [24] conv:(24.67)
 3. petallength='(-inf-1.59]' petalwidth='(-inf-0.34]' 33 ==> class=Iris-setosa 33    <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22)
 4. petalwidth='(1.06-1.3]' 21 ==> class=Iris-versicolor 21    <conf:(1)> lift:(3) lev:(0.09) [14] conv:(14)
 5. petallength='(5.13-5.72]' 18 ==> class=Iris-virginica 18    <conf:(1)> lift:(3) lev:(0.08) [12] conv:(12)
 6. sepallength='(4.66-5.02]' petalwidth='(-inf-0.34]' 17 ==> class=Iris-setosa 17    <conf:(1)> lift:(3) lev:(0.08) [11] conv:(11.33)
 7. sepalwidth='(2.96-3.2]' class=Iris-setosa 16 ==> petalwidth='(-inf-0.34]' 16    <conf:(1)> lift:(3.66) lev:(0.08) [11] conv:(11.63)
 8. sepalwidth='(2.96-3.2]' petalwidth='(-inf-0.34]' 16 ==> class=Iris-setosa 16    <conf:(1)> lift:(3) lev:(0.07) [10] conv:(10.67)
 9. petallength='(3.95-4.54]' 26 ==> class=Iris-versicolor 25    <conf:(0.96)> lift:(2.88) lev:(0.11) [16] conv:(8.67)
10. petalwidth='(1.78-2.02]' 23 ==> class=Iris-virginica 22    <conf:(0.96)> lift:(2.87) lev:(0.1) [14] conv:(7.67)
```

Figura 18: Iris - Resultados do Algoritmo Apriori com confiança maior ou igual a 90% e suporte mínimo de 0.1

```
Best rules found:

 1. petalwidth='(-inf-0.34]' 41 ==> class=Iris-setosa 41    <conf:(1)> lift:(3) lev:(0.18) [27] conv:(27.33)
 2. petallength='(-inf-1.59]' 37 ==> class=Iris-setosa 37    <conf:(1)> lift:(3) lev:(0.16) [24] conv:(24.67)
 3. petallength='(-inf-1.59]' petalwidth='(-inf-0.34]' 33 ==> class=Iris-setosa 33    <conf:(1)> lift:(3) lev:(0.15) [22] conv:(22)
 4. petallength='(3.95-4.54]' 26 ==> class=Iris-versicolor 25    <conf:(0.96)> lift:(2.88) lev:(0.11) [16] conv:(8.67)
 5. petallength='(-inf-1.59]' 37 ==> petalwidth='(-inf-0.34]' 33    <conf:(0.89)> lift:(3.26) lev:(0.15) [22] conv:(5.38)
 6. petallength='(-inf-1.59]' class=Iris-setosa 37 ==> petalwidth='(-inf-0.34]' 33    <conf:(0.89)> lift:(3.26) lev:(0.15) [22] conv:(5.38)
 7. petallength='(-inf-1.59]' 37 ==> petalwidth='(-inf-0.34]' class=Iris-setosa 33    <conf:(0.89)> lift:(3.26) lev:(0.15) [22] conv:(5.38)
 8. class=Iris-setosa 50 ==> petalwidth='(-inf-0.34]' 41    <conf:(0.82)> lift:(3) lev:(0.18) [27] conv:(3.63)
 9. petalwidth='(-inf-0.34]' 41 ==> petallength='(-inf-1.59]' 33    <conf:(0.8)> lift:(3.26) lev:(0.15) [22] conv:(3.43)
10. petalwidth='(-inf-0.34]' class=Iris-setosa 41 ==> petallength='(-inf-1.59]' 33    <conf:(0.8)> lift:(3.26) lev:(0.15) [22] conv:(3.43)
11. petalwidth='(-inf-0.34]' 41 ==> petallength='(-inf-1.59]' class=Iris-setosa 33    <conf:(0.8)> lift:(3.26) lev:(0.15) [22] conv:(3.43)
```

Figura 19: Iris - Resultados do Algoritmo Apriori com confiança maior ou igual a 80% e suporte mínimo de 0.15

3.4.2 Classificação

Como o objetivo do estudo deste *dataset* é a previsão da classe da uma planta consoante as características desta, devem-se utilizar algoritmos de classificação. Neste caso, o grupo optou por utilizar o algoritmo J48. É importante salientar que este algoritmo apresenta a capacidade de lidar com valores numéricos, logo, foi utilizada a versão do *dataset* sem discretização de atributos.

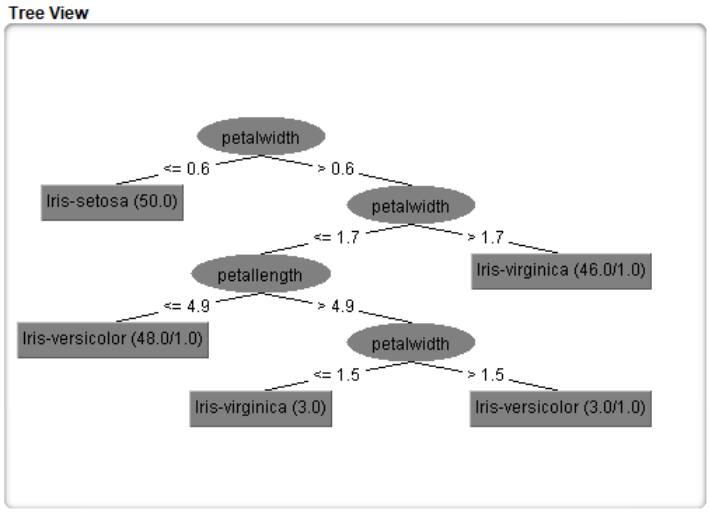

Figura 20: Iris - Árvore do Algoritmo J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.000    1.000      0.980    0.990      0.985    0.990     0.987     Iris-setosa
                 0.940    0.030    0.940      0.940    0.940      0.910    0.952     0.880     Iris-versicolor
                 0.960    0.030    0.941      0.960    0.950      0.925    0.961     0.905     Iris-virginica
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.968     0.924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

Figura 21: Iris - Resultados do Algoritmo J48

3.4.3 Segmentação

Para efeitos de teste, o grupo optou também por realizar um teste em que é usada a segmentação. Neste sentido, uma visto que a segmentação é um processo não supervisionado, optou-se por esconder o atributo objetivo (classe) para verificar se realmente cada espécie forma um grupo natural. Deste modo, foram consideradas a existências de três *clusters*.

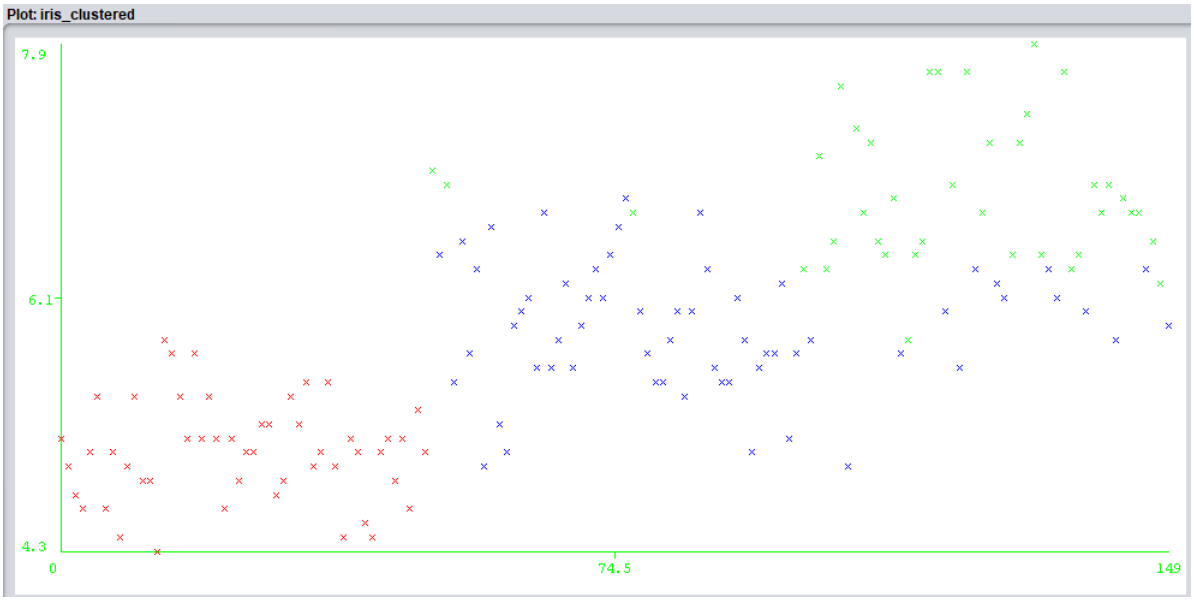

Figura 22: Iris - *Cluster*

```
Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                          Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
============================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795



Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

Figura 23: Iris - Resultados *Cluster*

3.5 Análise de Resultados

Utilizados os diversos algoritmos lecionados, podemos então concluir para este *dataset*, os mais adequados seriam os algoritmos de associação e classificação, uma vez que o objetivo principal é prever a classe, e sabemos previamente da existência de três classes.

Deste modo, podemos verificar que foram encontradas várias conclusões 100% corretas, através do algoritmo de associação utilizado. No mesmo sentido, é possível observar que o algoritmo J48 consegue 144 das 150 instâncias corretas. Por outro lado, é possível verificar que a classe *Setosa* forma um grupo natural muito destacado dos outros, sendo por isso fácil de calcular/prever. No entanto, apresenta falhas entre as classes *Versicolor* e *Virginica*, uma vez que existe sobreposição entre os dois, o que dificulta as classificações pertencentes a essa zona (daí os 39/61 e não os 50/50 pretendidos).

Assim, foram retiradas algumas conclusões importantes:

- Se o valor de Petal Width for inferior a 0.6, então a sua classe é *Setosa*;
- Se o valor de Petal Length for inferior a 1,59, então a sua classe é *Setosa*;
- Se o valor de Sepal Length estiver no intervalo [4.66-5.02] e Petal Length inferior a 0.34, a sua classe é *Setosa*;
- Se o valor de Sepal Width estiver no intervalo [2.96-3.2] e Petal Length inferior a 0.34, a sua classe é *Setosa*;
- Se o valor de Petal Width estiver no intervalo [1.06-1.3], a sua classe é *Versicolor*;
- Se o valor de Petal Length estiver no intervalo [5.13-5.72], a sua classe é *Virginica*;
- Se o valor de Petal Length for igual ou superior a 4.9 e Petal Width estiver no intervalo [0.6-1.5], a sua classe é *Virginica*;

4 Wine Quality

O terceiro dataSet escolhido foi o "Wine Quality" fornecido por (Paulo Cortez,2009). O objetivo principal deste estudo passa por prever a qualidade do vinho com base em dados físico-químicos. Este estudo também foi conduzido para identificar anormalidades ou anomalias no conjunto de amostras de vinho, a fim de detetar adulteração no vinho. Este *dataset* é composto por dois conjuntos de dados de análise química de vinhos: um conjunto de amostras de *White Wine* e outro de *Red Wine* do vinho português "Vinho Verde".

O conjunto de dados contém 1599 instâncias para *Red Wine* e 4989 instâncias para *White Wine*. Cada uma destas instâncias é composta por 12 variáveis físico-químicas:*Fixed Acidity, Volatile acidity, Citric Acid, residual sugar, Chlorides, Free Sulpher dioxide, Total sulfur dioxide, density, pH, sulfates, Alcohol* e uma qualidade de avaliação do respetivo vinho. A classificação de qualidade é baseada em um teste de sabor sensorial com valores entre 0 (muito mau) até 10 (excelente).

4.1 Descrição dos Atributos

Lista dos atributos com uma simples descrição de cada um deles:[1]

Atributo	**Descrição**	**Tipo**	**Valores**
Fixed Acidity	Acidez fixa	valor numérico	Continuo
Volatile acidity	Acidez volátil	valor numérico	Continuo
Citric acid	Ácido cítrico	valor numérico	Continuo
Residual sugar	Açúcar residual	valor numérico	Continuo
Chlorides	Cloretos	valor numérico	Continuo
Free sulpher dioxide	Dióxido de sulfato livre	valor numérico	Continuo
Total sulpher dioxide	Dióxido de sulfato total	valor numérico	Continuo
Density	Densidade	valor numérico	Continuo
pH	Nível de Acidez	valor numérico	Continuo
Sulphates	Sulfatos	valor numérico	Continuo
Alcohol	Álcool	valor numérico	Continuo

Tabela 2: Wine Quality *Dataset* - Descrição dos Atributos

4.2 Objetivos

Depois de efetuada leitura do artigo e de fazermos uma primeira análise ao *dataset*, foram estabelecidos os seguintes objetivos:

1. Usar os dados fornecidos para prever o tipo de vinho, ou seja, um processo de classificação do vinho com base nos dados de análise química;

2. Prever a qualidade do vinho;

4.3 Data Mining - Objetivo 1

4.3.1 Preparação dos Dados

Antes de podermos aplicar qualquer técnica de *Data Mining* no conjunto de dados, efetuaram-se algumas alterações nestes, realizando o pré-processamento. O pré-processamento de dados é uma questão de extrema importância pois os dados do mundo real tendem a ser incompletos, ruidosos e inconsistentes. Esta etapa inclui discretização, limpeza, integração, transformação e a redução dos dados.

[1] Ambos os *DataSets* contêm os mesmos atributos

Como o objetivo é prever o tipo de vinho e temos dois *datasets*, ou seja, duas fontes de dados, foi necessário fazer a integração dos dados de forma a compor a informação numa coleção coerente e integrada de dados. Deste modo, o grupo juntou todas as instâncias num único *DataSet*, acrescentando o atributo "Kind", que apresenta o tipo de vinho de determinada instância, de modo a diferenciar os dois tipos de vinho existentes.

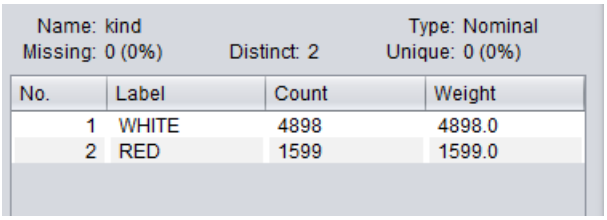

Figura 24: Integração dos Dados - Novo atributo "Kind"

Continuando a tarefa de pré-processamento, no processo de transformação dos dados decidimos restringir, já que era o objetivo do estudo, que o atributo de output "quality"apenas apresente valores inteiros entre 0-10, como podemos ver de seguida, de forma a ajudar no processo de mineração.

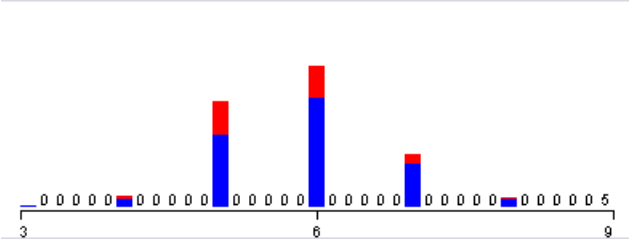

Figura 25: Análise atributo "quality"

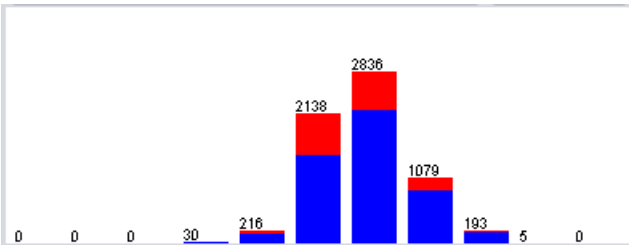

Figura 26: Análise atributo "quality"após transformação

Numa primeira abordagem, decidimos alterar todos os atributos numéricos em nominais de acordo com intervalos manualmente selecionados, reduzindo o número de valores possíveis de cada atributo nominal. O propósito era o de descomplicar os cálculos e tornar os resultados mais fáceis de interpretar. Na realidade, esta abordagem levou a que a capacidade de previsão fosse ligeiramente menor depois de aplicados os separadores, o que é resultado da perda de informação disponível no *dataset* integral.

Também não foi possível reduzir o número de valores o suficiente de forma a tornar uma árvore de decisão completamente legível. Os tempos de execução dos algoritmos aplicados não são muito grandes, não sendo, justificável a simplificação dos cálculos. Por tudo isto, decidimos não aplicar esta redução da informação presente no *dataset*.

4.3.2 Classificação

De forma a analisar o melhor possível o *dataset*, foram utilizados seis classificadores diferentes. Dois classificadores baseados em árvores de decisão (o J48 e o REPTree), três classificadores de regras de classificação (o PART, o JRip e o OneR) e o classificador probabilístico Naive Bayes.

Foram obtidos os seguintes resultados:

- REPTree
  - Tamanho da árvore (número de nodos): 41
  - Correctly Classified Instances: 6388 (98.3223 %)
  - Incorrectly Classified Instances: 109 (1.6777 %)
  - Kappa statistic: 0.9544
  - Mean absolute error: 0.025
  - Root mean squared error: 0.1236
  - Relative absolute error: 6.7307
  - Root relative squared error: 28.6873
- J48
  - Tamanho da árvore (número de nodos): 61
  - Correctly Classified Instances: 6414 (98.7225 %)
  - Incorrectly Classified Instances: 83 (1.2775 %)
  - Kappa statistic: 0.9654
  - Mean absolute error: 0.0161
  - Root mean squared error: 0.111
  - Relative absolute error: 4.3425 %
  - Root relative squared error: 25.7788 %
- PART
  - Regras: 25
  - Correctly Classified Instances: 6430 (98.9688 %)
  - Incorrectly Classified Instances: 67 (1.0312 %)
  - Kappa statistic: 0.9721
  - Mean absolute error: 0.0116
  - Root mean squared error: 0.0978
  - Relative absolute error: 3.1345 %
  - Root relative squared error: 22.6951 %
- JRip
  - Regras: 11
  - Correctly Classified Instances: 6414 (98.7225 %)
  - Incorrectly Classified Instances: 83 (1.2775 %)
  - Kappa statistic: 0.9655
  - Mean absolute error: 0.0163
  - Root mean squared error: 0.1107

  - Relative absolute error: 4.3917 %
  - Root relative squared error: 25.7 %

- OneR
  - Regra:

```
total sulfur dioxide:
        < 56.5  -> RED
        < 57.5  -> WHITE
        < 60.5  -> RED
        < 61.5  -> WHITE
        < 66.5  -> RED
        >= 66.5 -> WHITE
```

  - Correctly Classified Instances: 5944 (91.4884 %)
  - Incorrectly Classified Instances: 553 (8.5116 %)
  - Kappa statistic: 0.7623
  - Mean absolute error: 0.0851
  - Root mean squared error: 0.2917
  - Relative absolute error: 22.9345 %
  - Root relative squared error: 67.7307 %

- Naive Bayes
  - Correctly Classified Instances: 6338 (98.7225 %)
  - Incorrectly Classified Instances: 83 (2.4473 %)
  - Kappa statistic: 0.9348
  - Mean absolute error: 0.0301
  - Root mean squared error: 0.149
  - Relative absolute error: 8.103 %
  - Root relative squared error: 34.589 %

De um modo global, os resultados obtidos de todos os algoritmos apresentados são muito idênticos, obtendo-se um erro de classificação entre 1,0% e 2,4%. O algoritmo OneR é uma exceção, pois gera piores resultados, o que corresponde às expetativas uma vez que o principal objetivo deste classificador é a conceção de uma única regra de classificação, mostrando que o tipo do vinho é bastante influenciado pelo atributo "dióxido de sulfúrico total", ou seja, quando o valor deste é menor que 56.5, o tipo do vinho em questão é "RED" e quando é superior ou igual a 66.5 é do tipo "WHITE". Relativamente aos algoritmos de criação de árvores, os resultados foram bastante idênticos. O algoritmo J48 foi capaz de criar uma árvore mais pequena, obtendo com isto resultados ligeiramente menos precisos.

Para finalizar, para os algoritmos de criação de regras (à exceção do OneR) observam-se resultados também bastante idênticos. Com 11 regras, os resultados do JRip não difere praticamente em nada dos do PART, que criou 25 regras, apesar de neste caso o PART ter obtido resultados ligeiramente melhores.

4.3.3 Conclusões

Com objetivo de adquirir a capacidade de previsão do tipo de vinho em questão através dos dados físico-químicos, uma grande variedade de algoritmos de classificação foram testados e analisados ao pormenor. Como podemos ver em cima, o algoritmo com menor erro foi PART com apenas 67 instâncias da seleção defeituosamente classificadas através de 25 regras criadas. Os restantes algoritmos alcançaram resultados bastante próximos a este. Assim, podemos concluir que é possível prever com confiança perto de 100%.

4.4 Data Mining - Objetivo 2

4.4.1 Preparação dos Dados

Para este objetivo, o grupo decidiu não juntar os dois *datasets* disponibilizados, mas sim tratar individualmente cada um deles e seus atributos para os dois tipos de vinho. De seguida, verificou-se a presença de atributos redundantes, chegando à conclusão de que o atributo "pH"não traz informação importante pois todas as instâncias deste *dataset* contem o "pH"entre 2.7 e 4, o que mostra que todos os vinhos são ácidos, não havendo necessidade de fazer esta distinção. Decidimos então remover este atributo pois não perdemos informação necessária e importante. Na parte da limpeza de dados, o objetivo é preencher valores em falta, identificar valores atípicos ou corrigir inconsistências nos dados, no entanto os dados desta fonte já vinham tratados, evitando assim este passo.

Tal como na análise ao objetivo de cima, decidiu-se alterar os valores possíveis da variável de output "quality"uma vez que só pode haver valores inteiros entre 0-10 conforme explicado anteriormente.

Attributes	*Red Wine*			*White wine*		
	Min	*Max*	*Mean*	*Min*	*Max*	*Mean*
Fixed acidity	4.6	15.9	8.3	3.8	14.2	6.9
Volatile acidity	0.1	1.6	0.5	0.1	1.1	0.3
Citric acid	0.0	1.0	0.3	0.0	1.7	0.3
Residual sugar	0.9	15.5	2.5	0.6	65.8	6.4
Chlorides	0.01	0.61	0.08	0.01	0.35	0.05
Free sulpher dioxide	1	72	14	2	289	35
Total sulfur dioxide	6	289	46	9	440	138
Density	0.990	1.004	0.996	0.987	1.039	0.994
pH	2.7	4.0	3.3	2.7	3.8	3.1
Sulphates	0.3	2.0	0.7	0.2	1.1	0.5
Alcohol	8.4	14.9	10.4	8.0	14.2	10.4

Figura 27: Dados estatísticos físico-químicos por tipo de vinho

Para cada um dos dois *datasets*, procedeu-se o cálculo da relevância de cada um dos atributos. Para isto, foi usado o filtro de *attribute selection* de maneira a chegar a um resultado. Deste modo, utilizou-se o algoritmo de seleção *CfsSubsetEval* e o algoritmo de Procura *BestFirst*. Os resultados foram os seguintes:

Red wine dataset	White wine dataset
Volatile Acidity (2)	Volatile Acidity (2)
Total Sulfur Dioxide (7)	Citric Acid (3)
Sulphate (10)	Chlorides (5)
Alcohol (11)	Free Sulfur Dioxide (6)
	Density (8)
	Alcohol (11)

Figura 28: Ranking dos Atributos usando algoritmo attribute selection

4.4.2 Associação

Utilizar os dados no seu formato original para elaborar modelos é inadequado devido a algumas deficiências. Um dos problemas é a grande amplitude dos valores dos atributos devido à sua natureza diferente ou às diferentes unidades de medidas desses valores, por exemplo o "Total Sulfure Dioxide"(6 - 289) comparativamente ao atributo "sulphates"(0,3 - 2). Tal inconsistência pode afetar as habilidades preditivas dos modelos, fazendo com que alguns atributos sejam mais "influentes"do que outros.

Para encontramos regras de associação relevantes ao problema que decidimos estudar, foi feita uma discretizacão de igual altura e obtidas as classificações para cada atributo relevante, consoante os seus intervalos. Os intervalos estão representados na tabela seguinte.

Red Wine	Classificação		
Atributo	Baixa	Média	Alta
Volatile Acidity	[-inf-0.425]	[0.425-0.5975]	[0.5975- +inf]
Total Sulfur Dioxide	[-inf-26.5]	[26.5-51.5]	[51.5- +inf]
Sulphates	[-inf-0.575]	[0.575-0.685]	[0.685- +inf]
Alcohol	[-inf-9.75]	[9.75-10.85]	[10.85- +inf]

Figura 29: Discretização Red Wine

White Wine	Classificação		
Atributo	Baixa	Média	Alta
Volatile Acidity	[-inf-0.2275]	[0. 2275-0.2975]	[0. 2975- +inf]
Citric Acid	[-inf-0.285]	[0. 285-0.355]	[0. 355- +inf]
Chlorides	[-inf-0.0375]	[0. 0375-0. 0475]	[0. 0475- +inf]
Free Sulfur Dioxide	[-inf-26.5]	[26.5-40,75]	[40.75- +inf]
Density	[-inf-0.992395]	[0.992395-0.995345]	[0.995345- +inf]
Alcohol	[-inf-9.72]	[9.72-11.025]	[11.025- +inf]

Figura 30: Discretização White Wine

Depois de feita a discretização dos valores dos atributos, obteve-se as seguintes regras de associação:

- White Wine

```
Best rules found:

 1. free sulfur dioxide='(26.5-40.75]' alcohol='(11.025-inf)' 617 ==> density='(-inf-0.992395]' 530    <conf:(0.86)> lift:(2.59) lev:(0.07) [325] conv:(4.69)
 2. density='(0.995345-inf)' quality=5 720 ==> alcohol='(-inf-9.716666]' 611    <conf:(0.85)> lift:(2.48) lev:(0.07) [364] conv:(4.31)
 3. volatile acidity='(0.2975-inf)' density='(-inf-0.992395]' 612 ==> alcohol='(11.025-inf)' 519    <conf:(0.85)> lift:(2.66) lev:(0.07) [323] conv:(4.44)
 4. chlorides='(-inf-0.0375]' density='(-inf-0.992395]' 972 ==> alcohol='(11.025-inf)' 818    <conf:(0.84)> lift:(2.64) lev:(0.1) [508] conv:(4.27)
 5. citric acid='(0.285-0.355]' alcohol='(11.025-inf)' 608 ==> density='(-inf-0.992395]' 511    <conf:(0.84)> lift:(2.54) lev:(0.06) [309] conv:(4.15)
 6. free sulfur dioxide='(40.75-inf)' alcohol='(-inf-9.716666]' 858 ==> density='(0.995345-inf)' 719    <conf:(0.84)> lift:(2.51) lev:(0.09) [432] conv:(4.08)
 7. chlorides='(-inf-0.0375]' alcohol='(11.025-inf)' 984 ==> density='(-inf-0.992395]' 818    <conf:(0.83)> lift:(2.51) lev:(0.1) [491] conv:(3.94)
 8. free sulfur dioxide='(26.5-40.75]' density='(-inf-0.992395]' 638 ==> alcohol='(11.025-inf)' 530    <conf:(0.83)> lift:(2.61) lev:(0.07) [326] conv:(3.99)
 9. alcohol='(11.025-inf)' 1561 ==> density='(-inf-0.992395]' 1270    <conf:(0.81)> lift:(2.46) lev:(0.15) [752] conv:(3.57)
10. citric acid='(0.355-inf)' alcohol='(-inf-9.716666]' 685 ==> density='(0.995345-inf)' 555    <conf:(0.81)> lift:(2.43) lev:(0.07) [326] conv:(3.48)
```

Figura 31: Algoritmo de Associação com mínimo de confiança 0.8 e mínimo suporte 0.1

```
Best rules found:

 1. free sulfur dioxide='(26.5-40.75]' alcohol='(11.025-inf)' 617 ==> density='(-inf-0.992395]' 530    <conf:(0.86)> lift:(2.59) lev:(0.07) [325] conv:(4.
 2. density='(0.995345-inf)' quality=5 720 ==> alcohol='(-inf-9.716666]' 611    <conf:(0.85)> lift:(2.48) lev:(0.07) [364] conv:(4.31)
 3. volatile acidity='(0.2975-inf)' density='(-inf-0.992395]' 612 ==> alcohol='(11.025-inf)' 519    <conf:(0.85)> lift:(2.66) lev:(0.07) [323] conv:(4.44)
 4. chlorides='(-inf-0.0375]' density='(-inf-0.992395]' 972 ==> alcohol='(11.025-inf)' 818    <conf:(0.84)> lift:(2.64) lev:(0.1) [508] conv:(4.27)
 5. citric acid='(0.285-0.355]' alcohol='(11.025-inf)' 608 ==> density='(-inf-0.992395]' 511    <conf:(0.84)> lift:(2.54) lev:(0.06) [309] conv:(4.15)
 6. free sulfur dioxide='(40.75-inf)' alcohol='(-inf-9.716666]' 858 ==> density='(0.995345-inf)' 719    <conf:(0.84)> lift:(2.51) lev:(0.09) [432] conv:(4
 7. chlorides='(-inf-0.0375]' alcohol='(11.025-inf)' 984 ==> density='(-inf-0.992395]' 818    <conf:(0.83)> lift:(2.51) lev:(0.1) [491] conv:(3.94)
 8. free sulfur dioxide='(26.5-40.75]' density='(-inf-0.992395]' 638 ==> alcohol='(11.025-inf)' 530    <conf:(0.83)> lift:(2.61) lev:(0.07) [326] conv:(3.
 9. alcohol='(11.025-inf)' 1561 ==> density='(-inf-0.992395]' 1270    <conf:(0.81)> lift:(2.46) lev:(0.15) [752] conv:(3.57)
10. citric acid='(0.355-inf)' alcohol='(-inf-9.716666]' 685 ==> density='(0.995345-inf)' 555    <conf:(0.81)> lift:(2.43) lev:(0.07) [326] conv:(3.48)
11. citric acid='(0.285-0.355]' density='(-inf-0.992395]' 638 ==> alcohol='(11.025-inf)' 511    <conf:(0.8)> lift:(2.51) lev:(0.06) [307] conv:(3.4)
12. free sulfur dioxide='(-inf-26.5]' alcohol='(11.025-inf)' 630 ==> density='(-inf-0.992395]' 503    <conf:(0.8)> lift:(2.41) lev:(0.06) [294] conv:(3.29
13. alcohol='(11.025-inf)' quality=6 718 ==> density='(-inf-0.992395]' 571    <conf:(0.8)> lift:(2.4) lev:(0.07) [333] conv:(3.24)
14. citric acid='(0.355-inf)' density='(0.995345-inf)' 700 ==> alcohol='(-inf-9.716666]' 555    <conf:(0.79)> lift:(2.32) lev:(0.06) [315] conv:(3.15)
15. volatile acidity='(0.2975-inf)' alcohol='(11.025-inf)' 659 ==> density='(-inf-0.992395]' 519    <conf:(0.79)> lift:(2.38) lev:(0.06) [300] conv:(3.13)
16. free sulfur dioxide='(40.75-inf)' density='(0.995345-inf)' 918 ==> alcohol='(-inf-9.716666]' 719    <conf:(0.78)> lift:(2.29) lev:(0.08) [405] conv:(3
17. density='(-inf-0.992395]' 1623 ==> alcohol='(11.025-inf)' 1270    <conf:(0.78)> lift:(2.46) lev:(0.15) [752] conv:(3.12)
18. chlorides='(0.0475-inf)' density='(0.995345-inf)' 864 ==> alcohol='(-inf-9.716666]' 674    <conf:(0.78)> lift:(2.28) lev:(0.08) [378] conv:(2.98)
19. density='(-inf-0.992395]' quality=6 737 ==> alcohol='(11.025-inf)' 571    <conf:(0.77)> lift:(2.43) lev:(0.07) [336] conv:(3.01)
20. free sulfur dioxide='(-inf-26.5]' density='(-inf-0.992395]' 668 ==> alcohol='(11.025-inf)' 503    <conf:(0.75)> lift:(2.36) lev:(0.06) [290] conv:(2.7
```

Figura 32: Algoritmo de Associação com mínimo de confiança 0.7 e mínimo suporte 0.1 e máximo 20 regras

- Red Wine

```
Best rules found:

 1. total sulfur dioxide='(51.5-inf)' sulphates='(-inf-0.575]' alcohol='(-inf-9.75]' 113 ==> quality=5 99    <conf:(0.88)> lift:(2.06) lev:(0.03) [50] conv:(4.32)
 2. volatile acidity='(0.5975-inf)' total sulfur dioxide='(51.5-inf)' alcohol='(-inf-9.75]' 106 ==> quality=5 88    <conf:(0.83)> lift:(1.95) lev:(0.03) [42] conv:(3.2)
 3. volatile acidity='(0.5975-inf)' total sulfur dioxide='(51.5-inf)' sulphates='(-inf-0.575]' 100 ==> quality=5 81    <conf:(0.81)> lift:(1.9) lev:(0.02) [38] conv:(2.
 4. volatile acidity='(0.5975-inf)' sulphates='(-inf-0.575]' alcohol='(-inf-9.75]' 124 ==> quality=5 99    <conf:(0.8)> lift:(1.87) lev:(0.03) [46] conv:(2.74)
 5. sulphates='(-inf-0.575]' alcohol='(-inf-9.75]' 244 ==> quality=5 193    <conf:(0.79)> lift:(1.86) lev:(0.06) [89] conv:(2.69)
 6. total sulfur dioxide='(51.5-inf)' alcohol='(-inf-9.75]' 252 ==> quality=5 197    <conf:(0.78)> lift:(1.84) lev:(0.06) [89] conv:(2.58)
 7. volatile acidity='(-inf-0.425]' quality=7 132 ==> alcohol='(10.85-inf)' 100    <conf:(0.76)> lift:(2.35) lev:(0.04) [57] conv:(2.71)
 8. volatile acidity='(0.5975-inf)' alcohol='(-inf-9.75]' 222 ==> quality=5 166    <conf:(0.75)> lift:(1.76) lev:(0.04) [71] conv:(2.24)
 9. total sulfur dioxide='(51.5-inf)' sulphates='(-inf-0.575]' 191 ==> quality=5 141    <conf:(0.74)> lift:(1.73) lev:(0.04) [59] conv:(2.15)
10. sulphates='(0.685-inf)' quality=7 132 ==> volatile acidity='(-inf-0.425]' 97    <conf:(0.73)> lift:(2.29) lev:(0.03) [54] conv:(2.49)
```

Figura 33: Algoritmo de Associação com mínimo de confiança 0.7 e mínimo suporte 0.1

Assim, é possível retirar várias conclusões:

- Com grau de confiança de 86%, sempre que "Free Sulfur Dioxide"tiver valores entre 26.5 e 40.75 e o nível de álcool superior a 11.626, a densidade do vinho será inferior a 0.9923.
- Com grau de confiança de 85%, quando a densidade for superior a 0.9955 e for classificado com qualidade 5, o nível de álcool será inferior a 9.72.
- Com grau de confiança de 81%, sempre que valor do álcool for superior a 11.025, então a densidade será inferior a 0.9923.

4.4.3 Classificação

Para este objetivo, uma das análises que nos pareceu fazer mais sentido e ser bastante interessante de trabalhar foi a classificação. Escolhemos o algoritmo J48 baseado em árvores, pois permite-nos ver graficamente e de forma menos complexa os resultados. É de referir ainda, como já foi dito num dos temas anteriores, as razões para que seja utilizada a versão do *dataset* sem discretização dos atributos.

- Red Wine

```
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         951               59.4747 %
Incorrectly Classified Instances       648               40.5253 %
Kappa statistic                          0.3584
Mean absolute error                      0.0846
Root mean squared error                  0.2417
Relative absolute error                 72.1679 %
Root relative squared error             99.993  %
Total Number of Instances             1599


=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   <-- classified as
   0   0   0   0   0   0   0   0   0   0   0 |   a = 0
   0   0   0   0   0   0   0   0   0   0   0 |   b = 1
   0   0   0   0   0   0   0   0   0   0   0 |   c = 2
   0   0   0   0   2   6   2   0   0   0   0 |   d = 3
   0   0   0   2   6  31  13   1   0   0   0 |   e = 4
   0   0   0   3   9 487 163  18   1   0   0 |   f = 5
   0   0   0   1   7 178 372  75   5   0   0 |   g = 6
   0   0   0   1   2  17  91  86   2   0   0 |   h = 7
   0   0   0   0   0   0   5  13   0   0   0 |   i = 8
   0   0   0   0   0   0   0   0   0   0   0 |   j = 9
   0   0   0   0   0   0   0   0   0   0   0 |   k = 10
```

Figura 34: Resultados do Algoritmo J48 - Red Wine

- White Wine

```
Time taken to build model: 0.26 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2851               58.2074 %
Incorrectly Classified Instances      2047               41.7926 %
Kappa statistic                          0.3752
Mean absolute error                      0.0832
Root mean squared error                  0.2523
Relative absolute error                 67.6543 %
Root relative squared error            101.7995 %
Total Number of Instances             4898
```

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g    h    i    j    k   <-- classified as
   0    0    0    0    0    0    0    0    0    0    0 |    a = 0
   0    0    0    0    0    0    0    0    0    0    0 |    b = 1
   0    0    0    0    0    0    0    0    0    0    0 |    c = 2
   0    0    0    1    4    5    6    2    2    0    0 |    d = 3
   0    0    0    2   34   78   40    7    2    0    0 |    e = 4
   0    0    0    5   59  902  417   65    9    0    0 |    f = 5
   0    0    0    5   31  439 1414  262   47    0    0 |    g = 6
   0    0    0    2    8   62  331  447   30    0    0 |    h = 7
   0    0    0    1    0   15   50   56   53    0    0 |    i = 8
   0    0    0    0    0    0    2    3    0    0    0 |    j = 9
   0    0    0    0    0    0    0    0    0    0    0 |    k = 10
```

Figura 35: Resultados do Algoritmo J48 - White Wine

Depois de analisar os resultados deste algoritmo conseguimos tirar algumas conclusões relativamente ao atributo de output "quality". Inicialmente consegue-se ver uma taxa a rondar os 60% de instâncias classificadas corretamente para ambos os tipos de vinhos. O erro encontra-se a rondar os 68%. Já na *Confusion Matrix*, os valores que se encontram na diagonal são as instâncias corretas, no entanto valores que se encontram noutras células encontram se mal calculados, desviando-se estes algumas células da diagonal.

4.5 Análise de Resultados

A partir da precisão da classificação resultante, descobrimos que a taxa de precisão para o "White Wine"é influenciada por um maior número de atributo de físico-químicos. Enquanto isso, a qualidade do "Red Wine"é altamente correlacionada a apenas quatro atributos. Isso mostra que a qualidade do White Wine é afetada por atributos que não afetam o Red Wine em geral.

Após analise dos resultados verificamos que os resultados não eram propriamente os esperados relativamente ao objetivo 2, devido ao número elevado de casos que foram mal calculados. Isto deve-se à fraca correlação entre atributos, uma vez que o atributo qualidade deste estudo é baseado num teste de sabor sensorial, ou seja, a qualidade atribuída a cada instância não esta relacionada com as várias características apresentadas pelo vinho. Isto é normal uma vez que se trata de uma opinião relativa, baseando-se nos seus gostos pessoais.

5 Conclusão

Terminados todos os estudos e análises feitas aos três diferentes *datasets*, conclui-se que a contextualização, perceção do tema de trabalho para definição de objetivos e o pré-tratamento de dados representam a maioria do tempo de trabalho, sendo estes de extrema importância.

A correta realização desta primeira parte, revelou-se fundamental, pois permite alcançar melhores resultados nas análises feitas posteriormente. A redundância de certos atributos nos *datasets* também foi um aspeto interessante devido à influência que pode ter.

Apesar da falta de experiência e sendo o primeiro contacto dos elementos do grupo com este tipo de projeto, foi possível compreender a importância destas análises e quais as vantagens que estas podem ter em ambientes empresariais.

Neste sentido, com o objetivo de aprender o maior número de conceitos, o grupo, na escolha dos *datasets*, optou por tentar manter uma variedade não só em termos de tema, como também em termos de atributos e instâncias, de maneira a sermos expostos a diferentes contextos de análise de dados.

Em suma, o grupo ficou bastante satisfeito com o trabalho desenvolvido, uma vez que compreendeu os conceitos e metodologias associadas às diversas fases e algoritmos existentes para extração de conhecimento.

Referências

[1] Fisher,R.A. "The use of multiple measurements in taxonomic problems"Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics"(John Wiley, NY, 1950).

[2] https://archive.ics.uci.edu/ml/datasets/iris

[3] IJCSMC, Vol. 4 (2015), *APRIORI ALGORITHM AND FILTERED ASSOCIATOR IN ASSOCIATION RULE MINING*, International Journal of Computer Science and Mobile Computing.

[4] http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

[5] https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/

[6] http://weka.sourceforge.net/doc.dev/weka/

[7] http://www3.dsi.uminho.pt/pcortez/wine5.pdf

[8] https://www.ibm.com/developerworks/br/opensource/library/os-weka2/index.html