MATHEUS CÂNDIDO DE OLIVEIRA

ATIVIDADE PRÁTICA

DISCIPLINA: ALGORITMOS PARA CIÊNCIA DE DADOS

CARMO DO RIO VERDE, 2021.

EXERCÍCIO 1:

- **Você terá que analisar as características dos clusters gerados e relacioná-los com as regras geradas pelo apriori, descreva isso em um relatório e com as regras e clusters gerados.**

PROPOSTA DE RESOLUÇÃO:

- **O primeiro passo para chegarmos à conclusão do exercício, será preciso, de início, criarmos os experimentos que geraram os clusters a serem analisados mais adiante.**
- **Utilizando o software Weka, fiz uma sequência de experimentos usando os seguintes valores de clusters: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 25, 50, 75, 100. Para os valores supracitados, achei os valores que estão na tabela a seguir.**

Clusters	Erros RMS
1	2590
2	1861
3	1293
4	1010
5	909
6	760
7	712
8	594
9	674
10	628
25	286
50	192
75	134
100	106

Figura 1 – tabela comparando os valores de clusters x erros RMS.

- **Com base nos valores mostrados na tabela pude gerar o seguinte gráfico:**

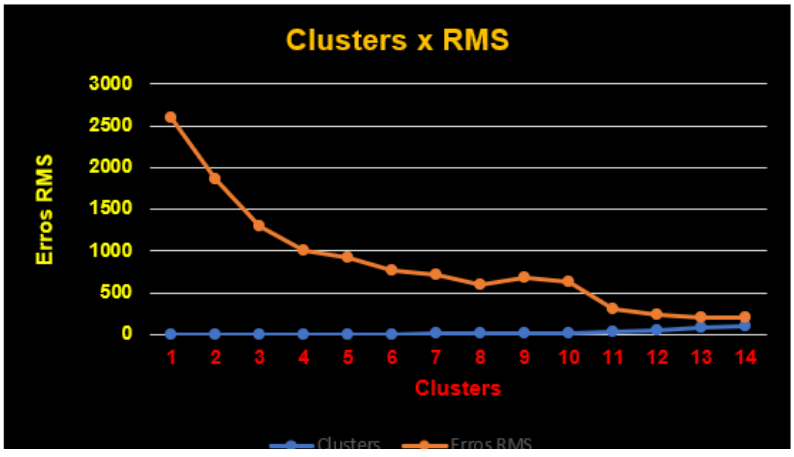

Figura 2 – gráfico comparando os valores x erros RMS.

- **Para continuar com o estudo para chegar ao objetivo de comparar as características das regras geradas pelo apriori com os resultados apresentados pelos clusters, irei usar o gráfico para escolher um cluster para ser nosso "objeto de estudo".**
- **Analisando o gráfico gerado, e baseando-se no que foi apresentado ao longo da disciplina, estamos procurando pelo "joelho da curva". Que neste caso me parece ser o ponto que representa o *cluster de valor 7.***
- **Votando ao Weka, e procurando pelo cluster 7 e gerando o gráfico que o próprio software nos permite criar pode ser observadas algumas características bastante interessantes.**

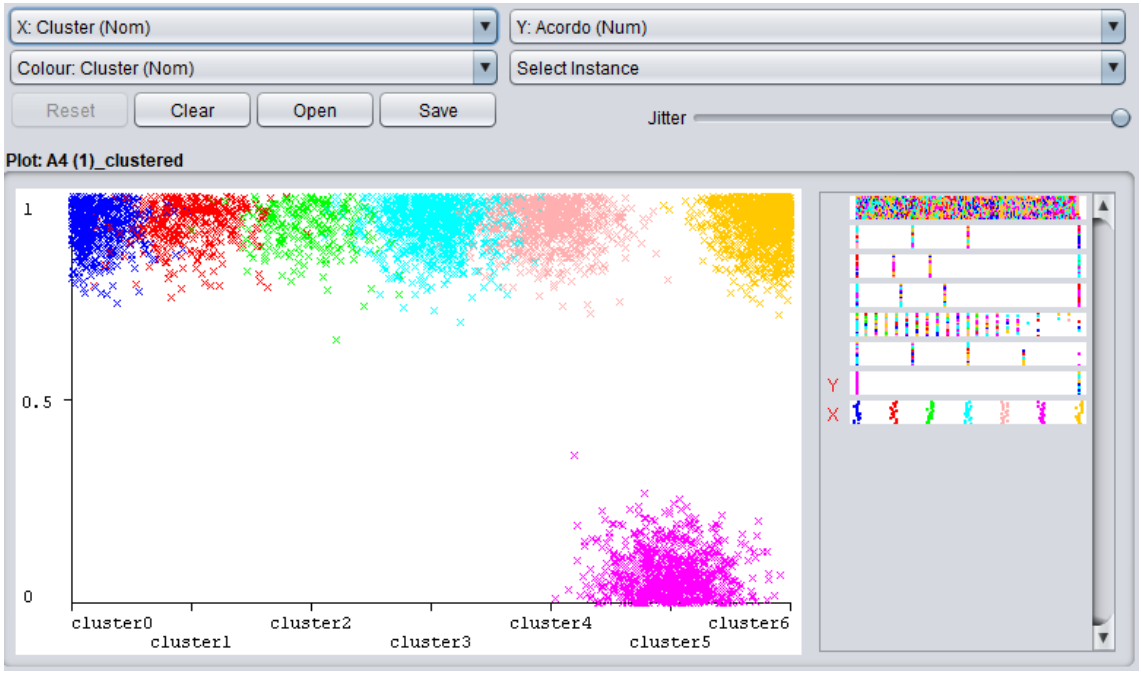

Figura 3 – gráfico gerado no Weka.

- **Com 7 clusters pude observar uma certa regularidade nos padrões gerados, pelo gráfico, tendo em vista a tendência de fechamentos de acordos.**
- **Apenas o cluster 5 apresenta maior tendência de não fechamento de acordos.**
- **Acredito que podendo ver de forma tão clara a "exceção" fica mais fácil conseguir fazer previsões e planejamentos acerca das estratégias mais eficazes para se conseguir fechar acordos.**
- **Agora já tendo escolhido um número de clusters para dar seguimento a conclusão do exercício, é hora de decodificar as informações descobertas no Weka e fazermos a comparação, de fato.**

Attribute	Full Data (4773.0)	0 (534.0)	1 (384.0)	2 (236.0)	3 (869.0)	4 (759.0)	5 (867.0)	6 (1124.0)
Idade	1.7461	4	2.1432	4	1.0311	0.2688	1.9146	1.4867
Atraso	2.4588	1.2772	0.7969	6	6	0.8814	2.045	1.4911
Valor	1.4718	1.2247	5	0.4576	0.3751	1.2648	2.0092	1.1699
CONTATO	4.7591	4.161	5.8203	3.9958	4.7135	4.8577	5.2549	4.427
EFETIVO	1.1672	1.1685	1.2422	1.2246	1.1749	1.1291	1.113	1.1904
Acordo	0.8184	1	1	1	1	1	0	1

```
Time taken to build model (full training data) : 0.06 seconds

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

Clustered Instances

0       534 ( 11%)
1       384 (  8%)
2       236 (  5%)
3       869 ( 18%)
4       759 ( 16%)
5       867 ( 18%)
6      1124 ( 24%)
```

Figura 5 – atributos dos clusters gerados.

CLUSTER	IDADE	ATRASO	VALOR	% – INSTÂNCIAS
6	26 a 35	15 a 30	200 a 500	24
3	26 a 35	>120	0 a 200	18
5	36 a 45	31 a 120	500 a 1000	18
4	0 a 25	15 a 30	200 a 500	16
0	>45	15 a 30	200 a 500	11
1	36 a 45	15 a 30	>1000	8
2	>45	>120	0 a 200	5

Figura 4 – informações "traduzidas".

- **Por fim, tendo feita a decodificação dos atributos trazidos pelo cluster gerado, podemos relacioná-los com as regras criadas com o apriori.**
- **Segue a comparação que fiz entre os modelos, nela mostro reproduzo a regra e faço um comentário de acordo com aquilo que creio ser pertinente:**
  - **1ª regra no apriori – Primeiramente débitos com valores de 0 a 200, independente da faixa de atraso; ->Num primeiro momento, olhando para minha planilha, essa regra até se mostra interessante, possível de ser aplicada no nosso modelo de estudo. Pois, o grupo dos débitos de 0 a 200 reais representa 23% do total de possibilidades de acordo. No entanto, uma coisa que me parece ser um ponto contra essa regra é a quantidade de tempo de atraso. Em ambos os clusters em que o débito é de 0 a 200 reais o atraso é de mais de 120 dias.**

- **2ª regra no apriori – Valores entre 200 e 500, com atraso de 31 a 120 dias; ->Neste caso o problema que observo é estritamente relacionado ao tempo de atraso. Os três clusters compreendidos dentro dessa faixa de débito de 200 a 500 reais, correspondem a 51% da base. E o tempo de atraso em todos os casos está entre 15 e 30 dias, portanto uma espera de até 120 dias, não me parece nada viável. Se levarmos em consideração que 51% das possibilidades de acordo "exigem" um tempo de espera para serem pagas muito menor.**
- **3ª regra no apriori – Faixa etária entre 0 e 25 anos; ->Esta regra me parece aplicável, no nosso modelo de estudo. Visto que o grupo compreendido entre as idades de 0 a 25 anos apresentam 16% das possibilidades de acordos, o que é um bom número. E ainda, o período de atraso de 15 a 30 dias, parece razoável.**
- **4ª regra no apriori – Depois débitos acima de 120 dias; ->De acordo com os valores exibidos na tabela, os débitos acima de 120 dias são os de 0 a 200 reais. Novamente o fato deste grupo compreender 23% das possibilidades de acordo, são um ponto a favor. Mas não acho que colocaria esta regram na quarta posição. Justamente pelo tempo de atraso e pelo valor do débito, portanto, acho que ela se aplica ao modelo desenvolvido na atividade, mas na quinta posição, talvez.**
- **5ª regra no apriori – Depois faixa etária entre 26 a 35 anos; ->Este grupo de idade abriga 42% das possibilidades de contrato da base. Portanto, creio que deva ser levado em consideração. Mas, o cluster 3, apresenta atraso de >120 dias e valor entre 0 a 200 reais, e sozinho abrange 18% das possibilidades de acordo. Deste modo, não deve ser desprezado. Mas é o ponto fraco para a aplicação desta regra.**
- **6ª regra no apriori – Depois o restante total da base. ->Creio que até aqui toda a base analisada já tenha sido explorada.**

- **E para concluir vou listar aqui as regras que eu acho que seriam mais vantajosas para o modelo que criamos.**
  - **1ª regra – Primeiramente os débitos em que o índice de atraso esteja entre 15 a 30 dias, independente do valor.**
  - **2ª regra – Valores entre 500 a 1000 reais e prazo entre 31 a 120 dias.**
  - **3ª regra – Valores entre 0 e 200 reais com prazos de >120 dias.**
  - **4ª regra – Todo o resto da base, caso haja.**

EXERCÍCIO 2:

- **No segundo experimento você deverá usar a base de dados "IrisDataSet" no arquivo "iris.csv" bastante conhecida para experimentos e clustering. Você deverá executar o experimento com o Kmeans no Weka e verificar qual é o melhor número de clusters para o modelo gerado, utilizando o erro RMS com um gráfico, como foi feito na unidade 6 com a base de dados "A".**

PROPOSTA DE RESOLUÇÃO:

- **Para resolver este exercício o primeiro passo será ir até o Weka, abrir a base e criar um experimento para gerar diferentes números clusters.**
- **Seguindo o "padrão do exercício" anterior, usei os seguintes números de clusters: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 25, 50, 75, 100. Para esses números de clusters, obtive os valores de erro RMS exibidos na tabela a seguir:**

Clusters	Erros RMS
1	141
2	62
3	7
4	6
5	6
6	6
7	5
8	4
9	4
10	4
25	1
50	0
75	0
100	0

Figura 6 –tabela de comparação entre o número de clusters e a quantidade de erros RMS.

- **Esta tabela feita no Excel será a base para que possamos fazer o gráfico que nos indicará o melhor número de clusters para ser usado.**

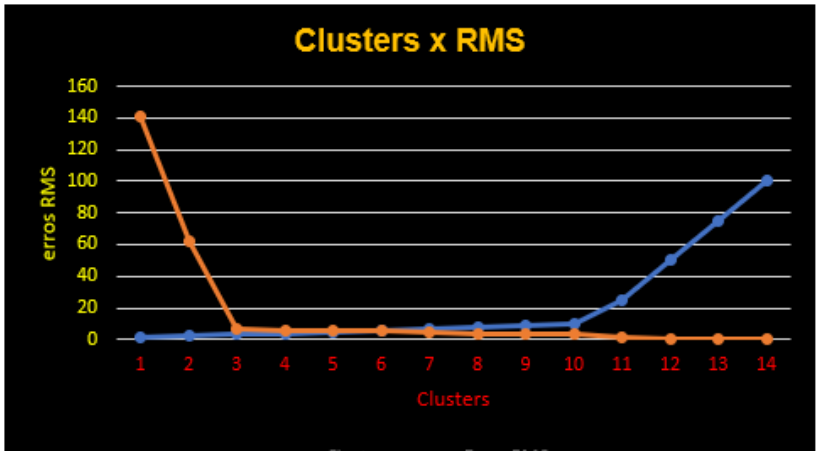

Figura 7 – gráfico de comparação entre o número de clusters e a quantidade de erros RMS

- **Com base na análise do gráfico acima pode-se concluir que o "joelho da curva" é encontrado no ponto 3, que indica um número de 7 erros RMS e a quantidade de 3 clusters.**
- **Sendo assim, posso dizer que para o modelo proposto, a quantidade de clusters ideal para se trabalhar é 3.**