

Sistemas de Suporte a Decisão - Mineração de dados (Data Mining) - Previsão de demanda

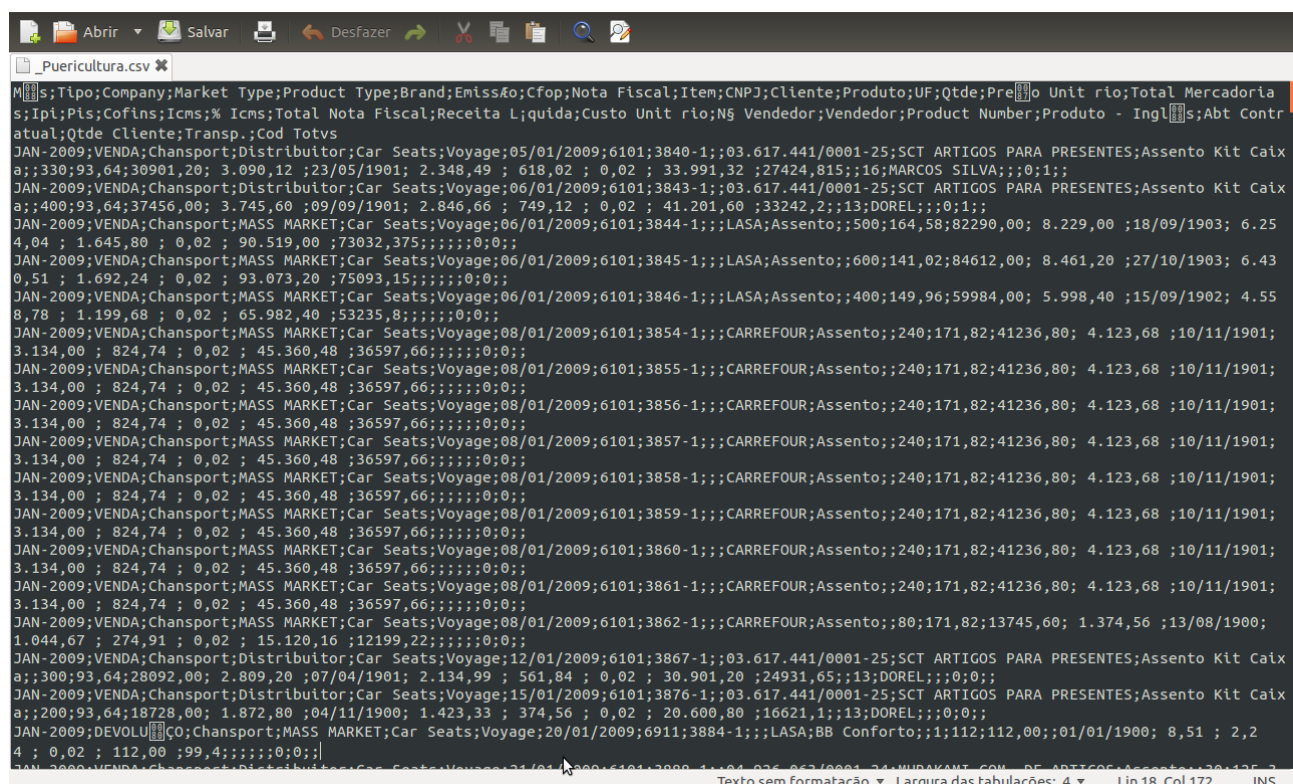
Introdução

Buscando o contexto de mineração de dados, nosso trabalho se baseou em prever demandas de uma determinada empresa, baseada nas demandas anteriores. Através desses dados, é possível estudar e analisar as previsões geradas, ajudando assim no planejamento futuro.

O trabalho consiste em dados retirados de uma base dados, contendo informações de vendas de produtos. Nesse contexto, decidimos então que seria mais relevante prever quantos produtos são vendidos em um determinado mês.

Tratamento dos dados

Adquirimos os dados em formato .csv com a seguinte formatação:



```
M[?];Tipo;Company;Market Type;Product Type;Brand;Emiss[?];Cfop;Nota Fiscal;Item;CNPJ;Cliente;Produto;UF;Qtde;Pre[?]o Unit rio;Total Mercadoria
s;Ipi;Pis;Cofins;Icms;% Icms;Total Nota Fiscal;Receita L[?]quida;Custo Unit rio;Ns Vendedor;Vendedor;Product Number;Produto - Ingl[?]s;Abt Contr
atual;Qtde Cliente;Transp.;Cod Totvs
JAN-2009;VENDA;Chansport;Distribuidor;Car Seats;Voyage;05/01/2009;6101;3840-1;;03.617.441/0001-25;SCT ARTIGOS PARA PRESENTES;Assento Kit Caix
a;;330;93,64;30901,20; 3.090,12 ;23/05/1901; 2.348,49 ; 618,02 ; 0,02 ; 33.991,32 ;27424,815;;16;MARCOS SILVA;;0;1;;
JAN-2009;VENDA;Chansport;Distribuidor;Car Seats;Voyage;06/01/2009;6101;3843-1;;03.617.441/0001-25;SCT ARTIGOS PARA PRESENTES;Assento Kit Caix
a;;400;93,64;37456,00; 3.745,60 ;09/09/1901; 2.846,66 ; 749,12 ; 0,02 ; 41.201,60 ;33242,2;;13;DOREL;;0;1;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;06/01/2009;6101;3844-1;;LASA;Assento;;500;164,58;82290,00; 8.229,00 ;18/09/1903; 6.25
4,04 ; 1.645,80 ; 0,02 ; 90.519,00 ;73032,375;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;06/01/2009;6101;3845-1;;LASA;Assento;;600;141,02;84612,00; 8.461,20 ;27/10/1903; 6.43
0,51 ; 1.692,24 ; 0,02 ; 93.073,20 ;75093,15;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;06/01/2009;6101;3846-1;;LASA;Assento;;400;149,96;59984,00; 5.998,40 ;15/09/1902; 4.55
8,78 ; 1.199,68 ; 0,02 ; 65.982,40 ;53235,8;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3854-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3855-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3856-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3857-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3858-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3859-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3860-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3861-1;;CARREFOUR;Assento;;240;171,82;41236,80; 4.123,68 ;10/11/1901;
3.134,00 ; 824,74 ; 0,02 ; 45.360,48 ;36597,66;;;;;0;0;;
JAN-2009;VENDA;Chansport;MASS MARKET;Car Seats;Voyage;08/01/2009;6101;3862-1;;CARREFOUR;Assento;;80;171,82;13745,60; 1.374,56 ;13/08/1900;
1.044,67 ; 274,91 ; 0,02 ; 15.120,16 ;12199,22;;;;;0;0;;
JAN-2009;VENDA;Chansport;Distribuidor;Car Seats;Voyage;12/01/2009;6101;3867-1;;03.617.441/0001-25;SCT ARTIGOS PARA PRESENTES;Assento Kit Caix
a;;300;93,64;28092,00; 2.809,20 ;07/04/1901; 2.134,99 ; 561,84 ; 0,02 ; 30.901,20 ;24931,65;;13;DOREL;;0;0;;
JAN-2009;VENDA;Chansport;Distribuidor;Car Seats;Voyage;15/01/2009;6101;3876-1;;03.617.441/0001-25;SCT ARTIGOS PARA PRESENTES;Assento Kit Caix
a;;200;93,64;18728,00; 1.872,80 ;04/11/1900; 1.423,33 ; 374,56 ; 0,02 ; 20.600,80 ;16621,1;;13;DOREL;;0;0;;
JAN-2009;DEVOLU[?];CO;Chansport;MASS MARKET;Car Seats;Voyage;20/01/2009;6911;3884-1;;LASA;BB Conforto;;1;112;112,00;01/01/1900; 8,51 ; 2,2
4 ; 0,02 ; 112,00 ;99,4;;;;;0;0;;
JAN-2009;VENDA;Chansport;Distribuidor;Car Seats;Voyage;21/01/2009;6101;3888-1;;04.826.862/0001-24;MURAKAMI COM DE ARTIGOS PARA PRESENTES;20;125,3
```

Ilustração 1: Arquivo aberto pelo Editor de texto padrão do Ubuntu 11.10.

Como é possível ver, o arquivo original contém dados com algumas informações vazias e codificação de caracteres diferente do que é lido pelo Editor de texto. A primeira dificuldade

encontrada no trabalho realizado foi a formatação desses dados, de modo que pudesse ser lido pelo *Weka*, software utilizado para mineração desses dados.

A formatação foi feita através do programa padrão do *Ubuntu 11.10* para manipulação de planilhas eletrônicas, também conhecido como *Calc*. Basicamente, essa parte do tratamento dos dados foi realizado em duas etapas:

- Exclusão de colunas irrelevantes para a previsão e também colunas que continham campos vazios.
- Definição dos tipos de dados de cada coluna, e também manipulação dos valores com ponto flutuando, que foram transformados em tipos inteiro.

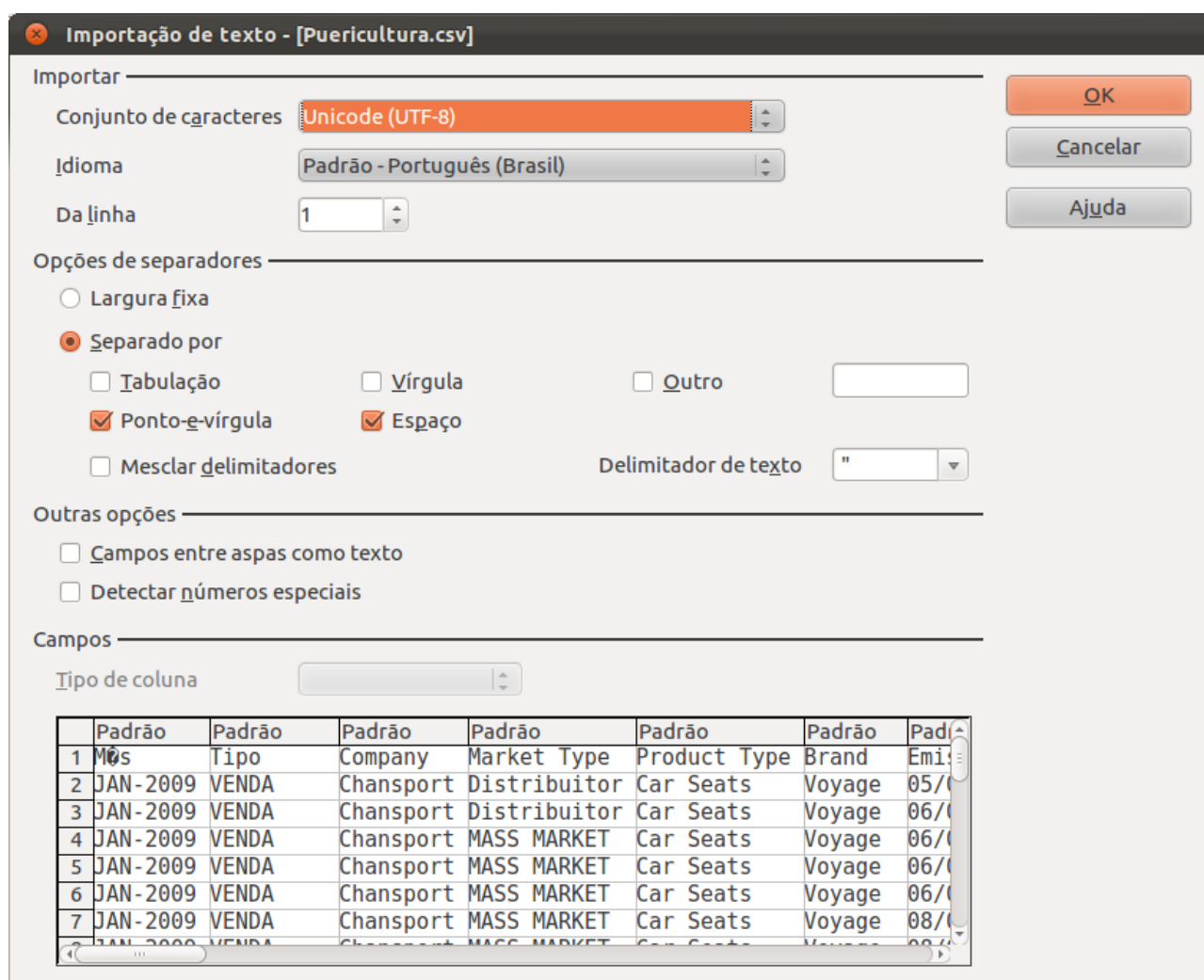


Ilustração 2: Interface inicial ao abrir o arquivo com o *Calc*.

Nesse ponto do trabalho, encontramos algumas barreiras, sendo necessária várias tentativas para a obtenção do resultado esperado. Primeiro, tentamos realizar a conversão do arquivo por sites encontrados na *Internet*. Sem sucesso, tentamos realizar a conversão pelo próprio *Weka*, seguindo um tutorial também encontrado através de buscas pela *Internet*. Nessa segunda tentativa,

conseguimos com que fosse feita a conversão, no entanto o *Weka* determinou a maioria dos campos como classes, tipo de variável usado nos arquivos *.arff*.

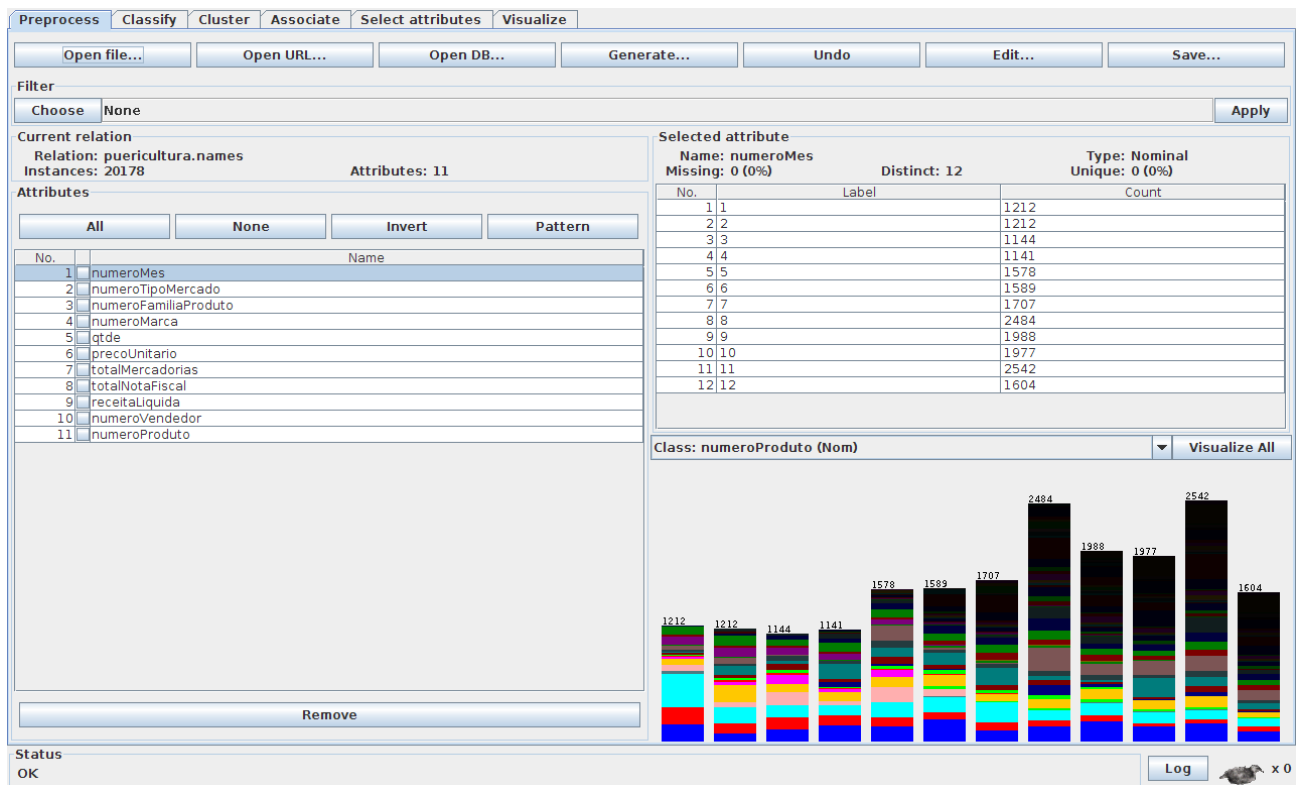
Com isso, optou-se na conversão manual do arquivo, que também demandou um esforço para que pudesse ser feito. Problemas com manipulação de valores do tipo *string*, impactou o andamento do trabalho. A solução encontrada foi a mudança dos valores *string* para *numeric*. Para que isso pudesse ser feito, foi criado um legenda para esses campos, de modo que continuasse sendo possível manter a representação correta dos números dentro do contexto da base de dados. Tomemos como exemplo o campo “Mês”, que originalmente, continha nos campos valores como: “*jan/11*”, “*fev/11*”, “*mar/11*”, e assim por diante. Substitui-se então o “*jan/11*” por 1, “*fev/11*” para 2 e “*mar/11*” para 3. A mesma lógica foi usada para os campos “*Vendedores*”, “*Market type*”, “*Product Families*”, “*brand*” e “*Produtos*”.

Por último, é preciso inserir um cabeçalho no arquivo para que esse possa ser lido como *.arff*. Primeiro definimos um nome para o conjunto de dados através da linha “*@relation 'puericultura.names'*”. Então definimos as colunas dos valores e seus respectivos tipos:

```
@attribute numeroMes { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
@attribute numeroTipoMercado { 1, 2, 3, 4}
@attribute numeroFamiliaProduto { 1, 2, 3, 4, 5, 6}
@attribute numeroMarca { 1, 2, 3, 4, 5, 6}
@attribute qtde real
@attribute precoUnitario real
@attribute totalMercadorias real
@attribute totalNotaFiscal real
@attribute receitaLiquida real
@attribute numeroVendedor { 24, 43, 44, 20, 19, 22, 40, 25, 36, 13, 9, 50, 11, 14, 8, 41, 18,
29, 34, 33, 26, 6, 5, 16, 47, 2, 49, 39, 42, 12, 38, 3, 31, 27, 35, 28}
@attribute numeroProduto { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59}
```

Dessa forma, fica descrito no arquivo quais campos serão números, e quais fazem parte de um contexto, sendo definidos como classes. Por último, colocamos a linha “*@data*” para indicar que a partir dessa linha, estará o conjunto de dados.

Agora o *Weka* é capaz de abrir o arquivo para que seja possível a mineração dos dados:



Colocando o Weka para trabalhar

Escolhemos uma técnica de mineração de dados de fácil entendimento e simples para a execução dos testes. Esse modelo é conhecido como regressão linear, usado para prever o resultado de uma variável dependente desconhecida, dados os valores das variáveis independentes.

Pelo *Weka*, temos a opção de excluir colunas que não serão relevantes para o teste, portanto retiramos as colunas “totalNotaFiscal”, “receitaLiquida”, “numeroVendedor” e “NumeroProduto”. Então na aba “Classify”, na primeira opção escolhemos qual será a função de classificação, no nosso caso, “LinearRegression”. Definimos então qual será a coluna dependente, em outras palavras, a coluna que estamos tentando prever, que no nosso caso será a coluna “qtde”. Após a execução do teste, o *Weka* gera uma saída, que em um primeiro momento parece confusa. A parte importante dessa saída é a que mostra os pesos que o *Weka* determinou para cada coluna. No nosso exemplo, tivemos a seguinte saída:

$$\begin{aligned}
 qtde = & -6.6417 * numeroMes=7,11,9,5,6,2,12,10,3,4,1 + \\
 & 4.3163 * numeroMes=11,9,5,6,2,12,10,3,4,1 + \\
 & 2.7932 * numeroMes=5,6,2,12,10,3,4,1 + \\
 & -7.2352 * numeroMes=2,12,10,3,4,1 + \\
 & 4.5444 * numeroMes=12,10,3,4,1 + \\
 & 3.7422 * numeroMes=10,3,4,1 + \\
 & -4.0199 * numeroMes=4,1 +
 \end{aligned}$$

$$\begin{aligned}
& -10.7287 * \text{numeroTipoMercado}=4,3,1 + \\
& 20.6859 * \text{numeroTipoMercado}=1 + \\
& -3.6365 * \text{numeroFamiliaProduto}=2,6,3,1,4 + \\
& 14.8499 * \text{numeroFamiliaProduto}=3,1,4 + \\
& 3.0704 * \text{numeroFamiliaProduto}=1,4 + \\
& 9.9735 * \text{numeroFamiliaProduto}=4 + \\
& -10.2595 * \text{numeroMarca}=5,2,1,4,6 + \\
& 4.2997 * \text{numeroMarca}=2,1,4,6 + \\
& -35.4625 * \text{numeroMarca}=1,4,6 + \\
& 4.3159 * \text{numeroMarca}=4,6 + \\
& 43.5883 * \text{numeroMarca}=6 + \\
& -0.1238 * \text{precoUnitario} + \\
& 0.0119 * \text{totalMercadorias} + \\
& 23.1343
\end{aligned}$$

Podemos interpretar o que o Weka quis dizer com esses valores. Para a variável “*numeroMes*”, ele nos diz para diminuirmos 7.2352 da variável “*qtde*” caso o “*numeroMes*” seja 2, 12, 10, 3, 4 ou 1. Caso seja 12, 10, 3, 4 ou 1, somamos 4.5444 e assim por diante. O Weka nos disse também que as variáveis “*precoUnitario*” e “*totalMercadoria*” não são relevantes para o que queremos saber, já que possuem um peso menor.

Para validarmos essa equação gerada pelo *Weka*, separamos uma linha da base de dados para substituímos pelas variáveis. Escolhemos os valores 9, 2, 6, 2, 63, 63, onde 9 é o “*numeroMes*”, 2 o “*tipoMercado*”, 6 a “*familiaProduto*”, 2 o “*numeroMarca*”, 63 o “*precoUnitario*” e 63 o “*totalMercadoria*”. Aplicando os pesos, temos o seguinte cálculo:

$$\begin{aligned}
qtde = & -6.6417 * 1 + 4.3163 * 1 + 2.7932 * 0 + -7.2352 * 0 + 4.5444 * 0 + 3.7422 * 0 + - \\
& 4.0199 * 0 + -10.7287 * 0 + 20.6859 * 0 + -3.6365 * 1 + 14.8499 * 0 + 3.0704 * 0 + 9.9735 * 0 + \\
& -10.2595 * 1 + 4.2997 * 1 + -35.4625 * 0 + 4.3159 * 0 + 43.5883 * 0 + -0.1238 * 63 + 0.0119 * \\
& 63 + 23.1343
\end{aligned}$$

Como resultado, temos 4.1629. Já na base de dados, para essa venda a variável “*qtde*” contém 1.

Podemos realizar diversos testes com essa equação, tomando sempre como base qualquer linha da nossa base de dados. Para as variáveis que são classes definimos 1 ou 0 na equação, 1 caso o valor esteja no conjunto de dados definidos pelo *Weka* ou 0 caso ele não esteja.

Ex: “*numeroMes*” igual a 12 então multiplicamos 4.5444 por 1, já que 12 está no conjunto de dados “*numeroMes*=12,10,3,4,1”.

Conclusão

Ao final da soma de todos os pesos referentes a um exemplo encontramos um resultado próximo da quantidade de determinado produto, que faz parte de uma determinada família de produtos, que possui uma determinada marca, que foi vendido em um determinado tipo de mercado, em um determinado mês.

A regressão linear faz uma média da “*qtde*” baseada nos dados passados como peso e o resultado gerado é aproximado, como no nosso exemplo que temos um resultado de 4 produtos vendidos, entretanto na linha da nossa base de dados temos apenas 1 produto. Podendo assim algumas vendas possuir produtos abaixo ou acima da media, mas com valores sempre próximos.

Vale ressaltar também, que devido a base de dados ser referente a vendas realizadas, essa previsão tende a ser menos precisa, podendo então ter valores fora da área proposta pelo *Weka*.

Referências Bibliográficas

<http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>

Anexos

1. Legenda

Meses:

1 - jan/11

2 - fev/11

3 - mar/11

4 - abr/11

5 - mai/11

6 - jun/11

7 - jul/11

8 - ago/11

9 - set/11

10 - out/11

11 - nov/11

12 - dez/11

Vendedores:

24 - ADRIANA VIANNA
43 - AGUIA REPRESENTACOES
28 - AILTON LUIS
44 - ALESSANDRA GARCIA
20 - ANDREA
19 - CAROLINA NOBREGA
22 - CESAR AUGUSTO DEA
40 - DAYANE ROBERTA
25 - DIEGO TAVARES
36 - DIONNE CRISTINA
13 - DOREL
9 - ELIAS CAMPANHA
50 - FABRICIO WESTFAL
11 - GILSON RIBEIRO
14 - GUILHERME PAVANI
8 - ISAAC RIBEIRO
41 - JAMES CLAUS
18 - JOSE CARLOS
29 - JOSE GERALDO CAROL
34 - JOSE GERALDO DOREL
33 - JOSE MOACYR
26 - JOSUE ORNELAS
6 - LUIZ FILHO
5 - MARCOS ANTONIO
16 - MARCOS SILVA
47 - MARCUS SERRA
2 - MARIA DAS GRACAS
49 - MATEUS SILVA
39 - MAURO LIMA
42 - ROBSON
12 - ROGER SOUZA
38 - RONY JOSE SCARABELLI

3 - RUBIA

31 - STELIO FERRAZ

27 - SUELI CAROL

35 - SUELI DOREL

Market type (tipoMercado):

1 - MASS MARKET

2 - SPECIALIST SHOP

3 - DISTRIBUITOR

4 - DOTCOM

Product Families (familiasProduto):

1 - CAR SEATS

2 - PLAYARD

3 - OTHERS

4 - STROLLERS

5 - TRAVEL

6 - WALKER

brand (marcas):

1 - VOYAGE

2 - COSCO

3 - BEBE CONFORT

4 - STILLO

5 - SAFETY FIRST

6 - DISNEY

Produtos:

1 - CV30000VOY

2 - YY060CSC

3 - CV3000II0CSC

4 - YY061CSC

5 - C661CSC

6 - 27301BBC

7 - CV20000VOY

8 - B-81STL
9 - 12351BBC
10 - 13041BBC
11 - 26801BBC
12 - 85211BBC
13 - CD200TS0CSC
14 - 85201SF1
15 - H6001CSC
16 - 27551BBC
17 - G750H1SF1
18 - LS20571STL
19 - 416720CSC
20 - CV20000CSC
21 - 86891BBC
22 - 16711BBC
23 - 86881BBC
24 - C326M0CSC
25 - I000010VOY
26 - 28231BBC
27 - I000020VOY
28 - 16061BBC
29 - 13021BBC
30 - 13011BBC
31 - 1235B1BBC
32 - 17211BBC
33 - 416730CSC
34 - 24541BBC
35 - 14581BBC
36 - LM2041CSC
37 - CV40000CSC
38 - 417670CSC
39 - 12601SF1
40 - 15231BBC
41 - B-81VOY
42 - C66T1CSC

43 - 87660BBC
44 - HC0511SF1
45 - 11341SF1
46 - LS20571VOY
47 - CV30010VOY
48 - 27691BBC
49 - WA0331SF1
50 - LM2041VOY
51 - WA0111CSC
52 - YY060DIS
53 - 113891SF1
54 - 417800CSC
55 - C681VOY
56 - SC9001CSC
57 - LS20581VOY
58 - SC4081SF1
59 - TLDB1006041VOY