# Extração de Conhecimento de Dados Estruturados

## Grupo 2

André Almeida Gonçalves A75625, Rogério Gomes Lopes Moreira A74634,
Tiago Filipe Oliveira Sá A71835

Aprendizagem e Extração de Conhecimento, Perfil de Sistemas Inteligentes,
Universidade do Minho

**Resumo** Este relatório serve como descrição dos procedimentos de extração de conhecimento usados em dois conjuntos de dados distintos. São apresentas as diferente decisões tomadas e exploradas ao longo do desenvolvimento do processo, assim como detalhes relevantes da sua implementação.

# 1 Introdução

O ritmo de recolha e armazenamento de dados a partir de contextos reais aumenta de forma dramática, dia após dia. Torna-se assim praticamente impossível tratar e explorar estes dados manualmente. Não só porque os dados recolhidos a partir do mundo real não podem ser imediatamente utilizados para a extração de conhecimento já que muitas vezes apresentam diversos problemas de incoerências, inconsistências ou até mesmo dados inválidos. Surge assim a necessidade de tratar os dados recolhidos e só posteriormente extrair o conhecimento possível, de forma digital. Este tratamento segue uma série de passos que vão desde a seleção, o pré-processamento, a transformação, a mineração dos dados e posterior interpretação dos resultados. Extração de Conhecimento é assim, o processo de descoberta de conhecimento a partir de dados de fontes. O processo de mineração de dados é o mais importante já que consiste em aplicar algoritmos específicos para extrair padrões dos dados, contudo este passo na maioria das vezes é aquele que ocupa menos tempo. De entre os algoritmos para este processo destacam-se a Regressão, a Classificação, a Segmentação e a Associação.
Neste trabalho será realizado um processo de Extração de Conhecimento em dois conjuntos de dados distintos, com suporte da ferramenta WEKA.

## 2 AirQuality

O número grupo é par, por isso, um dos conjunto de dados atribuídos foi o *AirQuality*. Este dataset contém 9358 instâncias de respostas médias horárias de uma matriz de 5 sensores químicos de óxido metálico dentro de um dispositivo multi-sensor químico de qualidade do ar. O dispositivo estava localizado numa área significativamente poluída, ao nível da estrada, dentro de uma cidade italiana. Os dados foram registados de março de 2004 a fevereiro de 2005 (um ano). A cada hora são registadas as concentrações médias de CO, Hidrocarbonetos Não Metánicos, Benzeno, Óxidos de Nitrogénio Total (NOx) e Dióxido de Nitrogénio (NO2) fornecidos por um analisador certificado de referência colocalizado. Contém 15 atributos e os valores em falta são representados por -200.

### 2.1 Objetivos do estudo

Uma das primeiras coisas que o grupo detetou neste conjunto de dados foi a presença de valores nulos em grande quantidade, indicando que ou o valor registado pelo sensor em questão não é válido (por exemplo má calibração do sensor) ou o valor simplesmente não foi registado. Este foi o ponto de partida para o nosso estudo, uma vez que definimos como um dos objetivos encontrar um modelo de Regressão Linear que nos permitisse estimar os valores em falta, não só para os dados presentes no conjunto como para eventuais falhas futuras. Por outro lado, achamos também pertinente tentar perceber de que maneira os poluentes no ar se relacionam entre si, para assim entender, por exemplo, formas efetivas de colmatar o problema. Para isto vai ser usado um modelo de Associação que permite estabelecer relações entre os diferentes atributos.

### 2.2 Atributos do conjunto de dados

De seguida listam-se os atributos originais do conjunto de dados e a sua descrição.

Pela primeira análise aos atributos conseguimos facilmente perceber que o conjunto de dados é de bastante valor um vez que temos todos os atributos de valor numérico e, por isso, facilmente analisados.

| Atributo | Descrição | Tipo |
|---|---|---|
| Date | Data do registo | Date |
| Time | Hora do registo | Date |
| Concentração de CO[1] | Concentração média de CO em mg/metro3 | Valor numérico |
| PT08.S1(CO) | Resposta média do sensor 1 por hora | Valor numérico |
| Concentração de NMHC[2] | Concentração média de NMHC em microg/metro3 | Valor numérico |
| Concentração de C6H6 | Concentração média de C6H6 em microg/metro3 | Valor numérico |
| PT08.S2(NMHC) | Resposta média do sensor 2 por hora | Valor numérico |
| Concentração de NOx[3] | Concentração média de Nox em ppb | Valor numérico |
| PT08.S3(NOx) | Resposta média do sensor 3 por hora | Valor numérico |
| NO2[4] | Concentração média de NO2 em ppb | Valor numérico |
| PT08.S4(NO2) | Resposta média do sensor 4 por hora | Valor numérico |
| PT08.S5(O3) | Resposta média do sensor 5 por hora | Valor numérico |
| T | Temperatura em graus Celsius | Valor numérico |
| RH | Percentagem de humidade relativa | Valor numérico |
| AH | Humidade Absoluta | Valor numérico |

### 2.3 Preparação do conjunto

Para uma posterior correta análise dos dados foi necessário , antes de partir para a exploração dos dados, tratar os dados tendo em vista a sua uniformização. Para isso foram utilizadas duas ferramentas: Excel e o Weka.

O primeiro passo neste conjunto de dados foi separar a data para ser possível retirar conclusões quanto a dias, meses e anos em separado. Para isso dividiu-se o atributo Date em três novos atributos: Day, Month e Year.

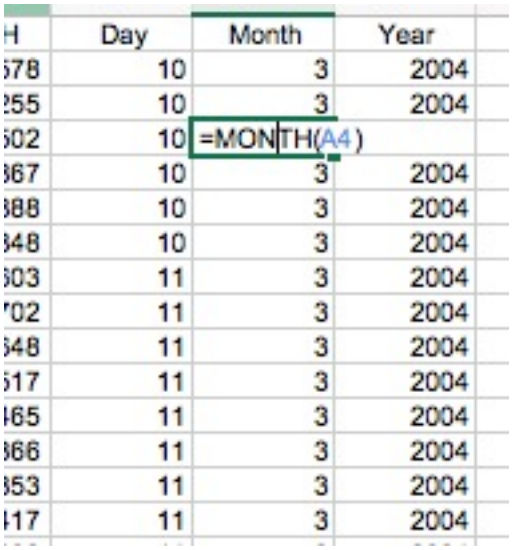


**Figura 1.** Fórmula utilizada no Excel

Um dos primeiros problemas detetados pelo grupo neste dataset foi a presença de valores nulos identificados com o valor -200. Como o Weka interpreta isso como mais um valor do dataset e não como um valor nulo todos estes valores foram alterados para ?. Foi também necessário alterar os field separator e line separator do ficheiro .csv. De seguida dividiu-se o problema em duas instâncias. A primeira instância seria aquela à qual se iria aplicar Regressão Linear e para isso foi necessário o seguinte conjunto de passos:

1. Transformar o separador decimal em ponto
2. Retirar o line separator do ċsv
3. Transformar o atribute separator para vírgula.
4. Carregar os dados para o Weka e remover os atributos relacionados com o tempo já que pretendiamos um modelo onde o tempo e a hora não fosse influência dos restantes fatores
5. Passar os atributos do tipo String para Nominal e através da edição do ficheiro .arff mudar para numeric

A segunda instância seria aquela onde se iria aplicar métodos de Associação e para isso foi necessário o seguinte conjunto de passos:

1. Passar o atributo Time para Hour uma vez que todas as horas representadas eram absolutas, passando este a ser um atributo do tipo numérico
2. Transformar o separador decimal em ponto
3. Retirar o line separator do ċsv
4. Transformar o atribute separator para
5. Carregar os dados para o Weka
6. Passar os atributos do tipo String para Nominal
7. Aplicar o filtro Discretize aos dados Nominal
8. Remover os atributos Year e Month

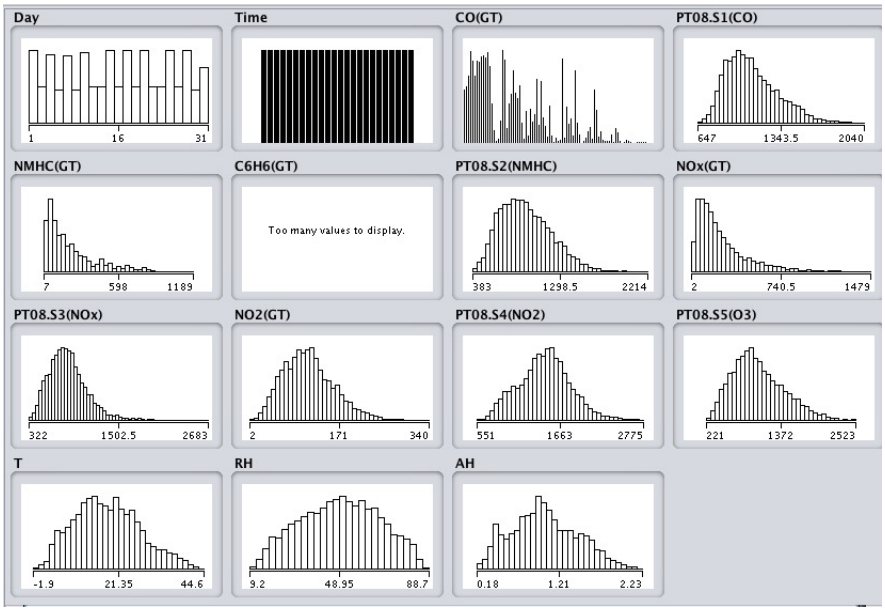


**Figura 2.** AirQuality - conjunto de dados para regressão

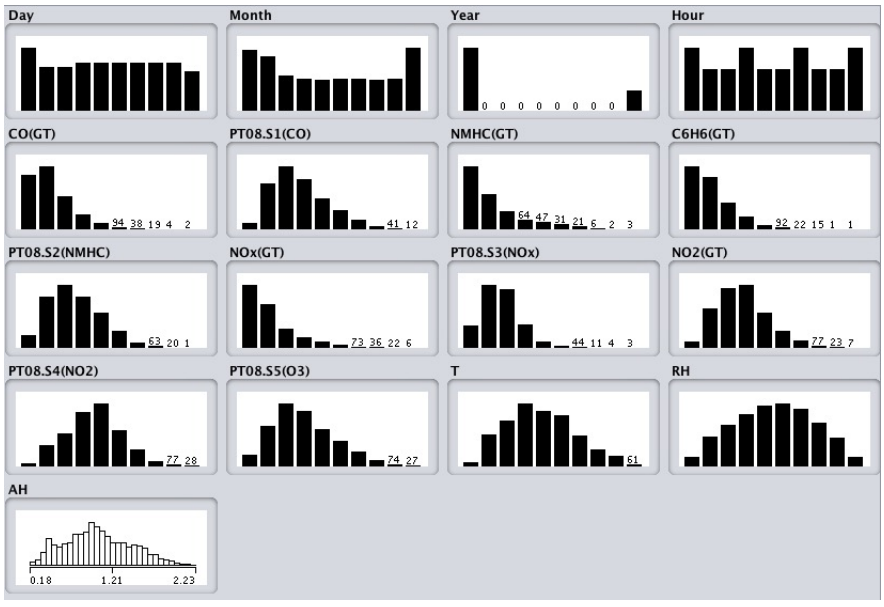


**Figura 3.** AirQuality - conjunto de dados para associação

### 2.4 Regressão linear

Utilizou-se o classificador de Regressão Linear do Weka para estimar os diversos atributos através dos restantes. Para isto as configurações definidas foram o algoritmo M5 eliminando os atributos co-lineares.

**Função estimada pela regressão para calcular o atributo CO(GT)**

```
=== Classifier model (full training set) ===


Linear Regression Model

CO(GT) =

      0.0013 * PT08.S1(CO) +
      0.001  * NMHC(GT) +
      0.0822 * C6H6(GT) +
      0.0025 * NOx(GT) +
      0.0002 * PT08.S3(NOx) +
      0.0022 * NO2(GT) +
      0.0011 * PT08.S4(NO2) +
     -0.0005 * PT08.S5(O3) +
     -0.0226 * T +
     -0.0069 * RH +
     -0.0853 * AH +
     -1.6358

Time taken to build model: 0.07 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.07 seconds

=== Summary ===

Correlation coefficient                  0.9427
Mean absolute error                      0.3077
Root mean squared error                  0.4847
Relative absolute error                 27.5315 %
Root relative squared error             33.352  %
Total Number of Instances             7674
Ignored Class Unknown Instances               1683
```

**Função estimada pela regressão para calcular o atributo NMHC(GT)**

```
Linear Regression Model

NMHC(GT) =

      88.4515 * CO(GT) +
      -0.507  * PT08.S1(CO) +
       0.625  * PT08.S2(NMHC) +
       0.0798 * PT08.S3(NOx) +
       0.1362 * PT08.S4(NO2) +
      -0.0317 * PT08.S5(O3) +
      -9.5206 * T +
      -3.5289 * RH +
     247.2342 * AH +
    -116.041

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correlation coefficient                  0.9125
Mean absolute error                     61.9972
Root mean squared error                 83.5736
Relative absolute error                 39.6808 %
Root relative squared error             40.8977 %
Total Number of Instances              914
Ignored Class Unknown Instances               8443
```

**Função estimada pela regressão para calcular o atributo C6H6(GT)**

```
=== Classifier model (full training set) ===


Linear Regression Model

C6H6(GT) =

      0.2704 * CO(GT) +
      0.001  * PT08.S1(CO) +
      0.0016 * NMHC(GT) +
      0.028  * PT08.S2(NMHC) +
      0.0031 * NOx(GT) +
      0.0037 * PT08.S3(NOx) +
     -0.0103 * NO2(GT) +
      0.0005 * PT08.S4(NO2) +
     -0.0003 * PT08.S5(O3) +
     -0.0963 * T +
     -0.0276 * RH +
      1.2435 * AH +
    -19.5928

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correlation coefficient                  0.9883
Mean absolute error                      0.8159
Root mean squared error                  1.138
Relative absolute error                 14.1102 %
Root relative squared error             15.277  %
Total Number of Instances             8991
Ignored Class Unknown Instances                366
```

**Função estimada pela regressão para calcular o atributo NMHC(GT)**

```
=== Classifier model (full training set) ===


Linear Regression Model

PT08.S2(NMHC) =

      -2.3276 * CO(GT) +
      -0.0412 * NMHC(GT) +
      23.8649 * C6H6(GT) +
       0.0163 * NOx(GT) +
      -0.1706 * PT08.S3(NOx) +
       0.1317 * NO2(GT) +
       0.1318 * PT08.S4(NO2) +
       0.0477 * PT08.S5(O3) +
       2.2441 * T +
     -76.4193 * AH +
     632.8796

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 1.29 seconds

=== Summary ===

Correlation coefficient                  0.9922
Mean absolute error                     24.4286
Root mean squared error                 33.1979
Relative absolute error                 11.3125 %
Root relative squared error             12.4422 %
Total Number of Instances             8991
Ignored Class Unknown Instances                366
```

**Função estimada pela regressão para calcular o atributo NOx(GT)**

```
=== Classifier model (full training set) ===


Linear Regression Model

NOx(GT) =

      64.8127 * CO(GT) +
      -0.0461 * PT08.S1(CO) +
      -0.2584 * NMHC(GT) +
      12.2255 * C6H6(GT) +
       0.2433 * PT08.S2(NMHC) +
       0.071  * PT08.S3(NOx) +
       1.2564 * NO2(GT) +
      -0.492  * PT08.S4(NO2) +
       0.0685 * PT08.S5(O3) +
       3.3026 * T +
       3.2582 * RH +
     104.3763 * AH +
     -24.4399

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correlation coefficient                  0.9246
Mean absolute error                     57.5821
Root mean squared error                 81.1059
Relative absolute error                 36.2008 %
Root relative squared error             38.0841 %
Total Number of Instances             7718
Ignored Class Unknown Instances               1639
```

**Função estimada pela regressão para calcular o atributo NO2(GT)**

```
=== Classifier model (full training set) ===


Linear Regression Model

NO2(GT) =

      5.6621 * CO(GT) +
      0.0219 * PT08.S1(CO) +
     -5.0209 * C6H6(GT) +
      0.1107 * PT08.S2(NMHC) +
      0.0997 * NOx(GT) +
     -0.0472 * PT08.S3(NOx) +
      0.0149 * PT08.S5(O3) +
     -0.5216 * RH +
    -36.8543 * AH +
     83.9743

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correlation coefficient                  0.8821
Mean absolute error                     16.9002
Root mean squared error                 22.786
Relative absolute error                 44.2808 %
Root relative squared error             47.1107 %
Total Number of Instances             7715
Ignored Class Unknown Instances               1642
```

**Função estimada pela regressão para calcular o atributo T**

```
=== Classifier model (full training set) ===


Linear Regression Model

T =

      -0.1851 * CO(GT) +
      -0.0006 * NMHC(GT) +
      -0.3977 * C6H6(GT) +
       0.0107 * PT08.S2(NMHC) +
       0.0029 * NOx(GT) +
       0.0004 * PT08.S3(NOx) +
      -0.0018 * NO2(GT) +
       0.0058 * PT08.S4(NO2) +
      -0.0028 * PT08.S5(O3) +
      -0.3396 * RH +
      13.9848 * AH +
       8.6789

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correlation coefficient                  0.965
Mean absolute error                      1.7173
Root mean squared error                  2.3153
Relative absolute error                 23.7038 %
Root relative squared error             26.2166 %
Total Number of Instances             8991
Ignored Class Unknown Instances                366
```

**Função estimada pela regressão para calcular o atributo RH**

```
=== Classifier model (full training set) ===


Linear Regression Model

RH =

      -0.3998 * CO(GT) +
       0.0095 * PT08.S1(CO) +
      -0.002  * NMHC(GT) +
      -0.772  * C6H6(GT) +
      -0.0008 * PT08.S2(NMHC) +
       0.0187 * NOx(GT) +
      -0.0018 * PT08.S3(NOx) +
      -0.0405 * NO2(GT) +
       0.0159 * PT08.S4(NO2) +
      -0.0015 * PT08.S5(O3) +
      -2.2981 * T +
      33.2042 * AH +
      36.6834

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correlation coefficient                  0.9376
Mean absolute error                      4.6628
Root mean squared error                  6.0204
Relative absolute error                 32.3253 %
Root relative squared error             34.7681 %
Total Number of Instances             8991
Ignored Class Unknown Instances                366
```

**Função estimada pela regressão para calcular o atributo AH**

```
=== Classifier model (full training set) ===


Linear Regression Model

AH =

      -0.0098 * CO(GT) +
      -0.0002 * PT08.S1(CO) +
       0.0147 * C6H6(GT) +
      -0.001  * PT08.S2(NMHC) +
       0.0002 * NOx(GT) +
      -0.0004 * PT08.S3(NOx) +
      -0.0009 * NO2(GT) +
       0.0005 * PT08.S4(NO2) +
       0.0001 * PT08.S5(O3) +
       0.0401 * T +
       0.0141 * RH +
       0.3426

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correlation coefficient                  0.9518
Mean absolute error                      0.095
Root mean squared error                  0.1239
Relative absolute error                 28.9207 %
Root relative squared error             30.6792 %
Total Number of Instances             8991
Ignored Class Unknown Instances                366
```

Não aplicamos o modelo de regressão linear aos atributos PT08.Sx uma vez que entendemos que estes representavam características dos sensores e por isso não deveriam ser abrangidas pelo modelo.

## 2.5 Associação

Foi de seguida aplicado um algoritmo de Associação para retirar conhecimento a partir dos dados. O algoritmo utilizado foi o Apriori com as seguintes especificações:


**Figura 4.** Filtro Apriori

```
Apriori
=======

Minimum support: 0.1 (936 instances)
Minimum metric <confidence>: 0.65
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 48

Size of set of large itemsets L(2): 24

Best rules found:

 1. NO2(GT)='(35.8-69.6]' 1298 ==> NOx(GT)='(-inf-149.7]'
    1192    <conf:(0.92)> lift:(2.65) lev:(0.08) [741]
    conv:(7.92)
 2. PT08.S5(O3)='(1141.8-1372]' 1298 ==> PT08.S3(NOx)
    ='(558.1-794.2]' 1015    <conf:(0.78)> lift:(2.28)
    lev:(0.06) [570] conv:(3.01)
 3. PT08.S1(CO)='(1064.9-1204.2]' 1951 ==> PT08.S3(NOx)
    ='(558.1-794.2]' 1435    <conf:(0.74)> lift:(2.15)
    lev:(0.08) [766] conv:(2.48)
 4. NO2(GT)='(69.6-103.4]' 2001 ==> NOx(GT)='(-inf
    -149.7]' 1348    <conf:(0.67)> lift:(1.94) lev:(0.07)
    [654] conv:(2)
```

De referir que depois de algumas tentativas só conseguimos obter quatro regras para os dados, contudo tivemos que baixar o grau de confiança para >65 % .

- Sempre que o valor de NO2 está entre 35.8 e 69.6 então o valor de NOx é sempre inferior a 149.7, com um grau de confiança de 92%.
- Sempre que o valor do sensor PT08.S5 está entre 1141.8 e 1372 então o valor do sensor PT08.S3(NOx) está entre 558.1 e 794.2, com um grau de confiança de 78%.
- Sempre que o valor do sensor PT08.S1(CO) está entre 1064.9 e 1204.2 então o valor do sensor PT08.S3(NOx) está entre 558.1 e 794.2, com um grau de confiança de 74%.

### 2.6 Resultados e Recomendações

Conseguimos assim obter os resultados pretendidos, de modo a obter funções de regressão linear que de maneira simples permitem estimar a grande quantidade de valores em falta no conjunto de dados. Foi também possível obter

as regras de associação entre os diferentes atributos. Da análise dos resultados obtidos tiramos as seguintes recomendações:

- É possível diminuir a quantidade de tipos de medições já que alguns dos parâmetros são possíveis de ser estimados, ou seja, poderemos ter sensores de menor custo já que a quantidade de elementos a medir será menor;
- Com sensores de menor custo poderemos abranger mais área, permitindo perceber a poluição atmosférica de por exemplo uma cidade inteira;
- Sempre que a quantidade de Dióxido de Nitrogénio no ar está entre 35.8 e 69.6 então a quantidade de Óxido de Nitrogénio tem um máximo de 149.7, indicando que a libertação deste tipo de gases poluentes está relacionada;
- Sempre que os valores do sensor PT08.S1, calibrado para CO, está entre 1064.9 e 1204.2 então o valor do sensor PT08.S3, calibrado para NOx, está entre 558.1 e 794.2 indicando que as medições estão relacionadas.

# 3 Online News

O segundo dataset escolhido foi o Online News. Este conjunto de dados sumariza várias características de artigos publicados online no Mashable, durante um período de dois anos. A data da recolha é 8 de janeiro de 2015. Este conjunto de dados tem mais de 34.600 registos com 61 atributos cada. Não tem valores nulos, contudo existem valores não classificados. O objetivo primordial do conjunto de dados é estimar o número de partilhas de um artigo.

## 3.1 Objetivos do estudo

Aquando da escolha deste conjunto de dados uma das coisas que foi imediatamente referenciada pelo Professor foi a presença de grande número de atributos. Foi esse o nosso ponto de partida para o estudo deste conjunto de dados. Tentar perceber quais os atributos que realmente iriam contribuir para o estudo em causa e quais aqueles que poderiam ser descartados.
O segundo ponto do estudo será perceber quais as características principais que tornam um artigo mais partilhado ou menos partilhado e a forma como os próprios jornalistas podem usar técnicas para melhorar a performance social do artigo/notícia.

## 3.2 Atributos do conjunto de dados

De seguida listam-se os atributos originais do conjunto de dados e a sua descrição.

| Atributo | Descrição | Tipo |
|---|---|---|
| url | Endereço do artigo, atributo não preditivo | String |
| timedelta | Tempo entre a publicação do artigo e a aquisição do conjunto de dados, atributo não preditivo | Valor numérico |
| n_tokens_title | Número de palavras no título | Valor numérico |
| n_tokens_content | Número de palavras do artigo | Valor numérico |
| n_unique_tokens | Percentagem de palavras únicas no artigo | Valor numérico |
| n_non_stop_unique_tokens | Percentagem de "palavras vazias"no artigo | Valor numérico |
| num_hrefs | Número de links externos | Valor numérico |
| num_self_hrefs | Número de links para outros artigos do Mashable | Valor numérico |
| num_imgs | Número de imagens | Valor numérico |
| num_videos | Número de vídeos | Valor numérico |
| average_token_length | Média de tamanho das palavras no artigo | Valor numérico |
| num_keywords | Número de palavras-chave | Valor numérico |
| data_channel_is_lifestyle | Da categoria Lifestyle | Boleano |
| data_channel_is_entertainment | Da categoria Entretenimento | Boleano |
| data_channel_is_bus | Da categoria Business | Boleano |
| data_channel_is_socmed | Da categoria Social Media | Boleano |
| data_channel_is_tech | Da categoria Tecnologia | Boleano |
| data_channel_is_world | Da categoria Mundo | Boleano |
| kw_min_min | Mínimo de partilhas da pior palavra-chave | Valor numérico |
| kw_max_min | Máximo de partilhas da pior palavra-chave | Valor numérico |
| kw_avg_min | Média de partilhas da pior palavra-chave | Valor numérico |
| kw_min_max | Mínimo de partilhas da melhor palavra-chave | Valor numérico |
| kw_max_max | Máximo de partilhas da melhor palavra-chave | Valor numérico |
| kw_avg_max | Média de partilhas da melhor palavra-chave | Valor numérico |
| kw_min_avg | Mínimo de partilhas da palavra-chave média | Valor numérico |
| kw_max_avg | Máximo de partilhas da palavra-chave média | Valor numérico |
| kw_avg_avg | Média de partilhas da palavra-chave média | Valor numérico |
| self_reference_min_shares | Mínimo de partilhas dos artigos próprios referenciados | Valor numérico |
| self_reference_max_shares | Máximo de partilhas dos artigos próprios referenciados | Valor numérico |
| self_reference_avg_sharess | Média de partilhas dos artigos próprios referenciados | Valor numérico |
| weekday_is_monday | Publicado há segunda | Boleano |
| weekday_is_tuesday | Publicado há terça | Boleano |
| weekday_is_wednesday | Publicado há quarta | Boleano |
| weekday_is_thursday | Publicado há quinta | Boleano |
| weekday_is_friday | Publicado há sexta | Boleano |
| weekday_is_saturday | Publicado há sábado | Boleano |
| weekday_is_sunday | Publicado há domingo | Boleano |
| is_weekend | Publicado ao fim-de-semana | Boleano |
| LDA_00 | Proximidade ao tópico 0 do LDA | Valor numérico |
| LDA_01 | Proximidade ao tópico 1 do LDA | Valor numérico |
| LDA_02 | Proximidade ao tópico 2 do LDA | Valor numérico |
| LDA_03 | Proximidade ao tópico 3 do LDA | Valor numérico |
| LDA_04 | Proximidade ao tópico 4 do LDA | Valor numérico |
| global_subjectivity | Subjetividade global do artigo | Valor numérico |
| global_sentiment_polarity | Polaridade do sentimento do texto | Valor numérico |
| global_rate_positive_words | Percentagem de palavras positivas no artigo | Valor numérico |
| global_rate_negative_words | Percentagem de palavras negativas no artigo | Valor numérico |
| rate_positive_words | Rácio de palavras positivas e neutras | Valor numérico |
| rate_negative_words | Rácio de palavras negativas e neutras | Valor numérico |
| avg_positive_polarity | Média de polaridade das palavras positivas | Valor numérico |
| min_positive_polarity | Mínimo de polaridade das palavras positivas | Valor numérico |
| max_positive_polarity | Máximo de polaridade das palavras positivas | Valor numérico |
| avg_negative_polarity | Média de polaridade das palavras negativas | Valor numérico |
| min_negative_polarity | Mínimo de polaridade das palavras negativas | Valor numérico |
| max_negative_polarity | Máximo de polaridade das palavras negativas | Valor numérico |
| title_subjectivity | Subjetividade do título | Valor numérico |
| title_sentiment_polarity | Polaridade do sentimento do título | Valor numérico |
| abs_title_subjectivity | Grau de subjetividade global do título | Valor numérico |
| abs_title_sentiment_polarity | Grau de polaridade do título | Valor numérico |
| shares | Número de partilhas | Valor numérico |

Pela primeira análise aos atributos conseguimos facilmente perceber que o conjunto de dados é de bastante valor um vez que temos vários atributos de valor numérico e grande quantidade de atributos e informação. A tabela seguinte relaciona os atributos entre si:

Table 2: List of attributes by category.

| Feature | Type (#) |
|---|---|
| **Words** | |
| Number of words in the title | number (1) |
| Number of words in the article | number (1) |
| Average word length | number (1) |
| Rate of non-stop words | ratio (1) |
| Rate of unique words | ratio (1) |
| Rate of unique non-stop words | ratio (1) |
| **Links** | |
| Number of links | number (1) |
| Number of Mashable article links | number (1) |
| Minimum, average and maximum number of shares of Mashable links | number (3) |
| **Digital Media** | |
| Number of images | number (1) |
| Number of videos | number (1) |
| **Time** | |
| Day of the week | nominal (1) |
| Published on a weekend? | bool (1) |

| Feature | Type (#) |
|---|---|
| **Keywords** | |
| Number of keywords | number (1) |
| Worst keyword (min./avg./max. shares) | number (3) |
| Average keyword (min./avg./max. shares) | number (3) |
| Best keyword (min./avg./max. shares) | number (3) |
| Article category (Mashable data channel) | nominal (1) |
| **Natural Language Processing** | |
| Closeness to top 5 LDA topics | ratio (5) |
| Title subjectivity | ratio (1) |
| Article text subjectivity score and its absolute difference to 0.5 | ratio (2) |
| Title sentiment polarity | ratio (1) |
| Rate of positive and negative words | ratio (2) |
| Pos. words rate among non-neutral words | ratio (1) |
| Neg. words rate among non-neutral words | ratio (1) |
| Polarity of positive words (min./avg./max.) | ratio (3) |
| Polarity of negative words (min./avg./max.) | ratio (3) |
| Article text polarity score and its absolute difference to 0.5 | ratio (2) |

| Target | Type (#) |
|---|---|
| Number of article Mashable shares | number (1) |

**Figura 5.** Atributos por categoria

### 3.3 LDA

Um dos atributos que nos chamou mais à atenção, até porque não entendemos o que realmente significa foi o LDA (os 4 atributos relacionados com este tópico). Foi então necessário investigar o que significava. O LDA, Latent Dirichlet Allocation ou alocação latente de Dirichlet é um modelo estatístico generativo que permite explicar conjunto de observações através de grupos não observados, que explicam o porquê de algumas partes dos dados serem semelhantes. Por exemplo, à semelhança do modelo mostrado de seguida, se as observações são palavras adquiridas de vários documentos, isso deriva que cada documento é uma mistura de um pequeno número de tópicos e que a criação de cada palavra é atribuível a um dos tópicos do documento.
Assim, este atributo mede a proximidade do artigo a cada um dos cinco tópicos principais do texto identificados pelo algoritmo.

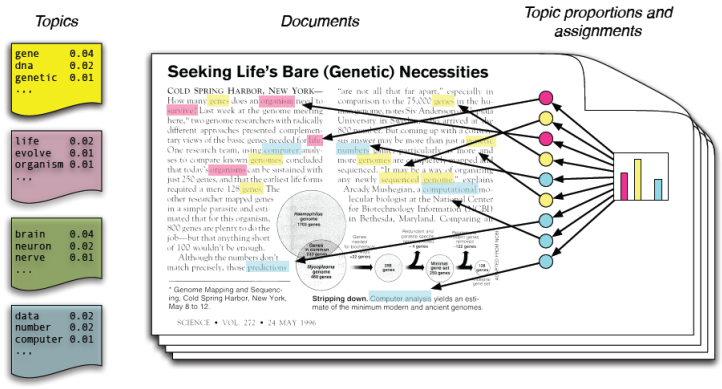


**Figura 6.** O algoritmo LDA

### 3.4 Preparação do conjunto

O primeiro ponto para a preparação do conjunto foi remover alguns atributos triviais. Os vários atributos data channel e date of publication foram convertidos em apenas dois com as seguintes associações:

**Tabela 1.** Data Channel

| | |
|---|---|
| 0 | Lifestyle |
| 1 | Entertainment |
| 2 | Business |
| 3 | Social Media |
| 4 | Technology |
| 5 | World |
| 6 | Other |

**Tabela 2.** Date of Publication

| | |
|---|---|
| 0 | segunda-feira |
| 1 | terça-feira |
| 2 | quarta-feira |
| 3 | quinta-feira |
| 4 | sexta-feira |
| 5 | sábado |
| 6 | domingo |

O passo seguinte será perceber quais os atributos d̈escartáveisë aqueles que mais contribuem para calcular o quão partilhável é um artigo. Contudo, neste aspeto já recorremos a algoritmos de Seleção de Atributos.

### 3.5 Seleção de Atributos

Como já foi indicado o primeiro passo foi selecionar os atributos que mais contribuem para a Class. Não só porque o conjunto de dados tinha um grande volume de atributos como também para perceber quais desses contribuem realmente para as partilhas do artigo e aqueles que menos contribuem. Para isto foi utilizam o algoritmo CorrelationAttributeEval que avalia o valor de um atributo medindo a correlação entre o atributo e a classe (número de partilhas). Este algoritmo precisa também de um método de pesquisa e o mais indicado e com o qual obtivemos melhores resultados foi o Ranker que classifica os atributos através das suas avaliações individuais.

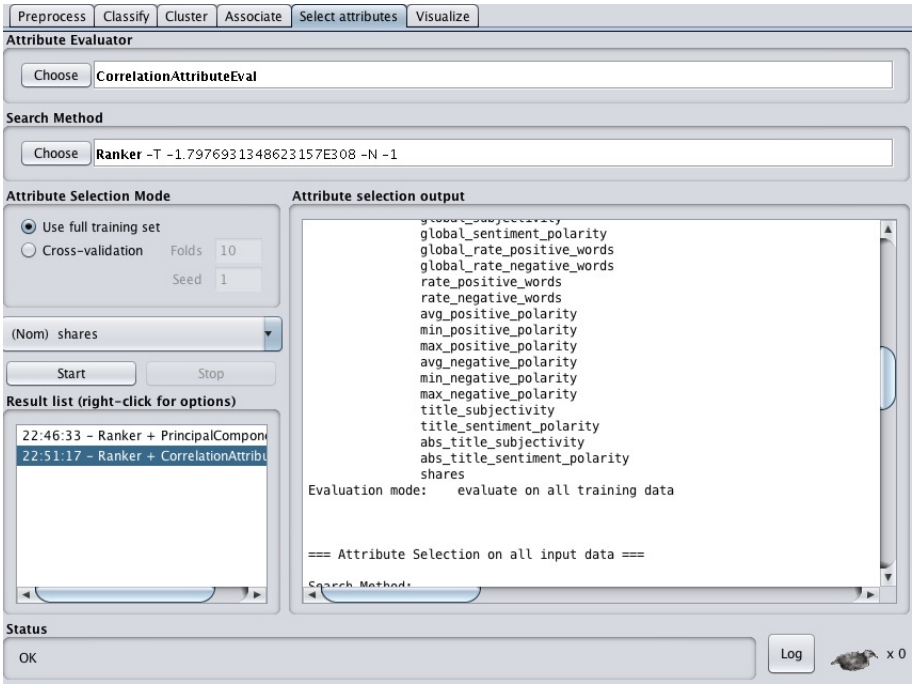


**Figura 7.** Configuração da Seleção de Atributos

```
=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 47
    shares):
  Correlation Ranking Filter
Ranked attributes:
 0.0163     28  LDA\_02
 0.01623    21  kw\_avg\_avg
 0.01259    11  num\_keywords
 0.01241    29  LDA\_03
 0.01208    12 data\_channel
 0.01205    30  LDA\_04
 0.01085    32  global\_sentiment\_polarity
 0.01081    36  rate\_negative\_words
 0.01043    26  LDA\_00
 0.0098     33  global\_rate\_positive\_words
 0.00962    20  kw\_max\_avg
 0.00935     6  num\_hrefs
 0.00912    27  LDA\_01
 0.00889     8  num\_imgs
 0.00879    19  kw\_min\_avg
 0.0084     43  title\_subjectivity
 0.00838    35  rate\_positive\_words
 0.00835    31  global\_subjectivity
 0.00785    46  abs\_title\_sentiment\_polarity
 0.00765    18  kw\_avg\_max
 0.00731    44  title\_sentiment\_polarity
 0.00697    10  average\_token\_length
 0.00688    38  min\_positive\_polarity
 0.00673    34  global\_rate\_negative\_words
 0.00668     7  num\_self\_hrefs
 0.0066     39  max\_positive\_polarity
 0.00631    23  self\_reference\_max\_shares
 0.00613    25 date\_publication
 0.00611    24  self\_reference\_avg\_sharess
 0.00609    40  avg\_negative\_polarity
 0.00606    41  min\_negative\_polarity
 0.00585     2  n\_tokens\_content
 0.00582    15  kw\_avg\_min
 0.00563    37  avg\_positive\_polarity
 0.00562    45  abs\_title\_subjectivity
```

```
0.00559     1   n\_tokens\_title
0.00538    14   kw\_max\_min
0.00513     9   num\_videos
0.00511    22   self\_reference\_min\_shares
0.00489    42   max\_negative\_polarity
0.00483    16   kw\_min\_max
0.00319    17   kw\_max\_max
0.00277    13   kw\_min\_min
0.00107     3   n\_unique\_tokens
0.00105     5   n\_non\_stop\_unique\_tokens
0.00101     4   n\_non\_stop\_words
```

Como podemos ver pela seleção anterior quase todos os atributos contribuem para as partilhas, contudo atributos como o LDA_02, o kw_avg_avg e o num_keywords contribuem mais para o número de partilhas que um artigo vai ter do que por exemplo atributos como o n_unique_tokens e o n_non_stop_words. Assim, esta lista serve como uma lista de prioridades a ter para um jornalista quando está a escrever um novo artigo.

### 3.6 Análise dos Atributos

De seguida analisam-se os atributos, referenciando quais os valores ótimos para se otimizar um artigo quanto ao número de partilhas.

**n_tokens_title**

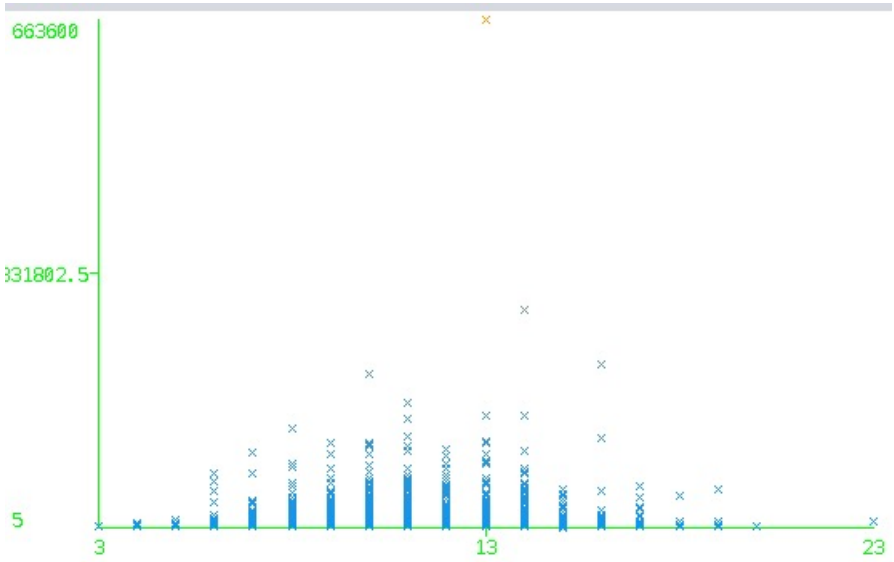


**Figura 8.** Relação entre o n_tokens_title e o número de partilhas

Como podemos observar o número de palavras ótimo para o título é até 13 palavras.

**n_tokens_content**

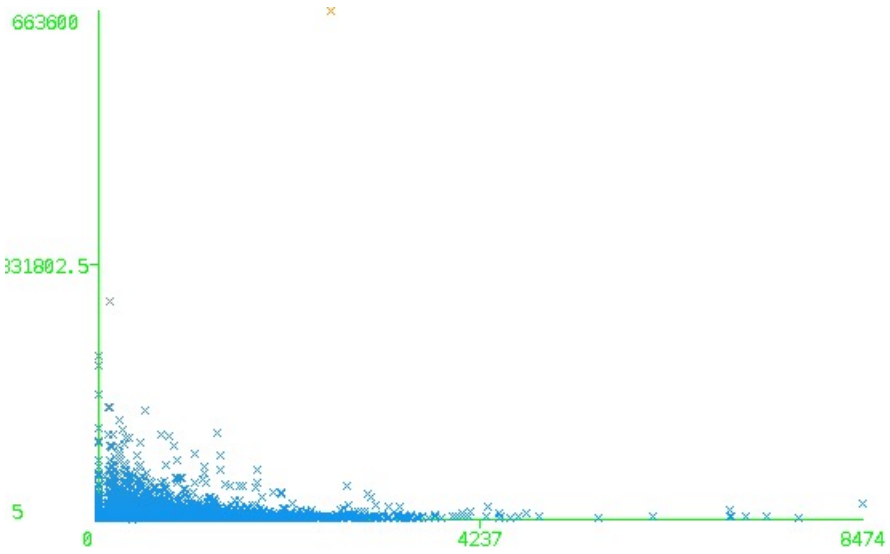


**Figura 9.** Relação entre o n_tokens_content e o número de partilhas

Artigos com menos palavras são mais partilhados. O ideal de palavras de um artigo é menos de 4237.

**num_hrefs**

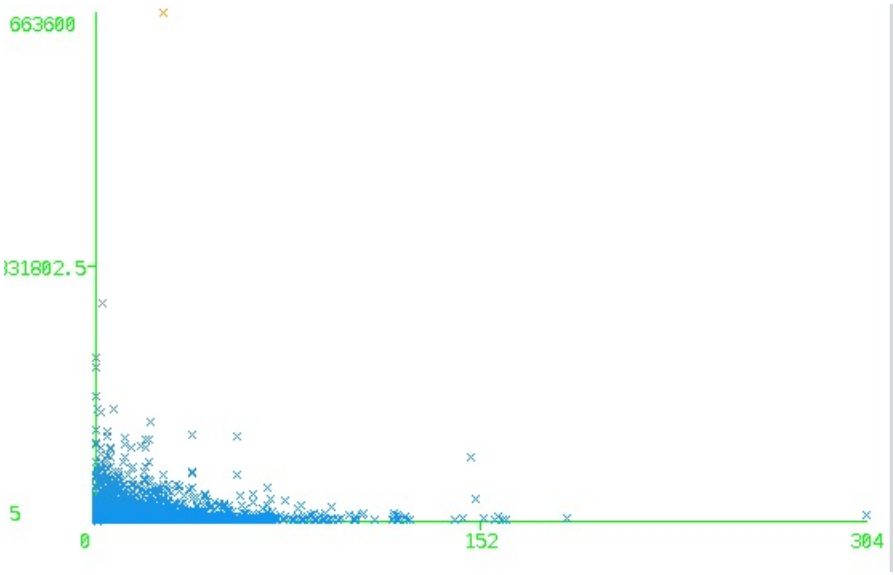


**Figura 10.** Relação entre o num_hrefs e o número de partilhas

Artigos com menos links externos são mais partilhados.
**num_self_hrefs**

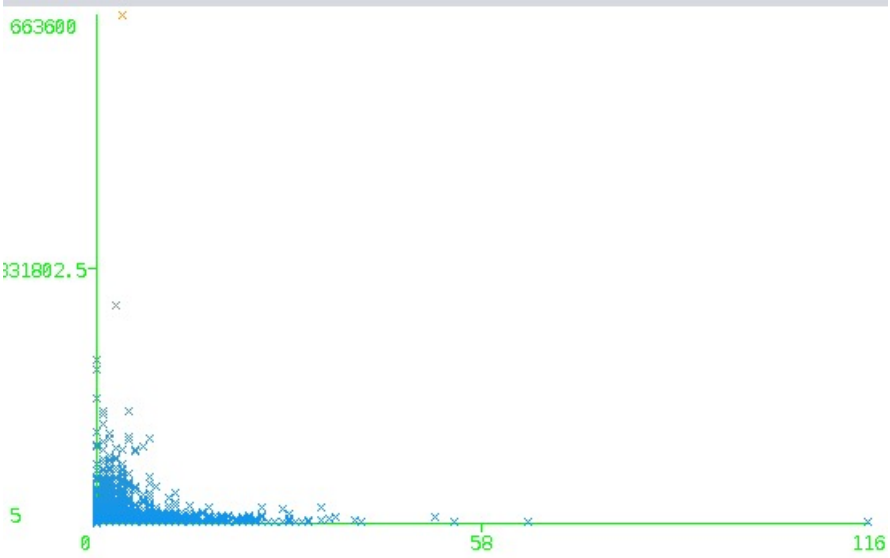


**Figura 11.** Relação entre o num_self_hrefs e o número de partilhas

Artigos com menos links internos são mais partilhados.

**num_imgs**

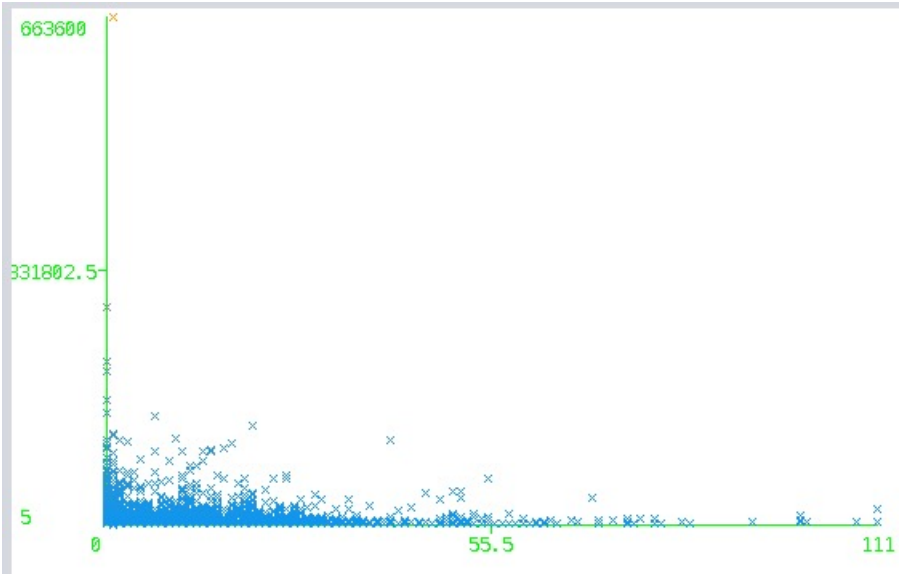


**Figura 12.** Relação entre o num_imgs e o número de partilhas

Um artigo deve ter menos de 55 imagens.

**n_keywords**

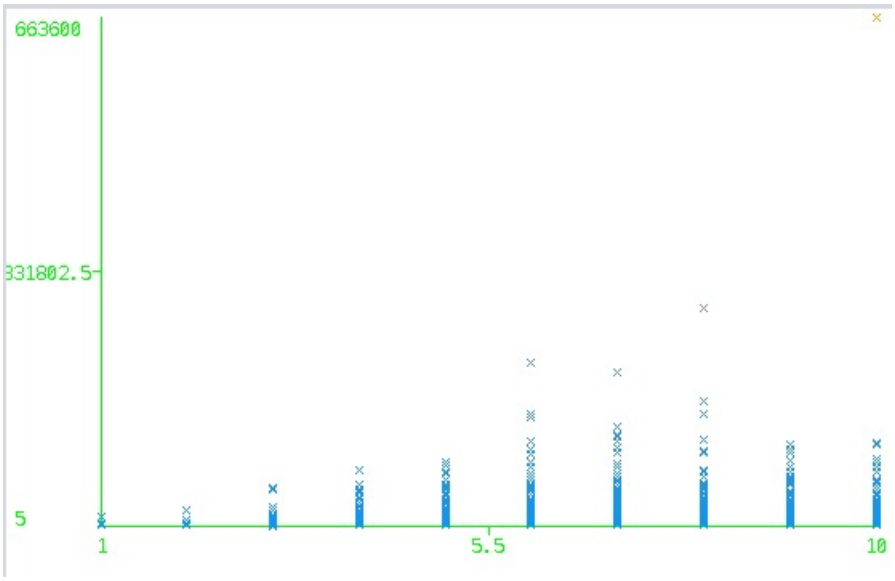


**Figura 13.** Relação entre o n_keywords e o número de partilhas

O artigo deverá ter mais de 5.5 keywords e menos de 10.
**data_channel**

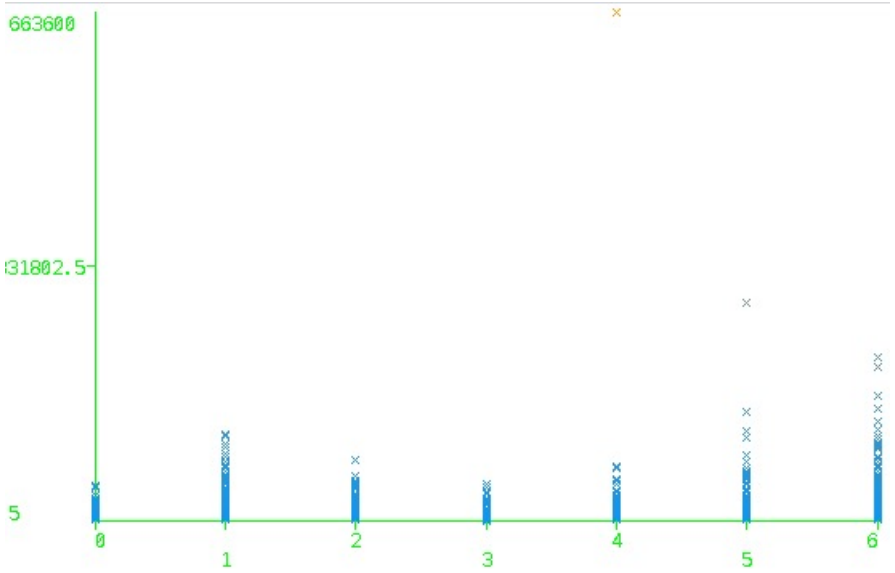


**Figura 14.** Relação entre o data_channel e o número de partilhas

Artigos de World, Other e Entertainment são mais partilhados.

**data_channel**

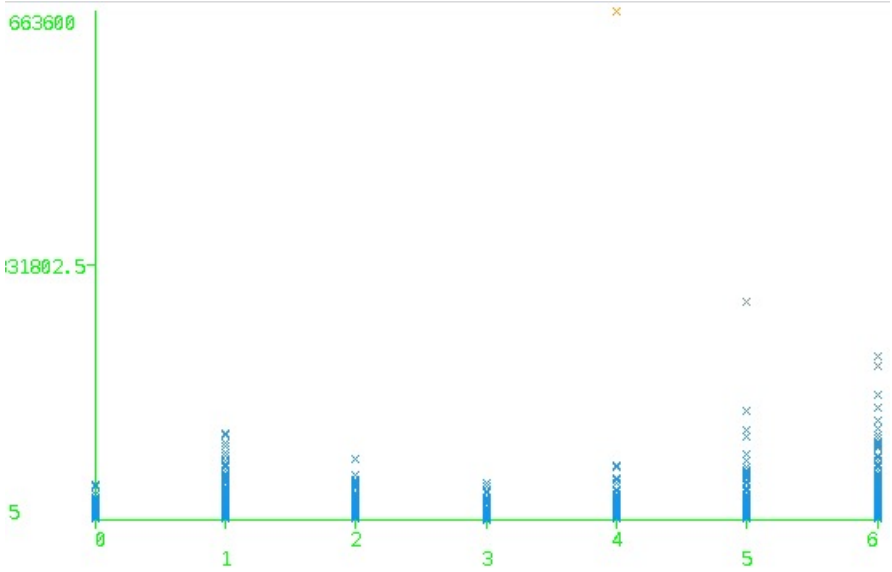


**Figura 15.** Relação entre o data_channel e o número de partilhas

Artigos de World, Other e Entertainment são mais partilhados.

**date_publication**

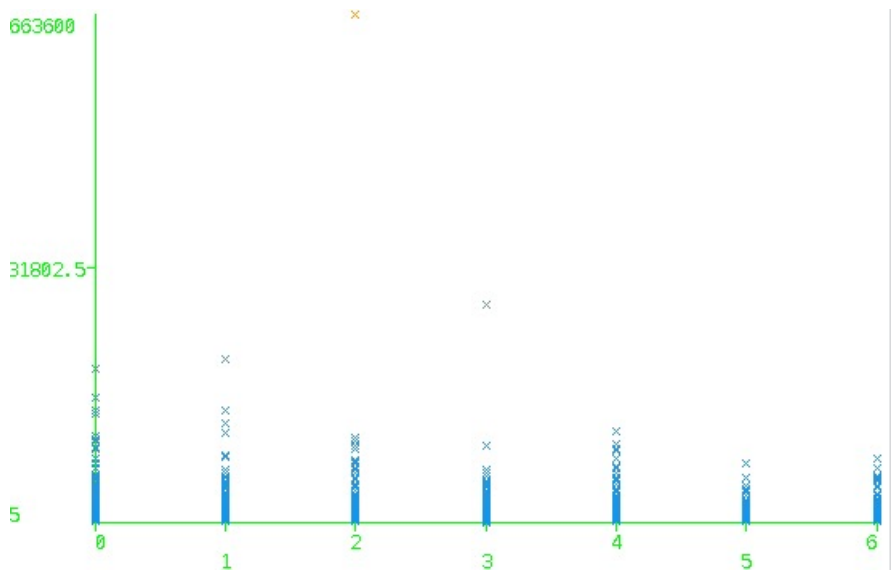


**Figura 16.** Relação entre o date_publication e o número de partilhas

Artigos publicados à segunda-feira e terça-feira são os mais publicados.

### 3.7 Resultados e Recomendações

Conseguimos assim obter os resultados pretendidos, de modo a obter noção das características um artigo deverá ter para ser mais viral, tornando-se mais partilhado pelos leitores. Da análise dos resultados obtidos tiramos ainda as seguintes recomendações, neste caso para o Mashable:

- Publicar mais artigos durante a semana do que ao fim-de-semana;
- Publicar mais artigos sobre World e Entertainment e evitar artigos sobre Social Media e Business;
- Tentar focar e manter o artigo o mais possível perto do título;
- Publicar artigos com mais pequenos, com menos palavras;
- Títulos de artigos mais curtos resultam em mais partilhas;

# 4 Iris

Um dos datasets que o grupo escolheu foi o Iris. Este é, segundo os autores, o maior conjunto de dados conhecido disponível para ser aplicado em processos de extração de conhecimento. O conjunto de dados contém 3 classes com 50 instâncias cada uma, onde cada classe refere-se a um tipo de planta de Íris. Uma classe é linearmente separável dos outros 2 e estes não são linearmente separáveis uns dos outros.

## 4.1 Objetivos do estudo

Uma das razões pelas quais o grupo escolheu este dataset foi o grande número e variedade de artigos científicos publicados de estudos que o utilizaram. Com o estudo deste conjunto de dados e, como o autor refere, pretende-se descobrir a relações entre os diferentes atributos tendo em vista a previsão da classe de Íris, dentro das que existem dados. Seguindo as recomendações do autor do dataset esse foi o nosso grande foco para este estudo.

## 4.2 Atributos do conjunto de dados

De seguida listam-se os atributos originais do conjunto de dados e a sua descrição.

| Atributo | Descrição | Tipo |
|---|---|---|
| Sepal Length | Comprimento da sépala (peça constituinte da flor) em cm | Numérico |
| Sepal Width | Largura da sépala (peça constituinte da flor) em cm | Numérico |
| Petal Length | Comprimento da pétala em cm | Numérico |
| Petal Width | Largura da pétala em cm | Numérico |
| Class | Tipo de Íris (Setosa, Versicolour ou Virginica) | Set.,Vers. ou Virg. |

Pela análise aos atributos conseguimos perceber que o atributo classe é aquele a prever pelo cálculo dos restantes e que o domínio dos atributos é bastante simples.

## 4.3 Preparação do conjunto

Como já foi anteriormente referido houve alguns aspetos que nos chamaram atenção para este conjunto de dados, sendo um deles o facto de os dados estarem bem recolhidos. O primeiro passo para a preparação do conjunto foi verificar se existiriam diferentes fontes de dados para fazer a integração/uniformização dos dados, tal não foi necessário já que se verificou que apenas existia uma fonte de dados. O passo seguinte foi verificar a existência de atributos redundantes, chegando à conclusão que não existem. Ficou apenas a nota de que o atributo classe seria um atributo a prever e não deveria ser considerado como dado de

ëntrada:. Após estes dois passos prosseguimos para a discretização de todos os atributos, exceto a Class.

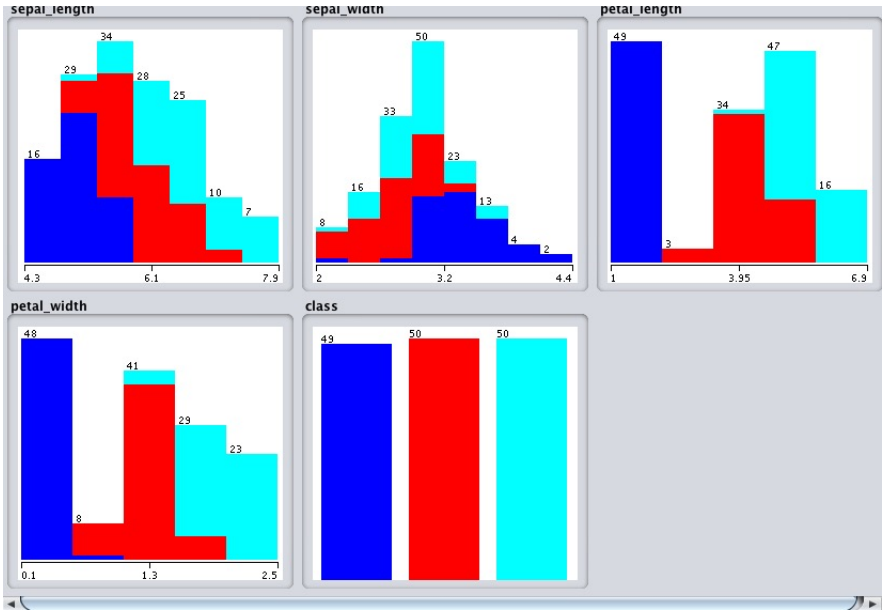


**Figura 17.** Gráfico dos dados antes da discretização

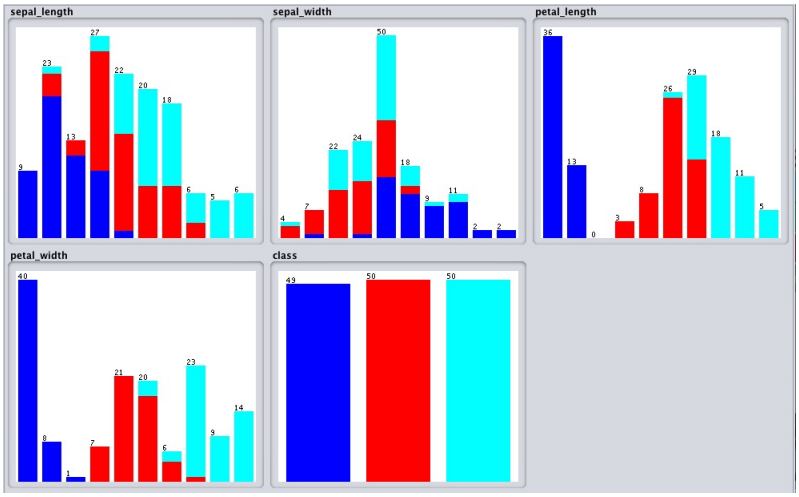


**Figura 18.** Gráfico dos dados após discretização

### 4.4 Classificação

**Algoritmo J48**

O primeiro Algoritmo de Classificação aplicado foi o J48, com a finalidade de construir uma árvore de decisão baseada no conjunto de dados. O algoritmo foi aplicado duas vezes uma delas usando o conjunto de dados como conjunto de treino com uma média absoluta de erro de 2% e utilizando c̈ross-validationc̈om uma média absoluta de erro de 3%. Preferimos então continuar a usar o conjunto como conjunto de treino.

```
=== Classifier model (full training set) ===

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

petal_width <= 0.6: Iris-setosa (49.0)
petal_width > 0.6
|   petal_width <= 1.7
|   |   petal_length <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petal_length > 4.9
|   |   |   petal_width <= 1.5: Iris-virginica (3.0)
|   |   |   petal_width > 1.5: Iris-versicolor (3.0/1.0)
|   petal_width > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :   5

Size of the tree :  9


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         142
   95.302  %
Incorrectly Classified Instances         7
                  4.698  %
Kappa statistic                          0.9295
Mean absolute error                      0.0387
Root mean squared error                  0.1715
Relative absolute error                  8.7015 %
Root relative squared error             36.3696 %
Total Number of Instances              149
```

```
=== 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       0.986
                        Iris-setosa
                 0.940     0.040     0.922       0.940
                    0.931       0.895     0.950       0.870
                        Iris-versicolor
                 0.940     0.030     0.940       0.940
                    0.940       0.910     0.948       0.902
                        Iris-virginica
Weighted Avg.    0.953     0.024     0.954       0.953
   0.953       0.930     0.963      0.919

=== Confusion Matrix ===

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

Destes resultados podemos tirar as seguintes conclusões:

1. Se a largura da pétala for inferior a 0.6cm então a Íris é do tipo Setosa;
2. Se a largura da pétala for maior do que 0.6cm então poderá ser do tipo Versicolor ou Virginica;
3. Se a largura da pétala for menor ou igual a 1.7cm e o seu comprimento for também inferior a 4.9cm então é do tipo Versicolor;
4. Caso a largura da pétala seja menor ou igual a 1.7cm, o comprimento for maior do que 4.9cm e a largura da pétala for menor ou igual a 1.5cm então é do tipo Virginica;
5. Caso cumpra todos os requisitos do parâmetro anterior mas a largura seja maior do que 1.5cm então é do tipo Versicolor;
6. Se a largura da pétala da Íris for maior do que 0.6cm mas o comprimento da pétala for maior do que 1.7cm então é do tipo Virginica.

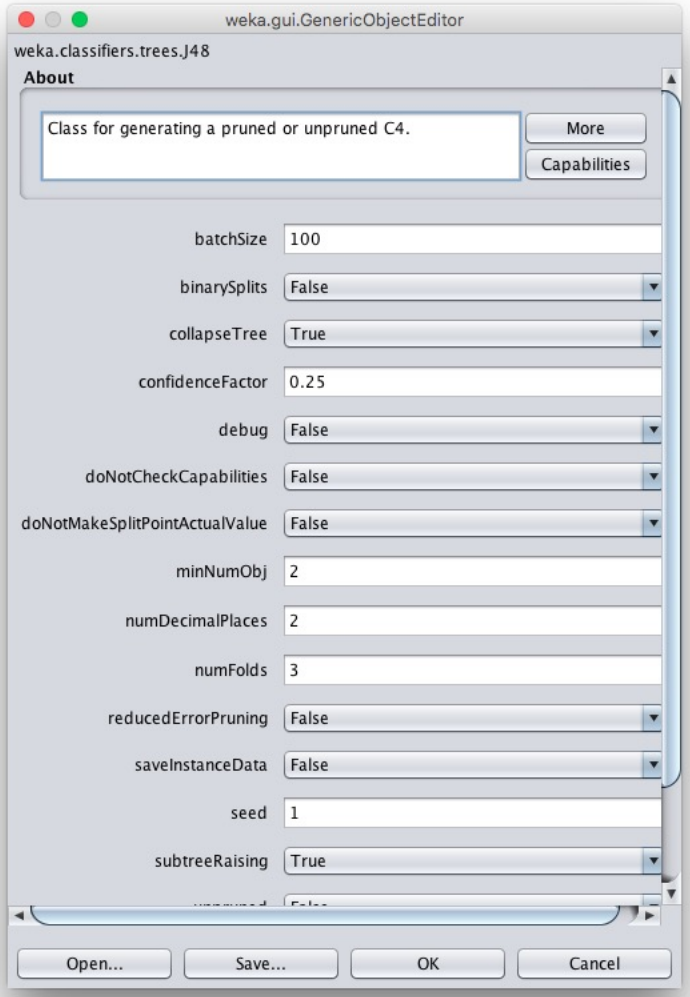


**Figura 19.** Configurações do Algoritmo J48

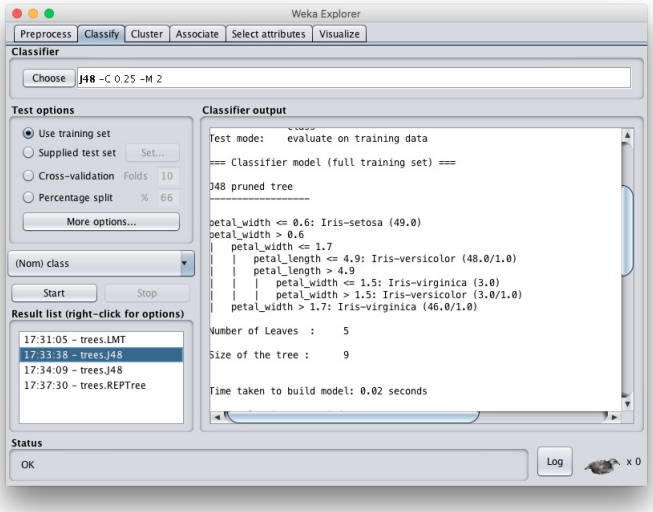


**Figura 20.** Algoritmo J48 usando o conjunto de treino

Foram também explorados os algoritmos LMT e REPTree mas em nenhum deles obtemos tão bons resultados como o J48, tendo em conta os erros absolutos e relativos.

### 4.5 Associação

Aplicamos também algoritmos de Associação para extrair o conhecimento a partir dos dados. O algoritmo utilizado foi o Apriori, para o qual testamos várias configurações até chegar a uma configuração ótima.

**Resultado para uma confiança maior que 60% e suporte mínimo de 0.25**

```
Apriori
=======

Minimum support: 0.25 (37 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 15

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 1
```

```
Best rules found:

 1. petal_width='(-inf-0.34]' 40 ==> class=Iris-setosa 40
        <conf:(1)> lift:(3.04) lev:(0.18) [26] conv
    :(26.85)
 2. class=Iris-setosa 49 ==> petal_width='(-inf-0.34]' 40
        <conf:(0.82)> lift:(3.04) lev:(0.18) [26] conv
    :(3.58)
```

**Resultado para uma confiança maior que 60% e suporte mínimo de 0.20**

```
Apriori
=======

Minimum support: 0.2 (30 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 16

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

 1. petal_width='(-inf-0.34]' 40 ==> class=Iris-setosa 40
        <conf:(1)> lift:(3.04) lev:(0.18) [26] conv
    :(26.85)
 2. petal_length='(-inf-1.59]' 36 ==> class=Iris-setosa
    36    <conf:(1)> lift:(3.04) lev:(0.16) [24] conv
    :(24.16)
 3. petal_length='(-inf-1.59]' petal_width='(-inf-0.34]'
    32 ==> class=Iris-setosa 32    <conf:(1)> lift:(3.04)
     lev:(0.14) [21] conv:(21.48)
 4. petal_length='(-inf-1.59]' 36 ==> petal_width='(-inf
    -0.34]' 32    <conf:(0.89)> lift:(3.31) lev:(0.15)
    [22] conv:(5.27)
 5. petal_length='(-inf-1.59]' class=Iris-setosa 36 ==>
    petal_width='(-inf-0.34]' 32    <conf:(0.89)> lift
    :(3.31) lev:(0.15) [22] conv:(5.27)
```

```
 6. petal_length='(-inf-1.59]' 36 ==> petal_width='(-inf
    -0.34]' class=Iris-setosa 32    <conf:(0.89)> lift
    :(3.31) lev:(0.15) [22] conv:(5.27)
 7. class=Iris-setosa 49 ==> petal_width='(-inf-0.34]' 40
        <conf:(0.82)> lift:(3.04) lev:(0.18) [26] conv
    :(3.58)
 8. petal_width='(-inf-0.34]' 40 ==> petal_length='(-inf
    -1.59]' 32    <conf:(0.8)> lift:(3.31) lev:(0.15)
    [22] conv:(3.37)
 9. petal_width='(-inf-0.34]' class=Iris-setosa 40 ==>
    petal_length='(-inf-1.59]' 32    <conf:(0.8)> lift
    :(3.31) lev:(0.15) [22] conv:(3.37)
10. petal_width='(-inf-0.34]' 40 ==> petal_length='(-inf
   -1.59]' class=Iris-setosa 32    <conf:(0.8)> lift
   :(3.31) lev:(0.15) [22] conv:(3.37)
```

Destes dados retiramos as seguintes conclusões:

- Se a largura das pétalas é menor que 0.34cm então a classe da Íris é Setosa, com um grau de confiança de 100%;
- Se o comprimento das pétalas é menor que 1.59cm então a classe da Íris é Setosa, com um grau de confiança de 100%;
- Se o comprimento das pétalas é menor que 1.59cm e a largura das pétalas é inferior a 0.34cm então a classe da Íris é Setosa, com um grau de confiança de 100%;
- Se o comprimento das pétalas é menor que 1.59cm então a largura das pétalas é também inferior a 0.34cm, com um grau de confiança de 89%.

### 4.6 Regressão Linear

Por curiosidade experimentamos se seria possível calcular a classe da Íris a partir dos restantes parâmetros, ou seja, uma equação matemática que dado as medidas das pétalas e da sépala resultasse no tipo de Íris (1,2,3). Tínhamos a noção que parecia inviável e até um pouco confuso, o que se veio a verificar quando aplicado o modelo de Regressão Linear e por isso nada conseguimos concluir.

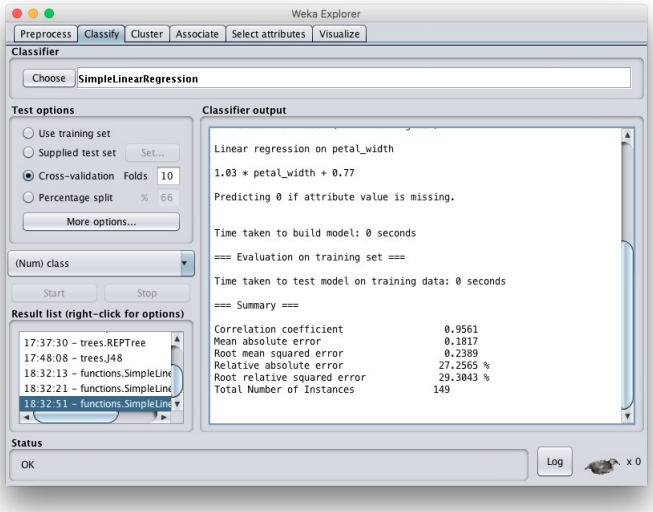


**Figura 21.** Regressão Linear

## 4.7 Resultados e Recomendações

Conseguimos assim obter os resultados inicialmente propostos, ou seja, a partir dos dados observáveis e mensuráveis obter a classe/tipo de Íris. No entanto, os dados analisados representam apenas três tipos da flor que, segundo as nossas pesquisas, apresenta mais de 50 tipos diferentes e por isso o modelo é limitado. De notar ainda que, os dados obtidos tanto por Associação como por Classificação são coerentes entre si.

## 5 Conclusões Finais

Ao longo de todo o processo do desenvolvimento deste projeto fomos-nos apercebendo aquilo que já tinha sido referido pelo professor durante as aulas: a fase de pré-processamento dos dados é de elevado valor uma vez que é o que vai permitir extrair conhecimento a partir dos dados. Outro dos pontos retirados é o facto que a primeira abordagem a um conjunto de dados, a fase de teste dos modelos é particularmente importante no sentido de identificar para cada conjunto de dados a abordagem a tomar na resolução dos problemas com as suas características. A fase de pré-processamento e preparação dos dados foi aquela que, como esperado, tomou mais tempo contudo foi o que permitiu a extração de conhecimento ser realmente eficaz. Durante todo o desenvolvimento foram tomados diversos caminhos, devidamente justificados no relatório, e que produziram os resultados explicitados. No entanto, admitimos que os caminhos não explorados poderiam ter sido melhores ou levar a melhores resultados finais. Todos as estratégias seguidas teriam as suas vantagens e desvantagens, coube ao grupo o balanceamento das duas.

## Referências


1. S. De Vito, G. Fattoruso, M. Pardo, F. Tortorella and G. Di Francia, 'Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction
2. Abernethy, Michael, Data Mining with Weka, https://www.ibm.com/developerworks/library/os-weka1/index.html
3. Vito, S. D. (2016). Air Quality Data Set. https://archive.ics.uci.edu/ml/datasets/Air+quality
4. S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical
5. https://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2017/MashableNews.html
6. https://rstudio-pubs-static.s3.amazonaws.com/122671_778c16d46da6489c9f88cd7c12b20ed3.html
7. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.