Knowledge Extraction of Structured Data to application of Classification, Clustering and Association

António Silva[1],
Marcelo Queirós Pinto[1],
[1] Universidade do Minho, Braga, Portugal
A73827@alunos.uminho.pt, mqspinto@gmail.com

Abstract. Ninety percent of the data in the world today has been created only in the last two years, according to IBM. Data is big, and getting bigger. In this context, Knowledge Extraction is the creation of knowledge from structured or unstructured data and allow the process of discovering useful knowledge from a collection of data. In this paper we will analyse 4 datasets and apply data mining algorithms to classify, segment and create associations in all 4 datasets.

Keywords: Knowledge Extraction, Data pre-processing, Data Mining, Classification, Clustering, Association, Artificial Intelligence.

1 Introdução

O produto final da extração de conhecimento são dados relevantes (informação) para ser utilizada por algoritmos de *Data Mining*, com o objetivo de ajudar o ser humano a produzir conhecimento, a partir dos resultados desses algoritmos. Assim, a extração de conhecimento é a área que se foca no pré processamento de dados, estudando o problema em questão, determinando as suas variáveis, permitindo escolher e tratar os atributos dos quais o nosso problema depende.

Normalmente, a extração de conhecimento é utilizada para se poder resolver um determinado problema. Três tipos comuns de problema são: classificação, segmentação e associação. A classificação permite desenvolver um modelo que prevê um determinado atributo a partir de outros; a segmentação permite dividir instâncias em subgrupos e a associação permite criar regras automaticamente a partir da análise estatística dos dados, abrindo portas para descobrir várias associações entre os dados. Estes tipos de problemas são tratados a partir do pré processamento dos dados e da consequente aplicação de algoritmos de *data mining* específicos para o tipo de problema em questão.

Neste trabalho decidiu-se fazer o pré processamento de 4 *datasets* individualmente para os 3 tipos de problema em questão. Assim, serão feitos 12 pré processamentos, cada um seguido da respetiva aplicação de vários algoritmos com o objetivo de solucionar um tipo de problema: classificação, segmentação ou associação.

2 Considerações

2.1 Boosting

No problema especifico classificação iremos abordar vários algoritmos de classificação. Optamos por utilizar Boosting em grande parte dos mesmos, de forma a otimizar os modelos. Assim, iremos explicar este processo:

Boosting é um processo que queria o mesmo modelo N vezes, testando-o e no modelo seguinte aumenta o peso dos casos que deram erro no primeiro modelo. Assim repete o processo com um maior peso nesses casos. Por exemplo: Num primeiro modelo temos a Figura 1, na qual existem 3 erros:

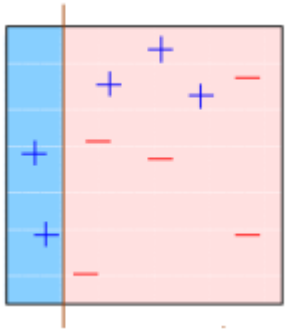


Figura 1 - 1º Iteração

O peso desses erros é aumentado e recalculado um novo modelo, que também pode dar errado:

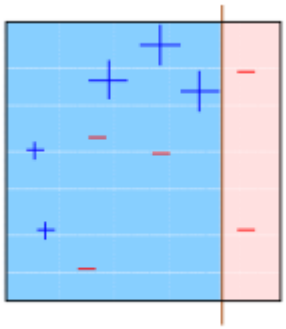


Figura 2 - 2º Iteração

No caso seguinte o peso das instências rodeadas é aumentado:

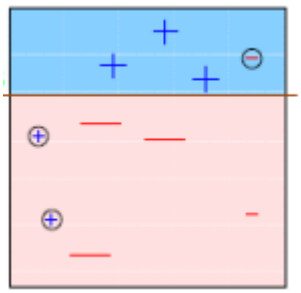


Figura 3 - 3º Iteração

E por fim temos um modelo provavelmente melhor:

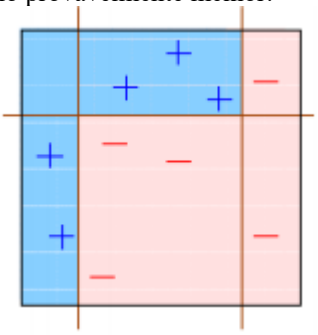


Figura 4 - Modelo Final

2.2. Comparação de Classificadores

A percentagem de acertos no total das instâncias revela-se um classificador de modelos insuficiente. Exemplo: na previsão de uma fraude queremos um classificador que não nos dê falsos negativos (haver fraude e o modelo dar negativo): assim queremos um Recall alto, não importando tanto a percentagem de acertos no total das instâncias (pode haver falsos positivos, não queremos mesmo é falsos negativos).

Assim, a comparação de classificadores irá ser feita com base em 4 parâmetros considerados essenciais:

- Percentagem de acertos no total das instâncias (Muito utilizado, mas insuficiente na comparação entre modelos);
- *Roc Area*: A *Roc Curve* é uma curva que nos permite visualizar todos os possíveis *thresholds* que separam as classes, dando o FT_Rate (FT = False Positive) e TP_Rate (TP = True Positive) para cada classificador. Assim a ROC área permite-nos avaliar o desempenho de um modelo para qualquer classificador, não tendo em conta um especifico;

- *Precision*: Percentagem de itens selecionados corretos;
- *Recall*: Percentagem de itens corretos selecionados.

De realçar que estes 4 parâmetros nos dão uma total visão sobre um modelo. O parâmetro *F-Measure* não será tido em conta uma vez que relaciona o parâmetro *Precision* e *Recall*. Assim, preferimos optar pela abordagem de avaliar modelos com classificadores mais específicos, podendo ser concluído onde um modelo falha exatamente.

3 *Dataset* ConcreteData

3.1 Objetivos de Estudo e Significado dos Dados

O betão é o material mais importante na área da construção. Resulta na mistura em proporção adequada de cimento, agregados e água, cujas características diferem substancialmente daquelas apresentadas pelos elementos que o constituem.

Assim, é de extrema importância a avaliação da sua resistência à compressão. Isto é, a partir de dados referentes à sua composição e idade, é possível extrair a sua força compressiva, uma das principais propriedades mecânicas do betão (evitando dezenas de ensaios em corpos-de-prova).

Para tal, existem variáveis quantitativas que são definitivas para a previsão deste parâmetro. São elas o cimento (componente 1); lâminas de alto nível (componente 2); cinzas de combustível pulverizado (componente 3); água (componente 4); superplastificante (componente 5); agregado grosseiro (componente 6) e agregado fino (componente 7) (estando todas elas em kg por m^3 de mistura). Por fim, tem-se em conta a idade do material (em dias).

3.2 Classificação

3.2.1 Pré-processamento

O dataset ConcreteData foi analisado com vista à classificação, sendo efetuados testes, optando-se pela classificação do atributo *Concrete compressive strength.*

Todas as outras variaveis são necessárias para a previsão deste atributo.

Assim, observou-se a distribuição do novo atributo de forma a criar 2 *thresholds* e criar 3 classes num novo atributo nominal: Má, média ou boa compressão. Estas transformações foram geradas no software *Microsoft Excel*.

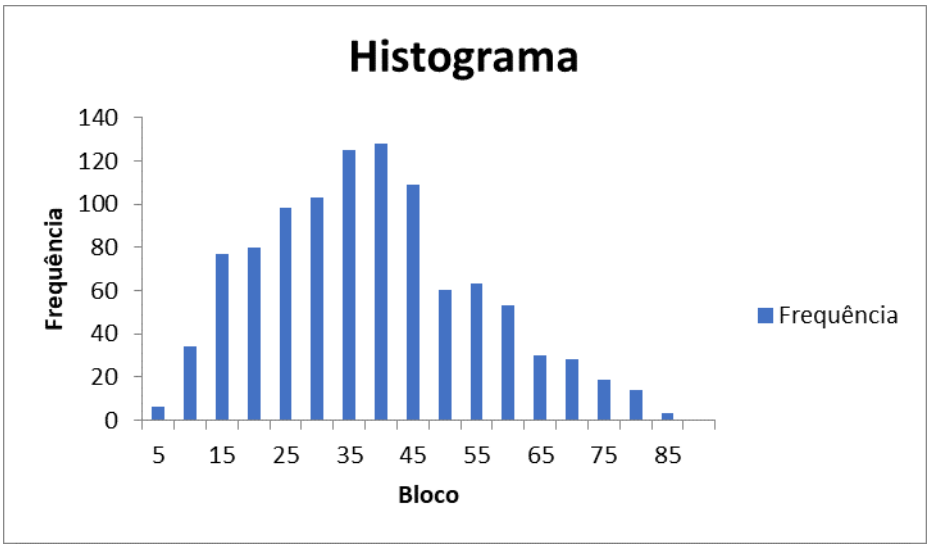


Figura 5 - Distribuição do atributo a prever

3.2.2 Algoritmos testados

Redes Neuronais

```
=== Summary ===

Correctly Classified Instances         907               88.0583 %
Incorrectly Classified Instances       123               11.9417 %
Kappa statistic                          0.7309
Mean absolute error                      0.0981
Root mean squared error                  0.2537
Relative absolute error                 33.3526 %
Root relative squared error             66.2072 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,681    0,031    0,688      0,681    0,684      0,653    0,946     0,728     Boa
                 0,912    0,199    0,921      0,912    0,916      0,708    0,928     0,968     Média
                 0,858    0,043    0,824      0,858    0,841      0,802    0,974     0,896     Má
Weighted Avg.    0,881    0,154    0,881      0,881    0,881      0,721    0,938     0,932
```

Figura 6 - Resultados - Multilayer Perceptron

```
=== Summary ===

Correctly Classified Instances         929               90.1942 %
Incorrectly Classified Instances       101                9.8058 %
Kappa statistic                          0.7734
Mean absolute error                      0.0741
Root mean squared error                  0.224
Relative absolute error                 25.2017 %
Root relative squared error             58.4635 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,713    0,018    0,798      0,713    0,753      0,731    0,971     0,824     Boa
               0,940    0,196    0,924      0,940    0,932      0,755    0,953     0,979     Média
               0,848    0,032    0,861      0,848    0,854      0,820    0,977     0,905     Má
Weighted Avg.  0,902    0,148    0,901      0,902    0,901      0,765    0,959     0,951
```

Figura 7 - Resultados - Multilayer Perceptron com AdaBoostM1

Support Vector Machine

```
=== Summary ===

Correctly Classified Instances         803               77.9612 %
Incorrectly Classified Instances       227               22.0388 %
Kappa statistic                          0.3185
Mean absolute error                      0.2716
Root mean squared error                  0.3514
Relative absolute error                 92.3794 %
Root relative squared error             91.7212 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,000    0,000      0,000    0,000      0,000    0,872     0,291     Boa
               0,989    0,753    0,769      0,989    0,866      0,398    0,618     0,769     Média
               0,365    0,010    0,900      0,365    0,520      0,523    0,797     0,503     Má
Weighted Avg.  0,780    0,542    0,724      0,780    0,720      0,386    0,676     0,674
```

Figura 8 - Resultados - SMO

```
=== Summary ===

Correctly Classified Instances         815               79.1262 %
Incorrectly Classified Instances       215               20.8738 %
Kappa statistic                          0.524
Mean absolute error                      0.1899
Root mean squared error                  0.3025
Relative absolute error                 64.5841 %
Root relative squared error             78.9645 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,585    0,050    0,539      0,585    0,561      0,516    0,833     0,465     Boa
                 0,858    0,378    0,852      0,858    0,855      0,482    0,847     0,923     Média
                 0,640    0,070    0,685      0,640    0,661      0,585    0,927     0,734     Má
Weighted Avg.    0,791    0,289    0,792      0,791    0,791      0,505    0,861     0,845
```

Figura 9 - Resultados - SMO com AdaBoostM1

Naive Bayes

```
=== Summary ===

Correctly Classified Instances         779               75.6311 %
Incorrectly Classified Instances       251               24.3689 %
Kappa statistic                          0.4834
Mean absolute error                      0.2322
Root mean squared error                  0.3428
Relative absolute error                 78.971  %
Root relative squared error             89.4754 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,532    0,038    0,581      0,532    0,556      0,514    0,767     0,447     Boa
                 0,788    0,323    0,861      0,788    0,823      0,440    0,794     0,879     Média
                 0,746    0,145    0,549      0,746    0,632      0,539    0,871     0,628     Má
Weighted Avg.    0,756    0,263    0,776      0,756    0,762      0,466    0,806     0,791
```

Figura 10 - Resultados – NaiveBayes

```
=== Summary ===

Correctly Classified Instances         760               73.7864 %
Incorrectly Classified Instances       270               26.2136 %
Kappa statistic                          0.4726
Mean absolute error                      0.215
Root mean squared error                  0.3513
Relative absolute error                 73.1063 %
Root relative squared error             91.7083 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,564    0,049    0,535      0,564    0,549      0,503    0,908     0,455     Boa
                 0,743    0,275    0,873      0,743    0,803      0,432    0,790     0,901     Média
                 0,802    0,173    0,523      0,802    0,633      0,544    0,904     0,712     Má
Weighted Avg.    0,738    0,235    0,775      0,738    0,747      0,460    0,822     0,824
```

Figura 11 - Resultados - NaiveBayes com AdaBoostM1

BayesNet

```
=== Summary ===

Correctly Classified Instances         850               82.5243 %
Incorrectly Classified Instances       180               17.4757 %
Kappa statistic                          0.5656
Mean absolute error                      0.1558
Root mean squared error                  0.2882
Relative absolute error                 52.9843 %
Root relative squared error             75.2172 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,532    0,033    0,617      0,532    0,571      0,534    0,941     0,688     Boa
                 0,927    0,433    0,845      0,927    0,884      0,544    0,862     0,931     Média
                 0,584    0,028    0,833      0,584    0,687      0,642    0,944     0,822     Má
Weighted Avg.    0,825    0,319    0,822      0,825    0,818      0,561    0,885     0,888
```

Figura 12 - Resultados - BayesNet

```
=== Summary ===

Correctly Classified Instances         911               88.4466 %
Incorrectly Classified Instances       119               11.5534 %
Kappa statistic                          0.731
Mean absolute error                      0.0847
Root mean squared error                  0.2434
Relative absolute error                 28.8087 %
Root relative squared error             63.5245 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,691    0,019    0,783      0,691    0,734      0,711    0,958     0,846     Boa
                 0,932    0,237    0,909      0,932    0,921      0,710    0,942     0,973     Média
                 0,797    0,038    0,831      0,797    0,813      0,771    0,968     0,904     Má
Weighted Avg.    0,884    0,179    0,883      0,884    0,883      0,722    0,949     0,948
```

Figura 13 - Resultados - BayesNet com AdaBoostM1

Árvores de Decisão – J48 (C4.5)

```
=== Summary ===

Correctly Classified Instances         917               89.0291 %
Incorrectly Classified Instances       113               10.9709 %
Kappa statistic                          0.7495
Mean absolute error                      0.0913
Root mean squared error                  0.2569
Relative absolute error                 31.0351 %
Root relative squared error             67.0507 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,702    0,026    0,733      0,702    0,717      0,690    0,918     0,660     Boa
                 0,927    0,199    0,922      0,927    0,924      0,731    0,900     0,938     Média
                 0,843    0,037    0,843      0,843    0,843      0,805    0,938     0,787     Má
Weighted Avg.    0,890    0,152    0,890      0,890    0,890      0,741    0,909     0,884
```

Figura 14 - Resultados - J48

```
=== Summary ===

Correctly Classified Instances         941               91.3592 %
Incorrectly Classified Instances        89                8.6408 %
Kappa statistic                          0.8018
Mean absolute error                      0.0569
Root mean squared error                  0.2275
Relative absolute error                 19.3625 %
Root relative squared error             59.3748 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,777    0,015    0,839      0,777    0,807      0,789    0,965     0,874     Boa
                 0,945    0,165    0,936      0,945    0,940      0,785    0,960     0,982     Média
                 0,863    0,032    0,863      0,863    0,863      0,831    0,978     0,931     Má
Weighted Avg.    0,914    0,126    0,913      0,914    0,913      0,794    0,964     0,963
```

Figura 15: Resultados - J48 com AdaBoostM1

Árvores de Decisão – *Random Florest*

```
=== Summary ===

Correctly Classified Instances         959               93.1068 %
Incorrectly Classified Instances        71                6.8932 %
Kappa statistic                          0.8388
Mean absolute error                      0.0781
Root mean squared error                  0.1908
Relative absolute error                 26.5761 %
Root relative squared error             49.8132 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,787    0,010    0,892      0,787    0,836      0,823    0,990     0,923     Boa
                 0,968    0,162    0,938      0,968    0,953      0,827    0,971     0,987     Média
                 0,863    0,018    0,919      0,863    0,890      0,866    0,985     0,938     Má
Weighted Avg.    0,931    0,120    0,930      0,931    0,930      0,834    0,975     0,972
```

Figura 16 - Resultados - Random Florest

```
=== Summary ===

Correctly Classified Instances         961               93.301  %
Incorrectly Classified Instances        69                6.699  %
Kappa statistic                          0.8433
Mean absolute error                      0.0458
Root mean squared error                  0.2101
Relative absolute error                 15.5848 %
Root relative squared error             54.8353 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,787    0,009    0,902      0,787    0,841      0,828    0,955     0,880     Boa
                 0,969    0,158    0,940      0,969    0,954      0,832    0,962     0,979     Média
                 0,868    0,018    0,919      0,868    0,893      0,869    0,972     0,928     Má
Weighted Avg.    0,933    0,118    0,932      0,933    0,932      0,839    0,963     0,960
```

Figura 17 - Resultados - Random Florest com AdaBoostM1

Árvores de Decisão – *Simple Cart*

```
=== Summary ===

Correctly Classified Instances         923               89.6117 %
Incorrectly Classified Instances       107               10.3883 %
Kappa statistic                          0.7578
Mean absolute error                      0.094
Root mean squared error                  0.2526
Relative absolute error                 31.9689 %
Root relative squared error             65.9292 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,702    0,017    0,805      0,702    0,750      0,729    0,907     0,685     Boa
                 0,942    0,216    0,917      0,942    0,929      0,742    0,883     0,923     Média
                 0,817    0,034    0,852      0,817    0,834      0,796    0,927     0,795     Má
Weighted Avg.    0,896    0,163    0,894      0,896    0,895      0,751    0,894     0,877
```

Figura 18 - Resultados - Random Florest

```
=== Summary ===

Correctly Classified Instances         953               92.5243 %
Incorrectly Classified Instances        77                7.4757 %
Kappa statistic                          0.8286
Mean absolute error                      0.0505
Root mean squared error                  0.2131
Relative absolute error                 17.1816 %
Root relative squared error             55.6325 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,809    0,014    0,854      0,809    0,831      0,814    0,979     0,910     Boa
               0,953    0,144    0,944      0,953    0,948      0,814    0,969     0,987     Média
               0,878    0,026    0,887      0,878    0,883      0,855    0,984     0,933     Má
Weighted Avg.  0,925    0,110    0,925      0,925    0,925      0,822    0,973     0,969
```

Figura 19 - Resultados - Random Florest com AdaBoostM1

Iremos sumariar estes resultados na penúltima secção deste dataset.

3.3 Segmentação

3.3.1 Pré-processamento

Não foi necessário qualquer pré processamento nos dados, utilizou-se o dataset de origem, sem qualquer atributo nominal.

De seguida utilizou-se o filtro *descretize*.

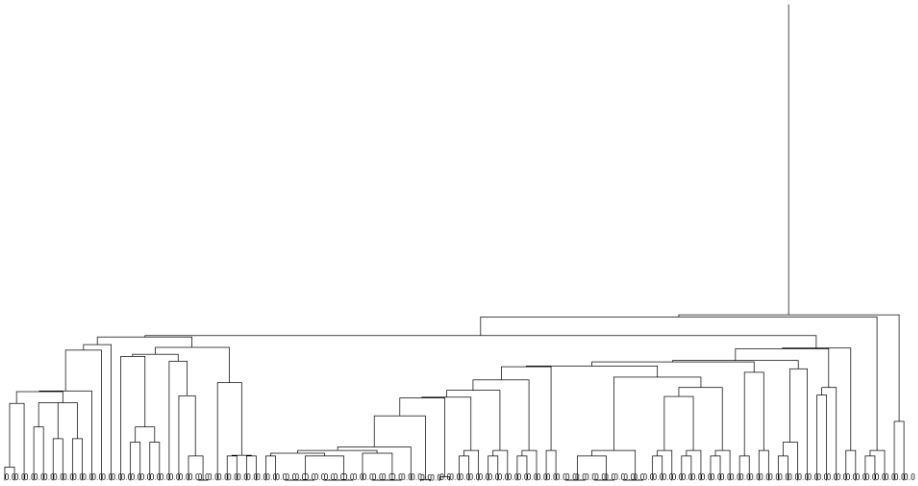
3.3.2 Algoritmos testados

EM

```
Clustered Instances

0       449 ( 44%)
1       387 ( 38%)
2       194 ( 19%)
```

Hierarchical Clusterer



Clustered Instances

0	94	(9%)
1	739	(72%)
2	197	(19%)

Simple K-Means

```
Final cluster centroids:
                                                        Cluster#
Attribute                                    Full Data          0          1
                                             (1030.0)    (831.0)    (199.0)
=============================================================================
Cement (component 1)(kg in a m^3 mixture)    281.1679   296.5433   216.9618
Blast Furnace Slag (component 2)(kg in a m^3 mixture)     73.8958    77.4794    58.9312
Fly Ash (component 3)(kg in a m^3 mixture)    54.1883    54.0508    54.7628
Water  (component 4)(kg in a m^3 mixture)    181.5673   180.5809   185.6864
Superplasticizer (component 5)(kg in a m^3 mixture)      6.2047     6.7804     3.8005
Coarse Aggregate  (component 6)(kg in a m^3 mixture)    972.9189   966.4882   999.7729
Fine Aggregate (component 7)(kg in a m^3 mixture)    773.5805   767.4248   799.2859
Age (day)                                    45.6621    53.8676    11.397
Concrete compressive strength(MPa. megapascals)      Média      Média      Má



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

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

Clustered Instances

0       831 ( 81%)
1       199 ( 19%)

Final cluster centroids:
                                                        Cluster#
Attribute                                    Full Data          0          1          2
                                             (1030.0)    (462.0)    (198.0)    (370.0)
=============================================================================================
Cement (component 1)(kg in a m^3 mixture)    281.1679   332.0355   217.1091   251.9322
Blast Furnace Slag (component 2)(kg in a m^3 mixture)     73.8958   103.3939    59.447    44.7951
Fly Ash (component 3)(kg in a m^3 mixture)    54.1883     1.1494    53.9061   120.5665
Water  (component 4)(kg in a m^3 mixture)    181.5673   185.2076   186.001   174.6492
Superplasticizer (component 5)(kg in a m^3 mixture)      6.2047     5.1251     3.7556     8.8632
Coarse Aggregate  (component 6)(kg in a m^3 mixture)    972.9189   967.3147   999.2626   965.8192
Fine Aggregate (component 7)(kg in a m^3 mixture)    773.5805   755.5102   799.2263    782.42
Age (day)                                    45.6621    66.6039    11.0303    38.0459
Concrete compressive strength(MPa. megapascals)      Média      Média      Má      Média



Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       462 ( 45%)
1       198 ( 19%)
2       370 ( 36%)
```

3.4 Associação

3.4.1 Pré-processamento

Não foi necessário qualquer pré processamento nos dados, utilizou-se o dataset de origem, sem qualquer atributo nominal.

3.4.2 Algoritmos testados

Apriori

Output:
Minimum support: 0.1 (103 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 5

Size of set of large itemsets L(4): 1

Best rules found:

1. Water (component 4)(kg in a m^3 mixture)=192,0 Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 116 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 116 <conf:(1)> lift:(1.82) lev:(0.05) [52] conv:(52.26)
2. Fly Ash (component 3)(kg in a m^3 mixture)=0,0 Water (component 4)(kg in a m^3 mixture)=192,0 116 ==> Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 116 <conf:(1)> lift:(2.72) lev:(0.07) [73] conv:(73.32)
3. Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 Age (day)='(-inf-37.4]' 258 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 254 <conf:(0.98)> lift:(1.79) lev:(0.11) [112] conv:(23.25)
4. Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 379 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 373 <conf:(0.98)> lift:(1.79) lev:(0.16) [164] conv:(24.39)
5. Water (component 4)(kg in a m^3 mixture)=192,0 118 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 116 <conf:(0.98)> lift:(1.79) lev:(0.05) [51] conv:(17.72)

6. Water (component 4)(kg in a m^3 mixture)=192,0 118 ==> Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 116 <conf:(0.98)> lift:(2.67) lev:(0.07) [72] conv:(24.86)

7. Water (component 4)(kg in a m^3 mixture)=192,0 118 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 116 <conf:(0.98)> lift:(2.71) lev:(0.07) [73] conv:(25.09)

8. Blast Furnace Slag (component 2)(kg in a m^3 mixture)=0,0 Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 Age (day)='(-inf-37.4]' 151 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 147 <conf:(0.97)> lift:(1.77) lev:(0.06) [64] conv:(13.6)

9. Blast Furnace Slag (component 2)(kg in a m^3 mixture)=0,0 Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 215 ==> Fly Ash (component 3)(kg in a m^3 mixture)=0,0 209 <conf:(0.97)> lift:(1.77) lev:(0.09) [90] conv:(13.84)

10. Blast Furnace Slag (component 2)(kg in a m^3 mixture)=0,0 Fly Ash (component 3)(kg in a m^3 mixture)=0,0 232 ==> Superplasticizer (component 5)(kg in a m^3 mixture)=0,0 209 <conf:(0.9)> lift:(2.45) lev:(0.12) [123] conv:(6.11)

3.5 Resultados

Os melhores algoritmos de classificação foram: Random Florest, Redes Neuronais, J48 e Simple Cart. Todos entre os 90% e 93% de acertos nos testes com cross-validation fold 10 e a utilizar addaboost M1. Neste Problema o recall para a classe "Má" tem importância porque não se quer que a força de resistência de um bloco de betão seja prevista com um falso negativo. Neste capitulo, não há grandes diferenças entre os 4 algoritmos falados neste paragrafo.

Quanto à segmentação, verificou-se que o algoritmo EM dá resultados equivalentes ao k-means, neste caso, no entanto é mais demorado. Este algoritmo permite descobrir o melhor N. Para o k-means fizeram-se 2 testes, ambos com boa distribuição a dividir o dataset em 2 e 3 clusters.

A associação não apresentou bons resultados neste caso, já que não se revelava sequer útil neste contexto. No entanto, permitiu a exploração desta ferramenta.

3.6 Recomendações

Os resultados da utilização das técnicas de extração de conhecimento neste *dataset* são bastante úteis no que toca à classificação.

Assim, podemos inferir um modelo de previsão que nos permite saber a força compressiva (resistência) de um bloco de betão a partir da sua constituição (dada por uma série de variáveis) e idade.

4 *Dataset* EnergyUse

4.1 Objetivos de Estudo e Significado dos Dados

Nos dias correntes, casas, escolas, escritórios e outros edifícios consomem mais energia do que a indústria ou os transportes, tornando-se o conceito de edifício de baixo consumo energético extremamente relevante.

Para tal, dever-se-á promover a melhoria da eficiência energética nos edifícios, cobrindo todos os tipos de consumo, desde a preparação de água quente sanitária, passando pela iluminação e pelos equipamentos e eletrodomésticos, sem esquecer a melhoria da envolvente tendo em conta o impacto desta nos consumos de climatização (aquecimento, arrefecimento e ventilação) para assegurar o conforto ambiente.

Assim, a partir de dados como a temperatura e a humidade relativa numa residência (T1 e RH_1 a T9 e RH_9) é possível extrapolar a decisão do uso de isolantes térmicos, de janelas e portas altamente eficientes (algo que pode implicar custos desnecessários sem esta extração de conhecimento prévia).

Por outro lado, a utilização de um conjunto de dados como a temperatura externa (T_out), a pressão do ar (press_mm_hg), a humidade relativa externa(RH_out), a temperatura do ponto de orvalho (Tdewpoint) e a velocidade do vento (Windspeed) revela-se benéfico na medida em que é possível prever um modelo de gasto energético, sendo extremamente vantajoso na otimização de edifícios de baixa energia.

Deste modo, foram analisados e tratados diferentes tipos de dados, sendo eles os seguintes:

- Eletrodomésticos – uso de energia Wh
- Luzes – uso de energia de lâmpadas na casa em Wh
- T1 – Temperatura na área da cozinha em Celcius
- RH_1 – Humidade na área da cozinha em %
- T2 – Temperatura na sala de estar em Celcius
- RH_2 – Humidade na sala de estar em %
- T3 – Temperatura na área de lavandaria em Celcius
- RH_3 – Humidade na área de lavandaria de estar em %
- T4 – Temperatura no escritório em Celcius
- RH_4 – Humidade no escritório em %
- T5 – Temperatura na casa de banho em Celcius

- RH_5 – Humidade na casa de banho em %
- T6 – Temperatura fora do prédio em Celcius
- RH_6 – Humidade fora do prédio em %
- T7 – Temperatura na sala de engomar em Celcius
- RH_7 – Humidade na sala de engomar em %
- T8 – Temperatura no quarto do adolescente em Celcius
- RH_8 – Humidade no quarto do adolescente em %
- T9 – Temperatura no quarto dos pais em Celcius
- RH_9 – Humidade no quarto dos pais em %
- T_out – Temperatura externa (da estação meteorológica de Chievres), em Celcius
- Press_mm_hg – Pressão (da estação meteorológica de Chievres), em mmHg
- RH_out – Humidade externa (da estação metereológica de Chievres), em %
- Windspeed – Velocidade do vento (da estação metereológica de Chievres), em m/s
- Visibility – Visibilidade (da estação metereológica de Chievres), em km
- Tdewpoint – Temperatura do ponto de orvalho (da estação metereológica de Chievres), em Celcius
- Rv1 – Variável aleatória 1, adimensional
- Rv2 – Variável aleatória 2, adimensional

4.2 Classificação

4.2.1 Pré-processamento

O dataset EnergyUse foi analisado com vista à classificação, sendo efetuado um grande conjunto de testes, optando-se, por fim, pela classificação de um novo atributo "Gasto de eletrodomésticos", obtido pela subtração do atributo *Appliances* (Gasto toal) e *lights* (gasto de luzes). Assim, obtém-se um atributo que está intimamente relacionado às variáveis:

- T_out – Temperatura externa (da estação meteorológica de Chievres), em Celcius
- Press_mm_hg – Pressão (da estação meteorológica de Chievres), em mmHg
- RH_out – Humidade externa (da estação metereológica de Chievres), em %
- Windspeed – Velocidade do vento (da estação metereológica de Chievres), em m/s
- Tdewpoint – Temperatura do ponto de orvalho (da estação metereológica de Chievres), em Celcius

uma vez que estas variáveis levam a uma maior ou menor utilização de dispositivos gerados de calor ou frio.

Assim, observou-se a distribuição do novo atributo de forma a criar 2 *thresholds* e criar 3 classes num novo atributo nominal: Pouca, média ou muita energia gasta. Estas transformações foram geradas no software *Microsoft Excel*.

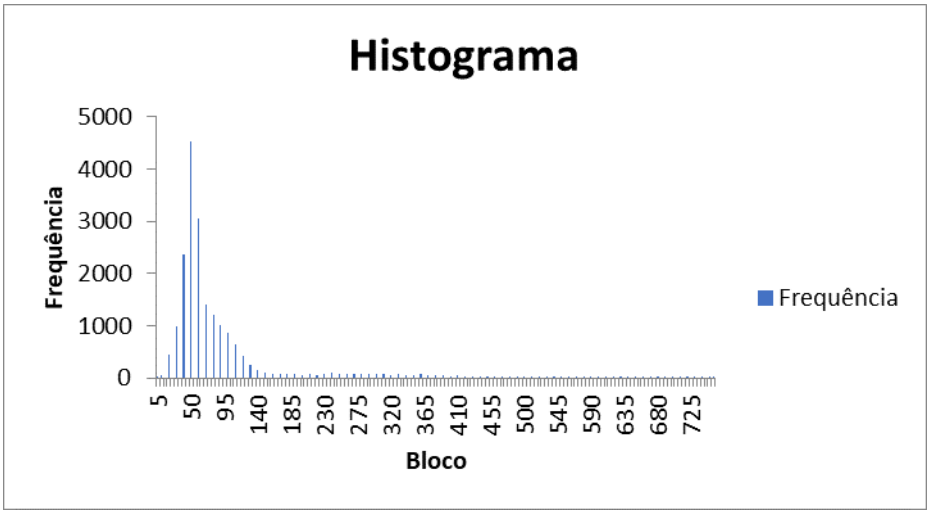


Figura 20 -Distribuição do atributo a prever

4.2.2 Algoritmos testados

Redes Neuronais

```
=== Summary ===

Correctly Classified Instances       15456               78.3177 %
Incorrectly Classified Instances      4279               21.6823 %
Kappa statistic                          0.0442
Mean absolute error                      0.2147
Root mean squared error                  0.3348
Relative absolute error                 91.6223 %
Root relative squared error             97.7994 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,992    0,964    0,786      0,992    0,877      0,101    0,633     0,848     Média
               0,000    0,000    0,000      0,000    0,000      0,000    0,628     0,041     Pouca
               0,041    0,008    0,560      0,041    0,076      0,111    0,657     0,327     Muita
Weighted Avg.  0,783    0,754    0,722      0,783    0,700      0,100    0,637     0,727
```

Figura 21 - Resultados - MultiLayer Perceptron

```
=== Summary ===

Correctly Classified Instances       15456               78.3177 %
Incorrectly Classified Instances      4279               21.6823 %
Kappa statistic                          0.0442
Mean absolute error                      0.2796
Root mean squared error                  0.3486
Relative absolute error                119.3082 %
Root relative squared error            101.852  %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,992    0,964    0,786      0,992    0,877      0,101    0,617     0,837     Média
                 0,000    0,000    0,000      0,000    0,000      0,000    0,395     0,021     Pouca
                 0,041    0,008    0,560      0,041    0,076      0,111    0,649     0,307     Muita
Weighted Avg.    0,783    0,754    0,722      0,783    0,700      0,100    0,617     0,714
```

Figura 22 – Resultados – MultiLayer Perceptron com adaboostM1

Support Vector Machine

```
=== Summary ===

Correctly Classified Instances       15421               78.1404 %
Incorrectly Classified Instances      4314               21.8596 %
Kappa statistic                          0
Mean absolute error                      0.2765
Root mean squared error                  0.3582
Relative absolute error                117.9697 %
Root relative squared error            104.6583 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,781      1,000    0,877      0,000    0,500     0,781     Média
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,026     Pouca
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,193     Muita
Weighted Avg.    0,781    0,781    0,611      0,781    0,686      0,000    0,500     0,648
```

Figura 23 - Resultados - SMO

```
=== Summary ===

Correctly Classified Instances       15421               78.1404 %
Incorrectly Classified Instances      4314               21.8596 %
Kappa statistic                          0
Mean absolute error                      0.2765
Root mean squared error                  0.3582
Relative absolute error                117.9697 %
Root relative squared error            104.6583 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,781      1,000    0,877      0,000    0,500     0,781     Média
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,026     Pouca
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,193     Muita
Weighted Avg.    0,781    0,781    0,611      0,781    0,686      0,000    0,500     0,648
```

Figura 24 - Resultados - SMO com AdaBoostM1

NaiveBayes

```
=== Summary ===

Correctly Classified Instances       15149               76.7621 %
Incorrectly Classified Instances      4586               23.2379 %
Kappa statistic                          0.0558
Mean absolute error                      0.2314
Root mean squared error                  0.343
Relative absolute error                 98.7306 %
Root relative squared error            100.1981 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,961    0,924    0,788      0,961    0,866      0,072    0,613     0,840     Média
                 0,000    0,000    0,000      0,000    0,000      0,000    0,676     0,043     Pouca
                 0,085    0,038    0,352      0,085    0,137      0,089    0,640     0,278     Muita
Weighted Avg.    0,768    0,730    0,684      0,768    0,703      0,074    0,620     0,712
```

Figura 25 - Resultados - Naive Bayes

```
=== Summary ===

Correctly Classified Instances       15149               76.7621 %
Incorrectly Classified Instances      4586               23.2379 %
Kappa statistic                          0.0558
Mean absolute error                      0.2916
Root mean squared error                  0.3576
Relative absolute error                124.4045 %
Root relative squared error            104.485  %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,961    0,924    0,788      0,961    0,866      0,072    0,572     0,810     Média
                 0,000    0,000    0,000      0,000    0,000      0,000    0,477     0,025     Pouca
                 0,085    0,038    0,352      0,085    0,137      0,089    0,585     0,256     Muita
Weighted Avg.    0,768    0,730    0,684      0,768    0,703      0,074    0,572     0,683
```

Figura 26 - Resultados - Naive Bayes com AdaBoostM1

BayesNet

```
=== Summary ===

Correctly Classified Instances       15471               78.3937 %
Incorrectly Classified Instances      4264               21.6063 %
Kappa statistic                          0.1005
Mean absolute error                      0.2233
Root mean squared error                  0.3348
Relative absolute error                 95.2584 %
Root relative squared error             97.8145 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,978    0,910    0,794      0,978    0,876      0,151    0,635     0,850     Média
                 0,000    0,000    0,000      0,000    0,000      -0,001   0,717     0,053     Pouca
                 0,101    0,021    0,531      0,101    0,169      0,167    0,668     0,346     Muita
Weighted Avg.    0,784    0,715    0,723      0,784    0,717      0,150    0,644     0,733
```

Figura 27 - Resultados - BayesNet

```
=== Summary ===

Correctly Classified Instances       15471               78.3937 %
Incorrectly Classified Instances      4264               21.6063 %
Kappa statistic                          0.1005
Mean absolute error                      0.2755
Root mean squared error                  0.3469
Relative absolute error                117.5529 %
Root relative squared error            101.343  %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,978    0,910    0,794      0,978    0,876      0,151    0,626     0,835     Média
                 0,000    0,000    0,000      0,000    0,000      -0,001   0,465     0,028     Pouca
                 0,101    0,021    0,531      0,101    0,169      0,167    0,652     0,330     Muita
Weighted Avg.    0,784    0,715    0,723      0,784    0,717      0,150    0,627     0,717
```

Figura 28 - Resultados - BayesNet com AdaBoostM1

Árvores de Decisão – J48 (C4.5)

```
=== Summary ===

Correctly Classified Instances       16377               82.9845 %
Incorrectly Classified Instances      3358               17.0155 %
Kappa statistic                          0.452
Mean absolute error                      0.1501
Root mean squared error                  0.308
Relative absolute error                 64.0358 %
Root relative squared error             89.9701 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,933    0,532    0,862      0,933    0,896      0,458    0,772     0,897     Média
                 0,061    0,004    0,307      0,061    0,102      0,128    0,746     0,095     Pouca
                 0,515    0,062    0,664      0,515    0,580      0,501    0,813     0,564     Muita
Weighted Avg.    0,830    0,428    0,810      0,830    0,815      0,457    0,779     0,812
```

Figura 29 - Resultados - J48

```
=== Summary ===

Correctly Classified Instances       16997               86.1262 %
Incorrectly Classified Instances      2738               13.8738 %
Kappa statistic                          0.5761
Mean absolute error                      0.0931
Root mean squared error                  0.2895
Relative absolute error                 39.7348 %
Root relative squared error             84.5623 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,936    0,402    0,893      0,936    0,914      0,573    0,870     0,948     Média
                 0,178    0,009    0,352      0,178    0,237      0,237    0,729     0,182     Pouca
                 0,651    0,053    0,747      0,651    0,696      0,632    0,911     0,751     Muita
Weighted Avg.    0,861    0,324    0,851      0,861    0,854      0,575    0,874     0,890
```

Figura 30 - Resultados - J48 com AdaBoostM1

Árvores de Decisão – *Random Florest*

```
=== Summary ===

Correctly Classified Instances       17222               87.2663 %
Incorrectly Classified Instances      2513               12.7337 %
Kappa statistic                          0.5944
Mean absolute error                      0.128
Root mean squared error                  0.2505
Relative absolute error                 54.5996 %
Root relative squared error             73.1732 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,956    0,425    0,889      0,956    0,922      0,599    0,894     0,956     Média
                 0,123    0,005    0,413      0,123    0,189      0,215    0,812     0,238     Pouca
                 0,634    0,037    0,804      0,634    0,709      0,656    0,933     0,797     Muita
Weighted Avg.    0,873    0,340    0,861      0,873    0,862      0,600    0,900     0,907
```

Figura 31 - Resultados - Random Florest

Árvores de Decisão – *Simple Cart*

```
=== Summary ===

Correctly Classified Instances       16545               83.8358 %
Incorrectly Classified Instances      3190               16.1642 %
Kappa statistic                          0.4839
Mean absolute error                      0.1447
Root mean squared error                  0.3059
Relative absolute error                 61.7304 %
Root relative squared error             89.3638 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,935    0,503    0,869      0,935    0,901      0,488    0,765     0,891     Média
               0,036    0,003    0,237      0,036    0,062      0,083    0,747     0,077     Pouca
               0,554    0,061    0,687      0,554    0,613      0,537    0,807     0,572     Muita
Weighted Avg.  0,838    0,404    0,818      0,838    0,824      0,487    0,773     0,808
```

Figura 32 - Resultados - Simple Cart

```
=== Summary ===

Correctly Classified Instances       16959               85.9336 %
Incorrectly Classified Instances      2776               14.0664 %
Kappa statistic                          0.5783
Mean absolute error                      0.0936
Root mean squared error                  0.2882
Relative absolute error                 39.9481 %
Root relative squared error             84.2042 %
Total Number of Instances            19735

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,928    0,385    0,896      0,928    0,912      0,572    0,872     0,948     Média
               0,208    0,012    0,320      0,208    0,252      0,242    0,750     0,203     Pouca
               0,667    0,056    0,740      0,667    0,702      0,636    0,910     0,746     Muita
Weighted Avg.  0,859    0,312    0,851      0,859    0,854      0,576    0,876     0,890
```

Figura 33 - Resultados - Simple Cart com AdaBoostM1

Iremos sumariar estes resultados na penúltima secção deste dataset.

4.3 Segmentação

4.3.1 Pré-processamento

O pré processamento utilizado foi a utilização do atributo data para a geração de um novo atributo "dia da semana". Consequentemente eliminou-se a variável data. As variáveis rv1 e rv2 também foram eliminados por não fazerem sentido neste problema.

4.3.2 Algoritmos testados

EM

```
Clustered Instances

0       449 ( 44%)
1       387 ( 38%)
2       194 ( 19%)
```

Figura 34: Resultados - EM

Hierarchical Clusterer

```
Clustered Instances     Clustered Instances

0      2061 ( 93%)      0      1629 ( 74%)
1       144 (  7%)      1       144 (  7%)
                        2       432 ( 20%)

Clustered Instances     Clustered Instances

0      1629 ( 74%)      0      1341 ( 61%)
1       144 (  7%)      1       288 ( 13%)
2       144 (  7%)      2       144 (  7%)
3       288 ( 13%)      3       144 (  7%)
                        4       288 ( 13%)
```

Figura 35 - Resultados Para Vários N's - Hierarchical Clusterer

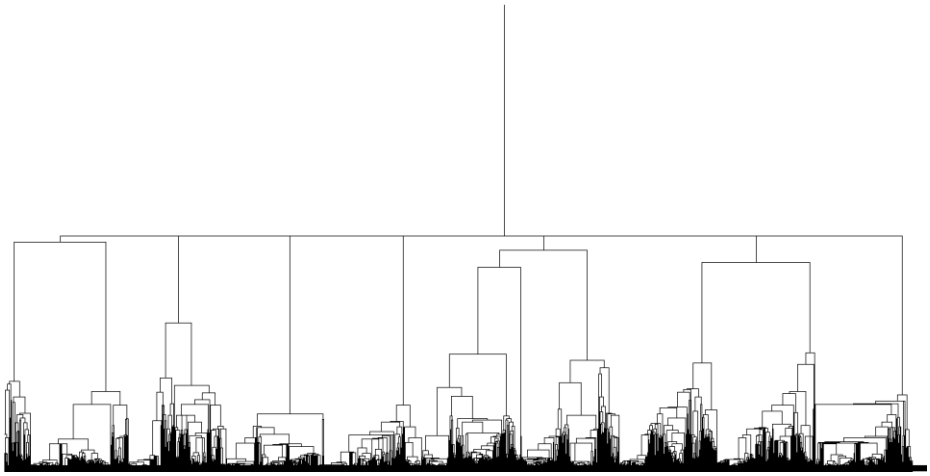


Figura 36 - Dendograma para N=5

Simple K-Means

```
Clustered Instances

0       11028  ( 56%)
1        8707  ( 44%)
```

```
Clustered Instances

0        5458  ( 28%)
1        8396  ( 43%)
2        5881  ( 30%)
```

```
Clustered Instances

0        6798  ( 34%)
1        5493  ( 28%)
2        3261  ( 17%)
3        4183  ( 21%)
```

```
Clustered Instances

0        4170  ( 21%)
1        3976  ( 20%)
2        2499  ( 13%)
3        3596  ( 18%)
4        5494  ( 28%)
```

Figura 37: Resultados Para Vários N's - Simple K-Means

4.4 Associação

4.4.1 Pré-processamento

O pré processamento utilizado foi a utilização do atributo data para a geração de um novo atributo "dia da semana". Consequentemente eliminou-se a variável data. As variáveis rv1 e rv2 também foram eliminados por não fazerem sentido neste problema. De seguida utilizou-se o filtro *descretize*.

4.4.2 Algoritmos testados

Apriori

Minimum support: 0.1 (1974 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 98

Size of set of large itemsets L(2): 229

Size of set of large itemsets L(3): 133

Size of set of large itemsets L(4): 13

Best rules found:

1. lights='(-inf-7]' T6='(0.806-4.2415]' 2864 ==> Appliances='(-inf-117]' 2662 <conf:(0.93)> lift:(1.13) lev:(0.02) [305] conv:(2.5)
2. 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)
3. lights='(-inf-7]' T2='(17.475667-18.851333]' 3122 ==> Appliances='(-inf-117]' 2900 <conf:(0.93)> lift:(1.13) lev:(0.02) [331] conv:(2.48)
4. lights='(-inf-7]' T6='(0.806-4.2415]' T_out='(1.22-4.33]' 2143 ==> Appliances='(-inf-117]' 1976 <conf:(0.92)> lift:(1.12) lev:(0.01) [212] conv:(2.26)
5. lights='(-inf-7]' RH_1='(37.924333-41.558]' RH_4='(37.032-39.375]' 2656 ==> Appliances='(-inf-117]' 2432 <conf:(0.92)> lift:(1.11) lev:(0.01) [246] conv:(2.09)
6. 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)

7. lights='(-inf-7]' T3='(19.6072-20.8108]' 2983 ==> Appliances='(-inf-117]' 2725 <conf:(0.91)> lift:(1.11) lev:(0.01) [270] conv:(2.04)

8. RH_6='(50.45-60.34]' 2456 ==> Appliances='(-inf-117]' 2236 <conf:(0.91)> lift:(1.11) lev:(0.01) [215] conv:(1.97)

9. lights='(-inf-7]' RH_6='(90.01-inf)' 2213 ==> Appliances='(-inf-117]' 2012 <conf:(0.91)> lift:(1.1) lev:(0.01) [191] conv:(1.94)

10. lights='(-inf-7]' Windspeed='(1.4-2.8]' 3278 ==> Appliances='(-inf-117]' 2966 <conf:(0.9)> lift:(1.1) lev:(0.01) [268] conv:(1.86)

4.5 Resultados

Os melhores algoritmos de classificação foram de longe árvores de decisão: Random Florest, J48 e Simple Cart, entre os 86% e 87% de acertos nos testes com cross-validation fold 10 todos exceto Random Florest a utilizar addaboost M1.

Quanto à segmentação, verificou-se que o algoritmo EM dá os resultados melhor distribuidos, no entanto é mais demorado. Este algoritmo permite descobrir o melhor N. Para o k-means fizeram-se 4 testes, nos quais a melhor distribuição se mostrou para 2 e 3 clusters.

A associação apresentou bons resultados, no entanto óbvios que não acrescentam informação útil, por exemplo a regra:

. lights='(-inf-7]' T6='(0.806-4.2415]' 2864 ==> Appliances='(-inf-117]' 2662 <conf:(0.93)> lift:(1.13) lev:(0.02) [305] conv:(2.5)

que nos diz que as o consumo de energia em luzes está relacionado com o consumo total de energia com grau de confiança 0.93 e suporte de 0.1 (número mínimo de casos de dataset que suportam a regra).

4.6 Recomendações

Os resultados da utilização das técnicas de extração de conhecimento neste *dataset* são bastante úteis no que toca à classificação.

Assim, podemos inferir um modelo de previsão que nos permite saber a quantidade de energia gasta em função das variáveis temperatura exterior, ponto de orvalho, vento, pressão e humidade.

5 *Dataset* Student Performance (Português)

5.1 Objetivos de Estudo e Significado dos Dados

Este *dataset* refere-se à *performance* em testes de Português de alunos de duas escolas, cujos atributos correspondem, não só aos parâmetros que influenciam diretamente o aproveitamento do aluno (por exemplo, tempo de estudo), mas também às condições socioeconómicas (meio onde vive, atividades, características da família, *etc*) que poderão representar uma influência indireta.

Uma classificação neste *dataset* é particularmente útil para descobrir em quais circunstâncias um aluno consegue a aprovação ou acaba reprovado.

5.2 Classificação

5.2.1 Pré-processamento

Estabeleceu-se um atributo *passed* que representa se o aluno obteve uma nota final (G3) maior ou igual a 10. Estabelecemos que uma classificação binária (*yes* ou *no*) seria mais adequada do que uma de 0 a 20, pois esta última faria com que a previsão tentasse adivinhar um valor demasiado específico e incorreria em erros maiores.

Como G3 depende diretamente de G1 e G2 e *passed* de G3, estes são removidos do *dataset*.

Numa das classificações foi também usado um filtro com uso de algoritmos genéticos para gerar atributos apropriados para Árvores de Decisão J48.

5.2.2 Algoritmos testados

Redes Neuronais

```
=== Summary ===

Correctly Classified Instances         535               82.4345 %
Incorrectly Classified Instances       114               17.5655 %
Kappa statistic                          0.2974
Mean absolute error                      0.1727
Root mean squared error                  0.378
Relative absolute error                 66.0315 %
Root relative squared error            104.7111 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.905    0.620    0.889      0.905    0.897      0.298    0.790     0.940     Yes
               0.380    0.095    0.422      0.380    0.400      0.298    0.790     0.443     No
Weighted Avg.  0.824    0.539    0.817      0.824    0.821      0.298    0.790     0.864
```

Figura 38 - Resultados - MultiLayer Perceptron

```
=== Summary ===

Correctly Classified Instances         536               82.5886 %
Incorrectly Classified Instances       113               17.4114 %
Kappa statistic                          0.2693
Mean absolute error                      0.1761
Root mean squared error                  0.4074
Relative absolute error                 67.3478 %
Root relative squared error            112.8348 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.916    0.670    0.882      0.916    0.899      0.272    0.689     0.909     Yes
                 0.330    0.084    0.418      0.330    0.369      0.272    0.689     0.328     No
Weighted Avg.    0.826    0.580    0.811      0.826    0.817      0.272    0.689     0.819
```

Figura 39 - Resultados - Mulr Perceptron com adaboostM1

Support Vector Machine

```
=== Summary ===

Correctly Classified Instances         560               86.2866 %
Incorrectly Classified Instances        89               13.7134 %
Kappa statistic                          0.3141
Mean absolute error                      0.1371
Root mean squared error                  0.3703
Relative absolute error                 52.4413 %
Root relative squared error            102.572  %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.971    0.730    0.880      0.971    0.923      0.350    0.620     0.879     Yes
                 0.270    0.029    0.628      0.270    0.378      0.350    0.620     0.282     No
Weighted Avg.    0.863    0.622    0.841      0.863    0.839      0.350    0.620     0.787
```

Figura 40 - Resultados - SMO

```
=== Summary ===

Correctly Classified Instances         547               84.2835 %
Incorrectly Classified Instances       102               15.7165 %
Kappa statistic                          0.2862
Mean absolute error                      0.1956
Root mean squared error                  0.3567
Relative absolute error                 74.8101 %
Root relative squared error             98.8142 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.700    0.881      0.942    0.910      0.297    0.735     0.917     Yes
                 0.300    0.058    0.484      0.300    0.370      0.297    0.735     0.384     No
Weighted Avg.    0.843    0.601    0.820      0.843    0.827      0.297    0.735     0.835
```

Figura 41 - Resultados - SMO COM AdaBoostM1

NaiveBayes

```
=== Summary ===

Correctly Classified Instances         537               82.7427 %
Incorrectly Classified Instances       112               17.2573 %
Kappa statistic                          0.3538
Mean absolute error                      0.1915
Root mean squared error                  0.3687
Relative absolute error                 73.247  %
Root relative squared error            102.1378 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.893    0.530    0.902      0.893    0.897      0.354    0.813     0.949     Yes
               0.470    0.107    0.443      0.470    0.456      0.354    0.813     0.438     No
Weighted Avg.  0.827    0.465    0.832      0.827    0.829      0.354    0.813     0.871
```

Figura 42 - Resultados - Naive Bayes

```
=== Summary ===

Correctly Classified Instances         535               82.4345 %
Incorrectly Classified Instances       114               17.5655 %
Kappa statistic                          0.2853
Mean absolute error                      0.2024
Root mean squared error                  0.3682
Relative absolute error                 77.4005 %
Root relative squared error            101.9799 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.909    0.640    0.886      0.909    0.897      0.286    0.763     0.934     Yes
               0.360    0.091    0.419      0.360    0.387      0.286    0.763     0.371     No
Weighted Avg.  0.824    0.555    0.814      0.824    0.819      0.286    0.763     0.848
```

Figura 43 - Resultados - Naive Bayes com AdaBoostM1

BayesNet

```
=== Summary ===

Correctly Classified Instances         553               85.208  %
Incorrectly Classified Instances        96               14.792  %
Kappa statistic                          0.4084
Mean absolute error                      0.1931
Root mean squared error                  0.3387
Relative absolute error                 73.8431 %
Root relative squared error             93.8242 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.922    0.530    0.905      0.922    0.913      0.409    0.803     0.947     Yes
               0.470    0.078    0.522      0.470    0.495      0.409    0.803     0.439     No
Weighted Avg.  0.852    0.460    0.846      0.852    0.849      0.409    0.803     0.869
```

Figura 44 - Resultados – BayesNet

```
=== Summary ===

Correctly Classified Instances         550               84.7458 %
Incorrectly Classified Instances        99               15.2542 %
Kappa statistic                          0.3873
Mean absolute error                      0.2004
Root mean squared error                  0.3556
Relative absolute error                 76.6393 %
Root relative squared error             98.4944 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.550    0.902      0.920    0.911      0.388    0.701     0.900     Yes
                 0.450    0.080    0.506      0.450    0.476      0.388    0.701     0.373     No
Weighted Avg.    0.847    0.478    0.841      0.847    0.844      0.388    0.701     0.819
```

Figura 45 - Resultados - BayesNet com AdaBoostM1

Árvores de Decisão – J48 (C4.5)

```
=== Summary ===

Correctly Classified Instances         549               84.5917 %
Incorrectly Classified Instances       100               15.4083 %
Kappa statistic                          0.3199
Mean absolute error                      0.1968
Root mean squared error                  0.362
Relative absolute error                 75.2519 %
Root relative squared error            100.2789 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.938    0.660    0.886      0.938    0.912      0.328    0.668     0.877     Yes
                 0.340    0.062    0.500      0.340    0.405      0.328    0.668     0.376     No
Weighted Avg.    0.846    0.568    0.827      0.846    0.833      0.328    0.668     0.799
```

Figura 46 - Resultados - J48

```
=== Summary ===

Correctly Classified Instances         540               83.2049 %
Incorrectly Classified Instances       109               16.7951 %
Kappa statistic                          0.2824
Mean absolute error                      0.159
Root mean squared error                  0.3784
Relative absolute error                 60.7909 %
Root relative squared error            104.8249 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.670    0.883      0.923    0.903      0.286    0.796     0.945     Yes
                 0.330    0.077    0.440      0.330    0.377      0.286    0.796     0.417     No
Weighted Avg.    0.832    0.579    0.815      0.832    0.822      0.286    0.796     0.864
```

Figura 47 - Resultados - J48 com AdaBoostM1

```
=== Summary ===

Correctly Classified Instances         565               87.057  %
Incorrectly Classified Instances        84               12.943  %
Kappa statistic                          0.4393
Mean absolute error                      0.1799
Root mean squared error                  0.3468
Relative absolute error                 68.8061 %
Root relative squared error             96.0527 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.949    0.560    0.903      0.949    0.925      0.447    0.679     0.883     Yes
               0.440    0.051    0.611      0.440    0.512      0.447    0.679     0.381     No
Weighted Avg.  0.871    0.482    0.858      0.871    0.862      0.447    0.679     0.805
```

Figura 48 - Resultados - J48 com filtro GPAttributeGeneration

Árvores de Decisão – *Random Florest*

```
=== Summary ===

Correctly Classified Instances         553               85.208  %
Incorrectly Classified Instances        96               14.792  %
Kappa statistic                          0.214
Mean absolute error                      0.2141
Root mean squared error                  0.3225
Relative absolute error                 81.8795 %
Root relative squared error             89.341  %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.974    0.820    0.867      0.974    0.918      0.258    0.821     0.957     Yes
               0.180    0.026    0.563      0.180    0.273      0.258    0.821     0.459     No
Weighted Avg.  0.852    0.698    0.820      0.852    0.818      0.258    0.821     0.880
```

Figura 49 - Resultados – RandomForest

Árvores de Decisão – *Simple Cart*

```
=== Summary ===

Correctly Classified Instances         547               84.2835 %
Incorrectly Classified Instances       102               15.7165 %
Kappa statistic                          0.2262
Mean absolute error                      0.2269
Root mean squared error                  0.3487
Relative absolute error                 86.7679 %
Root relative squared error             96.5953 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.956     0.780     0.871      0.956     0.911      0.248    0.645     0.886     Yes
               0.220     0.044     0.478      0.220     0.301      0.248    0.645     0.292     No
Weighted Avg.  0.843     0.667     0.810      0.843     0.817      0.248    0.645     0.795
```

Figura 50 - Resultados - Simple Cart

5.3 Segmentação

5.3.1 Pré-processamento

O pré processamento utilizado foi semelhante ao da classificação, ou seja, a criação do atributo *passed* e a remoção dos atributos que possuem uma relação direta com o mesmo.

5.3.2 Algoritmos testados

EM

```
Clustered Instances

0       145 ( 22%)
1       188 ( 29%)
2       300 ( 46%)
3        16 (  2%)
```

Simple K-Means

Attribute	Full Data (649.0)	Cluster# 0 (374.0)	1 (275.0)
school	GP	GP	GP
sex	F	M	F
age	'(16.4-17.1]'	'(15.7-16.4]'	'(16.4-17.1]'
address	U	U	U
famsize	GT3	GT3	GT3
Pstatus	T	T	T
Medu	'(1.6-2]'	'(1.6-2]'	'(3.6-inf)'
Fedu	'(1.6-2]'	'(1.6-2]'	'(0.8-1.2]'
Mjob	other	other	other
Fjob	other	other	other
reason	course	course	course
guardian	mother	mother	mother
traveltime	'(-inf-1.3]'	'(-inf-1.3]'	'(-inf-1.3]'
studytime	'(1.9-2.2]'	'(-inf-1.3]'	'(1.9-2.2]'
failures	'(-inf-0.3]'	'(-inf-0.3]'	'(-inf-0.3]'
schoolsup	no	no	no
famsup	yes	yes	yes
paid	no	no	no
activities	no	no	yes
nursery	yes	yes	yes
higher	yes	yes	yes
internet	yes	yes	yes
romantic	no	no	no
famrel	'(3.8-4.2]'	'(3.8-4.2]'	'(3.8-4.2]'
freetime	'(2.6-3]'	'(2.6-3]'	'(3.8-4.2]'
goout	'(2.6-3]'	'(2.6-3]'	'(3.8-4.2]'
Dalc	'(-inf-1.4]'	'(-inf-1.4]'	'(-inf-1.4]'
Walc	'(-inf-1.4]'	'(-inf-1.4]'	'(-inf-1.4]'
health	'(4.6-inf)'	'(4.6-inf)'	'(4.6-inf)'
absences	'(-inf-3.2]'	'(-inf-3.2]'	'(-inf-3.2]'
passed	Yes	Yes	Yes

Figura 51 - dados distribuidos por dois *clusters*

Clustered Instances		Clustered Instances	
0	374 (58%)	0	196 (30%)
1	275 (42%)	1	185 (29%)
		2	268 (41%)

Clustered Instances		Clustered Instances	
0	174 (27%)	0	161 (25%)
1	143 (22%)	1	134 (21%)
2	231 (36%)	2	215 (33%)
3	101 (16%)	3	76 (12%)
		4	63 (10%)

Figura 52 - Resultados Para Vários N's - Simple K-Means

5.4 Associação

5.4.1 Pré-processamento

O pré processamento utilizado foi semelhante ao da classificação e da segmentação, ou seja, a criação do atributo *passed* e a remoção dos atributos que possuem uma relação direta com o mesmo.

5.4.2 Algoritmos testados

Apriori

Minimum support: 0.75 (487 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 14

Best rules found:

1. failures=0 549 ==> paid=no 520 <conf:(0.95)> lift:(1.01) lev:(0.01) [3] conv:(1.1)
2. passed=Yes 549 ==> paid=no 519 <conf:(0.95)> lift:(1.01) lev:(0) [2] conv:(1.06)
3. schoolsup=no 581 ==> paid=no 548 <conf:(0.94)> lift:(1) lev:(0) [1] conv:(1.03)
4. Pstatus=T 569 ==> paid=no 534 <conf:(0.94)> lift:(1) lev:(-0) [0] conv:(0.95)
5. higher=yes 580 ==> paid=no 544 <conf:(0.94)> lift:(1) lev:(-0) [-1] conv:(0.94)
6. nursery=yes 521 ==> paid=no 488 <conf:(0.94)> lift:(1) lev:(-0) [-1] conv:(0.92)
7. failures=0 549 ==> higher=yes 513 <conf:(0.93)> lift:(1.05) lev:(0.03) [22] conv:(1.58)
8. passed=Yes 549 ==> higher=yes 513 <conf:(0.93)> lift:(1.05) lev:(0.03) [22] conv:(1.58)
9. passed=Yes 549 ==> failures=0 498 <conf:(0.91)> lift:(1.07) lev:(0.05) [33] conv:(1.63)
10. failures=0 549 ==> passed=Yes 498 <conf:(0.91)> lift:(1.07) lev:(0.05) [33] conv:(1.63)

5.5 Resultados

O algoritmo de classificação com o qual se obteve melhores resultados foi J48 com filtro de algoritmos genéticos com 87% de testes acertados. Esta prestação, embora não sendo perfeita, representa uma forma relativamente precisa de prever o aproveitamento de um aluno a partir das características e fatores relacionados com a vida do mesmo.

Quanto à segmentação, os melhores resultados em termos de distribuição dos dados pelos *clusters* verificaram-se com o algoritmo *Simple K-Means* com dois *clusters.*

A associação gerou regras úteis com grande valor de confiança, sendo as duas melhores:

- A regra 1 que define que, se um aluno não falhou noutras disciplinas no passado (*failures*=0), então nunca pagou aulas extra de apoio (*paid=no*)
- A regra 2 que define que, se um aluno passou, então nunca teve aulas de apoio extra.

5.6 Recomendações

Os resultados da utilização das técnicas de extração de conhecimento neste *dataset* são bastante úteis para possivelmente mudar hábitos/estratégias de estudo dos alunos.

As regras de associação, particularmente, permitem observar o tipo de alunos que pagam aulas de apoio que, neste caso, correspondem a alunos com dificuldades e que a prestação dos mesmos não melhora com as aulas de apoio (todos os alunos que tiveram aulas de apoio extra não passaram.

Poder-se-ia recomendar, portanto, por em prática esta informação de modo a melhorar a prestação dos alunos das escolas em questão.

6 *Dataset* Forest Fires

6.1 Objetivos de Estudo e Significado dos Dados

Os fogos florestais são acontecimentos cujos fatores de desencadeamento podem ser medidos de modo a avaliar o quão provável é o aparecimento de um.

Entre os fatores encontram-se índices de humidade, temperatura, vento, chuva, mês e dia da semana.

- X – valor x da posição a que se refere a medição
- Y – valor y da posição a que se refere a medição
- month – mês do ano
- day – dia da semana
- FFMC, DMC, DC, ISI, RH – índices de risco de incêndio
- Temp – temperatura
- wind – velocidade do vento
- rain – volume de chuva
- area – área ardida

6.2 Classificação

6.2.1 Pré-processamento

Foi criado um atributo *fire* que se refere a ter havido ou não (*yes* ou *no*) incêndio, ou seja, se a área ardida é maior do que zero e outro que representa se choveu ou não (*rain>0*).

Para além disto, foram retirados dois atributos que se relacionam com a posição à qual a medição se refere (*X* e *Y*), pois não são úteis para perceber se um fogo ocorrerá ou não no parque.

6.2.2 Algoritmos testados

Inicialmente tinha-se em mente o uso atributo *fire* como classe para obter uma previsão do desencadeamento de fogos com base nos fatores representados pelos atributos, que seria a classificação mais útil para o caso de estudo. No entanto, após diversos testes, o melhor resultado obtido apenas acertou 52% dos testes. Isto significa que a classificação não teve sucesso e que a probabilidade de conseguir prever um incêndio será quase aleatória, não sendo, portanto, útil.

```
=== Summary ===

Correctly Classified Instances         283               54.7389 %
Incorrectly Classified Instances       234               45.2611 %
Kappa statistic                          0.0392
Mean absolute error                      0.4883
Root mean squared error                  0.4941
Relative absolute error                 98.0024 %
Root relative squared error             98.998  %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.963    0.539      1.000    0.701      0.141    0.508     0.533     No
                 0.037    0.000    1.000      0.037    0.071      0.141    0.508     0.497     Yes
Weighted Avg.    0.547    0.510    0.756      0.547    0.405      0.141    0.508     0.516
```

Figura 53 - Melhor classificação encontrada para classe *fire*

Escolhemos como alternativa o atributo *rain* como classe para as classificações que se seguem.

Support Vector Machine

```
=== Summary ===

Correctly Classified Instances         511               98.8395 %
Incorrectly Classified Instances         6                1.1605 %
Kappa statistic                          0.5657
Mean absolute error                      0.0116
Root mean squared error                  0.1077
Relative absolute error                 35.7252 %
Root relative squared error             87.2294 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.500    0.992      0.996    0.994      0.572    0.748     0.992     No
                 0.500    0.004    0.667      0.500    0.571      0.572    0.748     0.341     Yes
Weighted Avg.    0.988    0.492    0.987      0.988    0.988      0.572    0.748     0.982
```

Figura 54 - Resultados – SMO

```
=== Summary ===

Correctly Classified Instances         511               98.8395 %
Incorrectly Classified Instances         6                1.1605 %
Kappa statistic                          0.5657
Mean absolute error                      0.0116
Root mean squared error                  0.1077
Relative absolute error                 35.7252 %
Root relative squared error             87.2294 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.500    0.992      0.996    0.994      0.572    0.748     0.992     No
                 0.500    0.004    0.667      0.500    0.571      0.572    0.748     0.341     Yes
Weighted Avg.    0.988    0.492    0.987      0.988    0.988      0.572    0.748     0.982
```

Figura 55 - Resultados - SMO com AdaBoostM1

NaiveBayes

```
=== Summary ===

Correctly Classified Instances         472               91.2959 %
Incorrectly Classified Instances        45                8.7041 %
Kappa statistic                          0.1596
Mean absolute error                      0.1047
Root mean squared error                  0.2484
Relative absolute error                322.3449 %
Root relative squared error            201.095  %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.917    0.375    0.994      0.917    0.954      0.233    0.817     0.996     No
               0.625    0.083    0.106      0.625    0.182      0.233    0.817     0.332     Yes
Weighted Avg.  0.913    0.370    0.980      0.913    0.942      0.233    0.817     0.986
```

Figura 56 - Resultados - Naive Bayes

```
=== Summary ===

Correctly Classified Instances         470               90.9091 %
Incorrectly Classified Instances        47                9.0909 %
Kappa statistic                          0.1222
Mean absolute error                      0.088
Root mean squared error                  0.2871
Relative absolute error                270.7629 %
Root relative squared error            232.4689 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.916    0.500    0.991      0.916    0.952      0.178    0.836     0.997     No
               0.500    0.084    0.085      0.500    0.145      0.178    0.838     0.095     Yes
Weighted Avg.  0.909    0.494    0.977      0.909    0.940      0.178    0.836     0.983
```

Figura 57 - Resultados - Naive Bayes com AdaBoostM1

BayesNet

```
=== Summary ===

Correctly Classified Instances         452               87.4275 %
Incorrectly Classified Instances        65               12.5725 %
Kappa statistic                          0.1087
Mean absolute error                      0.1394
Root mean squared error                  0.3114
Relative absolute error                429.2671 %
Root relative squared error            252.1155 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.878    0.375    0.993      0.878    0.932      0.185    0.821     0.996     No
               0.625    0.122    0.075      0.625    0.133      0.185    0.821     0.381     Yes
Weighted Avg.  0.874    0.371    0.979      0.874    0.920      0.185    0.821     0.987
```

Figura 58 - Resultados - BayesNet

```
=== Summary ===

Correctly Classified Instances         469               90.7157 %
Incorrectly Classified Instances        48                9.2843 %
Kappa statistic                          0.087
Mean absolute error                      0.0942
Root mean squared error                  0.2957
Relative absolute error                289.8496 %
Root relative squared error            239.4215 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.916    0.625    0.989      0.916    0.951      0.126    0.693     0.990     No
               0.375    0.084    0.065      0.375    0.111      0.126    0.692     0.062     Yes
Weighted Avg.  0.907    0.617    0.975      0.907    0.938      0.126    0.693     0.976
```

Figura 59 - Resultados - BayesNet com AdaBoostM1

Árvores de Decisão – J48 (C4.5)

```
=== Summary ===

Correctly Classified Instances         509               98.4526 %
Incorrectly Classified Instances         8                1.5474 %
Kappa statistic                          0
Mean absolute error                      0.0305
Root mean squared error                  0.1235
Relative absolute error                 93.8141 %
Root relative squared error             99.9863 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.985      1.000    0.992      0.000    0.399     0.981     No
                 0.000    0.000    0.000      0.000    0.000      0.000    0.399     0.015     Yes
Weighted Avg.    0.985    0.985    0.969      0.985    0.977      0.000    0.399     0.966
```

Figura 60 - Resultados - J48

```
=== Summary ===

Correctly Classified Instances         510               98.646  %
Incorrectly Classified Instances         7                1.354  %
Kappa statistic                          0.3582
Mean absolute error                      0.0134
Root mean squared error                  0.1152
Relative absolute error                 41.2518 %
Root relative squared error             93.2598 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.750    0.988      0.998    0.993      0.403    0.756     0.993     No
                 0.250    0.002    0.667      0.250    0.364      0.403    0.855     0.346     Yes
Weighted Avg.    0.986    0.738    0.983      0.986    0.983      0.403    0.758     0.983
```

Figura 61 - Resultados - J48 com AdaBoostM1

Árvores de Decisão – *Random Florest*

```
=== Summary ===

Correctly Classified Instances         510               98.646  %
Incorrectly Classified Instances         7                1.354  %
Kappa statistic                          0.3582
Mean absolute error                      0.0237
Root mean squared error                  0.1087
Relative absolute error                 72.9267 %
Root relative squared error             88.0021 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.750    0.988      0.998    0.993      0.403    0.767     0.994     No
                 0.250    0.002    0.667      0.250    0.364      0.403    0.767     0.396     Yes
Weighted Avg.    0.986    0.738    0.983      0.986    0.983      0.403    0.767     0.985
```

Figura 62 - Resultados - Random Forest

```
=== Summary ===

Correctly Classified Instances         509               98.4526 %
Incorrectly Classified Instances         8                1.5474 %
Kappa statistic                          0
Mean absolute error                      0.0305
Root mean squared error                  0.1235
Relative absolute error                 93.8141 %
Root relative squared error             99.9863 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.985      1.000    0.992      0.000    0.399     0.981     No
                 0.000    0.000    0.000      0.000    0.000      0.000    0.399     0.015     Yes
Weighted Avg.    0.985    0.985    0.969      0.985    0.977      0.000    0.399     0.966
```

Figura 63 - Resultados - SimpleCart

```
=== Summary ===

Correctly Classified Instances         510               98.646  %
Incorrectly Classified Instances         7                1.354  %
Kappa statistic                          0.3582
Mean absolute error                      0.0137
Root mean squared error                  0.115
Relative absolute error                 42.2524 %
Root relative squared error             93.0773 %
Total Number of Instances              517

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.998    0.750    0.988      0.998    0.993      0.403    0.774     0.993     No
               0.250    0.002    0.667      0.250    0.364      0.403    0.758     0.323     Yes
Weighted Avg.  0.986    0.738    0.983      0.986    0.983      0.403    0.773     0.983
```

Figura 64 - Resultados - SimpleCart com AdaBoostingM1

6.3 Segmentação

6.3.1 Pré-processamento

O pré processamento utilizado foi semelhante ao da classificação, ou seja, a criação do atributo *fire* e *rain* e a remoção dos atributos que possuem uma relação direta com os mesmos.

6.3.2 Algoritmos testados

EM

```
Clustered Instances

 0       72 ( 14%)
 1       86 ( 17%)
 2       37 (  7%)
 3       51 ( 10%)
 4       34 (  7%)
 5       18 (  3%)
 6       40 (  8%)
 7       33 (  6%)
 8       62 ( 12%)
 9       32 (  6%)
10       43 (  8%)
11        9 (  2%)
```

Figura 65 - Resultados - EM

Simple K-Means

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(517.0)	(167.0)	(80.0)	(67.0)	(90.0)	(113.0)
month	aug	sep	mar	sep	sep	aug
day	sun	thu	sat	tue	fri	sun
FFMC	91.6	91.6	91.7	91	92.4	90.2
DMC	99	99	35.8	129.5	117.9	142.4
DC	745.3	745.3	80.8	692.6	668	601.4
ISI	9.6	9.2	7.8	7	7.1	6.3
temp	17.4	16.8	15.2	24.1	19.6	16.6
RH	27	27	27	27	38	39
wind	2.2	2.2	4	3.1	4.5	3.6
fire	No	Yes	Yes	Yes	No	No
rainYN	No	No	No	No	No	No

Figura 66 - Dados distribuidos por cinco clusters

Clustered Instances		Clustered Instances	
0	345 (67%)	0	263 (51%)
1	172 (33%)	1	134 (26%)
		2	120 (23%)

Clustered Instances		Clustered Instances	
0	203 (39%)	0	167 (32%)
1	169 (33%)	1	80 (15%)
2	75 (15%)	2	67 (13%)
3	70 (14%)	3	90 (17%)
		4	113 (22%)

Figura 67 - Resultados para Vários N's - Simple K-Means

6.4 Associação

6.4.1 Pré-processamento

O pré processamento utilizado foi semelhante ao da classificação e segmentação, ou seja, a criação do atributo *fire* e *rain* e a remoção dos atributos que possuem uma relação direta com os mesmos.

6.4.2 Algoritmos testados

Apriori
=======

Minimum support: 0.1 (52 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 4

Best rules found:

1. wind=2.2 53 ==> rainYN=No 53 <conf:(1)> lift:(1.02) lev:(0) [0] conv:(0.82)
2. wind=3.1 53 ==> rainYN=No 53 <conf:(1)> lift:(1.02) lev:(0) [0] conv:(0.82)
3. fire=Yes 243 ==> rainYN=No 241 <conf:(0.99)> lift:(1.01) lev:(0) [1] conv:(1.25)

6.5 Resultados

Na classificação, o melhor resultado foi obtido com *Suport Vector Mahines*, com 98,8% de acerto nos testes. Esta classificação permite identificar com grande precisão quando irá chover com base nos atributos escolhidos.

Quanto à segmentação, verificou-se que os melhores resultados obtêm-se com SimpleKMeans de 5 *clusters.*

A associação obteve regras com bastante índice de confiança (acima de 98), no entanto estas não são particularmente úteis: as duas primeiras referem que para certos valores de velocidade do vento há chuva e a terceira que quando há incêndio, não há chuva.

6.6 Recomendações

Os resultados obtidos acima não são particularmente relevantes para o tema do *dataset* (incêndios florestais). Na classificação, a impossibilidade de fazer uma previsão precisa da ativação de fogos a partir dos fatores, sugere a falta de mais atributos/fatores relevantes ou poucas entradas/medições.

Recomenda-se, portanto, efetuar mais medições ou adicionar mais fatores de influência no desencadeamento de fogos, de modo a obter informações mais úteis para o tema.

7 Conclusões

A extração de conhecimento individual sobre os dados permitiu uma boa aplicação da classificação, segmentação e associação sobre os dados. Assim, fizeram-se bons modelos de previsão, divisões em clusters e descobriram-se regras úteis e que podem gerar conhecimento no caso real.

Assim, pensamos que este trabalho foi útil para o nosso desenvolvimento profissional na área de extração de conhecimento, e que também nos permitiu aplicar algoritmos no contexto de *data mining.*

Agradecimentos. Para este trabalho, foram de grande utilidade as aulas lecionadas pelo professor César Analide.

Referências

1. Luis M. Candanedo, Veronique Feldheim, Dominique Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788