UNIVERSIDADE ESTADUAL DE CAMPINAS

Curso de Bacharelado de Sistemas de Informação

Guilherme Masao Tsuyukubo 217250

Igor da Silva Gouvêa 217956

Leonardo Alberto da Ponte 220007

MINERAÇÃO DE DADOS: O USO DO MÉTODO DE ASSOCIAÇÃO E AGRUPAMENTO NUMA BASE

São Paulo

2019

Guilherme Masao Tsuyukubo 217250

Igor da Silva Gouvêa 217956

Leonardo Alberto da Ponte 220007

MINERAÇÃO DE DADOS: O USO DO MÉTODO DE ASSOCIAÇÃO E AGRUPAMENTO NUMA BASE

Trabalho apresentado para avaliação do rendimento escolar da disciplina Tópicos em Computação e Informática do curso de Bacharelado de Sistemas de Informação, ministrada pela Profa. Dr. Ana Estela Antunes da Silva.

São Paulo

2019

SUMÁRIO

1 INTRODUÇÃO

A mineração de dados é a prática de examinar dados que já foram coletados, utilizando diversos algoritmos a fim de obter determinados padrões e gerar informações úteis.

A mineração surgiu com o crescimento da economia da informação que gerou a indústria 4.0 no qual Big Data é gerada através da internet, inteligência artificial e machine learning. Essas informações podem ser usadas em diversas áreas desde saúde até dados internos de organizações.

Ela começou a ser muito utilizada devido ao grande crescimento da quantidade de dados coletados e armazenados e isso associado ao aumento na qualidade e na quantidade de ferramentas disponíveis possibilitou criar associações entre os dados. Podem ser citados alguns algoritmos como o agrupamento, classificação e associação.

Um exemplo que é muito citado é o caso de uma rede varejista (Walmart) que através do seu armazenamento de dados descobriu uma associação de venda de fraldas e cerveja, no caso os compradores eram homens e quando iam buscar as fraldas acabavam levando algumas latinhas e com o uso dessa informação a rede acabou colocando os dois produtos perto fazendo com que a venda dos dois tenham disparado.

Outro exemplo que pode ser citado é de uma empresa de telefonia norte americana chamada Sprint que através de seus dados conseguiu prever se um cliente iria desistir da companhia dentro de um período de dois meses e com isso a empresa investiu em um marketing agressivo conseguiu evitar a desistência de 120.000 clientes e consequente evitou uma perda de 35 milhões de dólares.

2 OBJETIVO

O objetivo do presente trabalho foi a execução de algoritmos de mineração de dados sobre uma base que possui como principal atributo a demanda hídrica.

3 TAREFAS DE MINERAÇÃO DE DADOS

Três tarefas podem ser citadas em mineração:

Agrupamento: seria basicamente a ideia de organizar características similares de alguns objetos como cor, forma, peso , velocidade, temperatura etc.

Classificação: tem como objetivo criar rótulos de classes que buscam prever qual será a saída para os registros desconhecidos. Um exemplo que pode ser citado é classificação dos registros de cartões de créditos fraudulentos ou não.

Associação: cada registro das bases de dados são chamadas de transação e assim essas bases são chamadas de transnacionais. É possível criar um exemplo analisando uma compra em um supermercado onde se pode criar associações entre as vendas de determinados objetos, no caso podemos referenciar o caso do Walmart que descobriu a associação entre cerveja e fralda.

4 DESCRIÇÃO DO PROBLEMA

O problema do trabalho relaciona-se a área de ambiental, mais especificamente, a parte hídrica que, apesar de ser um recurso renovável, encontra-se com os seus números diminuindo de forma assustadora, visto que o processo de renovação desse recurso é mais lento que o consumo humano.

Muitos estudos sobre a área demonstram que se o consumo humano hídrico não diminuir, num futuro próximo teremos problema de escassez desse recurso hídrico.

Isso é comprovável a partir de dados que demonstram que 97,2% da quantidade de água disponível, encontra-se nos oceanos, ou seja, é salgada e, portanto, não pode ser utilizada para o consumo humano. Dos 2,8% restantes, mais de dois terços encontra-se na forma de geleira, o que, novamente, impossibilita sua ingestão. E, de apesar de 0,4% apenas da água estar disponível para o uso, a demanda por ela não para de crescer.

A partir dessas informações, o grupo, utilizou de uma base de dados fornecida pela Profa. Dr. Lubienska Cristina Lucas Jaquiê Ribeiro para encontrar regras que pudessem ser utilizadas para minimizar esses problemas, no caso em uma região de Franca, relacionados a uma eventual ausência hídrica no futuro.

E, apesar de, a maior porção da utilização da água ser na área da agropecuária, 70% segundo a Organização das Nações Unidas para a Agricultura e Alimentação, a diminuição na utilização doméstica, ainda assim colabora com a preservação de um recurso tão necessário para a sobrevivência humana.

4.1 Explicação dos atributos

A base de dados utilizada pelo grupo é composta por 12 atributos que, no total, somam 12.576 instâncias. Os atributos são: Demanda, dia, mês, ano, hora, dia da semana, feriado, estação, temperatura, umidade do ar, velocidade do vento e precipitação.

A ‘demanda’ é calculada em L/s e, assim como mostrado na medida, demonstra a quantidade de Litros de água utilizada por segundo na região. Isso mostra o quanto a região necessita do recurso durante o período da análise. E, além disso, esse atributo é numérico.

Os atributos ‘dia’, ‘mês’, ‘ano’ e ‘hora’ demonstram a data e o momento em que o registro foi feito. Outro ponto é que esses registros são numéricos. No caso da ‘hora’, este é medido em horas.

O ‘dia da semana’, assim como os anteriores, demonstra que porção da semana o registro foi feito. Além disso, este atributo pode ser considerado nominal. Este pode ser: 1 - ‘domingo’, 2- ‘segunda’, 3- ‘terça’, 4- ‘quarta’, 5- ‘quinta’,6- ‘sexta’, 7- ‘sábado’.

O ‘feriado’, que é um atributo binário, demonstra se a data é uma comemoração especial ou é apenas um dia normal da semana. Este atributo pode ser: ‘0’ que se refere ao dia normal e ‘1’ que demonstra se a data é feriado.

A estação do ano é um atributo numérico divide-se em: ‘1’ - verão, ‘2’ - outono, ‘3’ - inverno , e, por fim, ‘4’ - primavera.

A ‘temperatura’ é um atributo numérico que demonstra o nível de agitação das moléculas na região. Este atributo é medido em °C.

A ‘umidade do ar’, ‘velocidade do vento’ e ‘precipitação’, são todos atributos numéricos referentes a condição climática no momento do registro na região. Estes podem ser medidos em, respectivamente, ‘%’, ‘m/s’ e ‘mm’.

4.2 Explicação do Pré-Processamento

Para a execução de qualquer tarefa de mineração de dados, todas as bases devem ser submetidas a um processo conhecido como Pré-Processamento. Alguns exemplos de métodos que são utilizados durante essa fase são: correção de dados ausentes e limpeza de dados inconsistentes.

A correção de dados ausentes atua sobre a falta de dados dentro de uma base de dados que, em diversos casos, é denotada como '?'. A fim de imputar um dado plausível que possa ser inserido nesse local, algumas técnicas são sugeridas, algumas delas são: ignorar o dado ausente, hot-deck ou usar a média ou moda.

O primeiro método tem como principal função a remoção do objeto que possui o dado ausente. Esta forma de imputação não é indicada em bases de dados pequenas pois, se o número de dados ausentes é alto,gera-se uma diminuição significativa nos dados analisados.

O segundo é o método hot-deck em que imputa-se um valor de um objeto similar ao ausente, este sendo aleatoriamente escolhido.

Por fim, o uso da média e da moda também pode ser utilizado para a imputação de dados ausentes, nesse caso, usa-se a média(no caso de atributos numéricos) e da moda(no caso de atributos nominais). O principal defeito desse método é que este desconsidera a diferença entre as classes, apesar deste ser o mais utilizado.

Já em relação a limpeza de dados inconsistentes temos que, dados inconsistentes podem influenciar na análise dos dados pelos algoritmos. Um erro muito comum é o de, dentro dos dados coletados, alguns dados estarem em outros sistemas de medida, o que entra nesse caso. Para a solução desse problema, basta-se fazer a conversão das medidas para um sistema único.

4.3 Explicação da escolha da tarefa para o problema

No caso da nossa base de dados, foi nos dado a função de encontrar algum tipo de relacionamento ou regra entre os atributos que pudesse explicar melhor a demanda hídrica de uma região que fica nos arredores do aeroporto de Franca.

Apesar de a base de dados não se encontrar na forma perfeita para a aplicação dessa tarefa, o grupo decidiu utilizar-se do método de associação pois, através deste, regras seriam feitas e, consequentemente, poderiam conter explicações de relações entre os atributos.

Além disso, a fim de comparação e visto que a base conta, em sua maioria, com atributos numéricos, decidiu-se utilizar também o método de agrupamento.

5 METODOLOGIA

5.1 Método utilizado durante o processo de Pré-Processamento

Durante o processo de Pré-Processamento da base, o método de imputação de dados foi o aspecto de ignorar os dados faltantes pois, por apresentar um número de instâncias muito extenso, a ausência desses dados não causaria impacto tão relevante na análise da base. Aliado a isso, tem-se o fato de que, quando colocada no Weka a base não apresentou ausência de dados visto que, os dados ausentes na base não eram simbolizados com ‘0’ ou ‘?’, o que fez com que o software considerasse que não houvesse dados faltantes.

FIGURA 1 - Dados faltantes (passa-se do mês 4 ao 6)

21.05	25	4	2013	18	5	0	2	24	53	3	0
21.25	25	4	2013	19	5	0	2	24	53	3	0
21.25	25	4	2013	20	5	0	2	24	53	3	0
18.82	25	4	2013	21	5	0	2	24	53	3	0
15.27	25	4	2013	22	5	0	2	24	53	3	0
12.72	25	4	2013	23	5	0	2	24	53	3	0
11.77	29	6	2013	0	7	0	3	19.4	84	0	0
9.27	29	6	2013	1	7	0	3	19.4	84	0	0
7.42	29	6	2013	2	7	0	3	19.4	84	0	0
6.77	29	6	2013	3	7	0	3	19.4	84	0	0
6.4	29	6	2013	4	7	0	3	19.4	84	0	0
6.17	29	6	2013	5	7	0	3	19.4	84	0	0
7.2	29	6	2013	6	7	0	3	19.4	84	0	0
11.31	29	6	2013	7	7	0	3	19.4	84	0	0

FONTE: Keynote

Apesar disso, como todos os atributos que se encontravam na base são numéricos e, algoritmos de associação como o “A PRIORI” não aceitam apenas esse tipo de atributo, foi necessária uma conversão para uma base transacional, ou seja, todos os atributos foram discretizados durante esse processo.

Em relação a isso, a discretização dos atributos ‘dia’, ‘mês’, ‘ano’ e ‘hora’ se tornaram inviáveis, visto que criamos “classes” para os atributos e, no caso desses citados, essa classificação não seria possível.

No processo de Pré-Processamento, os atributos ‘demanda’, ‘temperatura’, ‘umidade do ar’, ‘velocidade do vento’ e ‘precipitação’ foram discretizados da seguinte forma:

1. Calculou-se o desvio padrão e a média dos atributos.
2. Na média somou-se e diminui-se o desvio padrão.
3. Os números abaixo dessa subtração foram considerados baixos.
4. O intervalo entre a soma e a subtração foram considerados médio.
5. Por fim, os números que encontram-se acima da soma foram considerados altos, todos variando dependendo da medida utilizada para a classificação.

Os outros atributos foram substituídos pelos nomes que representam os números a que estavam atribuídos. Com isso os atributos foram classificados da seguinte forma na base transacional:

Demanda - { “LOW”, “MEDIUM”, “HIGH}
Dia da Semana - { “SEGUNDA”, “TERÇA”, “QUARTA”, “QUINTA”, “SEXTA”, “SÁBADO”, “DOMINGO” }
Feriado - { “NORMAL”, “FERIADO”}
Estação do Ano - { “VERÃO”, “OUTONO”, “INVERNO”, “PRIMAVERA”}
Temperatura - { “COLD”, “MILD”, “HOT”}
Umidade do ar - { “LOW”, “MEDIUM”, “HIGH”}
Velocidade do vento - { “CALM”, “LIGHT”, “STRONG”}
Precipitação - { “NULL”, “LOW”, “HIGH”}

FIGURA 2 - IMAGEM DA BASE ANTES DA MUDANÇA

21.05	25	4	2013	18	5	0	2	24	53	3	0
21.25	25	4	2013	19	5	0	2	24	53	3	0
21.25	25	4	2013	20	5	0	2	24	53	3	0
18.82	25	4	2013	21	5	0	2	24	53	3	0
15.27	25	4	2013	22	5	0	2	24	53	3	0
12.72	25	4	2013	23	5	0	2	24	53	3	0
11.77	29	6	2013	0	7	0	3	19.4	84	0	0
9.27	29	6	2013	1	7	0	3	19.4	84	0	0
7.42	29	6	2013	2	7	0	3	19.4	84	0	0
6.77	29	6	2013	3	7	0	3	19.4	84	0	0
6.4	29	6	2013	4	7	0	3	19.4	84	0	0
6.17	29	6	2013	5	7	0	3	19.4	84	0	0
7.2	29	6	2013	6	7	0	3	19.4	84	0	0
11.31	29	6	2013	7	7	0	3	19.4	84	0	0

Fonte: Keynote

FIGURA 3 - IMAGEM DA BASE DEPOIS DA MUDANÇA

DEMANDA	DIA DA SEMANA	UMIDADE	FERIADO	ESTAÇÃO DO ANO	TEMPERATURA	VELOCIDADE DO VENTO	PRECIPITAÇÃO
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
LOW	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
LOW	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
LOW	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
LOW	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
LOW	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
HIGH	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
HIGH	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	LIGHT	NULL
HIGH	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	STRONG	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	STRONG	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	STRONG	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	STRONG	NULL
MEDIUM	QUARTA	MEDIUM	NORMAL	OUTONO	COLD	STRONG	NULL

Fonte: EXCEL

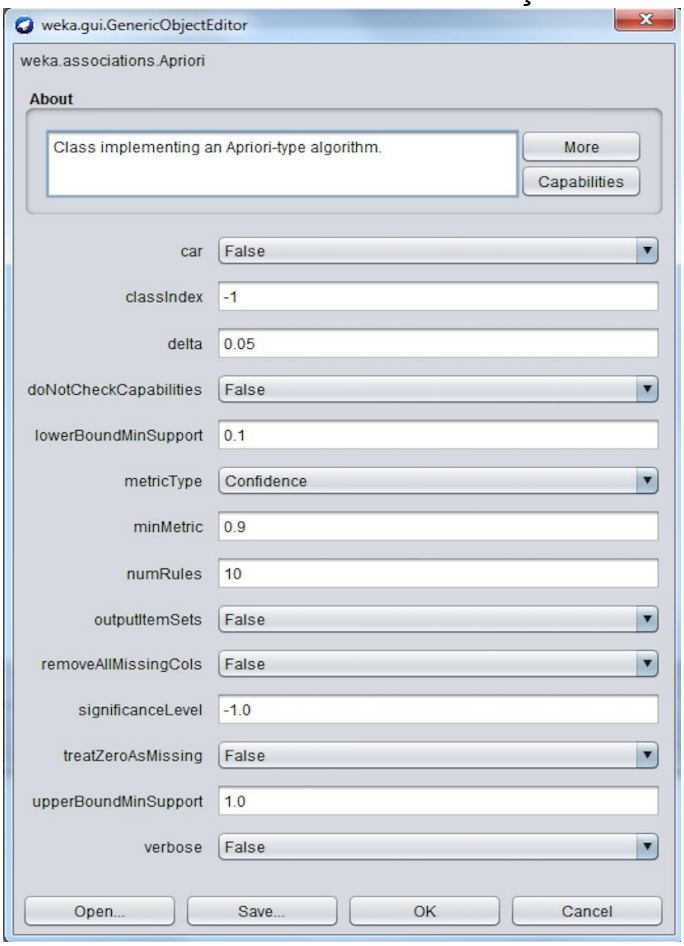
5.2 Método utilizado durante a tarefa de mineração

A partir da utilização da base de dados transacional criada durante o processo de Pré-Processamento, utilizou-se o método de associação, nesse caso, o algoritmo “A PRIORI” para a criação de regras que pudessem ser analisadas. Com esse algoritmo, o objetivo foi de encontrar relações que pudessem auxiliar na diminuição na demanda hídrica na região ou que, simplesmente, pudesse explicar essa demanda para que, a partir dessas regras, medidas possam ser criadas para se ter uma demanda menor.

Em relação ao algoritmo, ele foi executado com diversas confianças e suportes para encontrar, dentre diversas regras resultantes, algumas que trouxessem informações interessantes para a pesquisa.

Outro ponto é que, por conta de ser um atributo binário, o ‘feriado’ foi retirado em algumas iterações, visto que a ocorrência de ‘normal’ era alta o que refletia nas regras, visto que todas resultaram- em ‘feriado’ = ‘normal’.

IMAGEM 4 - IMAGEM DAS MUDANÇAS DOS PARÂMETROS (ASSOCIAÇÃO)



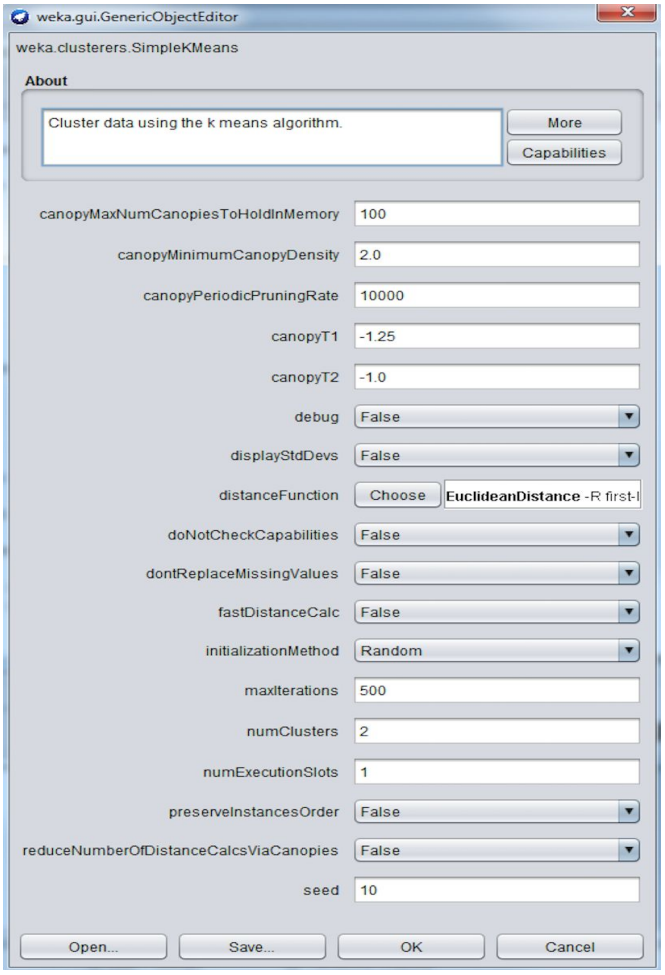
FONTE: WEKA

Outro método utilizado, a fim de possibilitar comparações foi o de agrupamento, nesse caso, mais especificamente o “K-MEANS” visto que este era um método secundário, simples e de conhecimento nosso por conta das aulas.

Durante a iteração por esse algoritmo foi utilizada a base original, ou seja, a com maioria numérica. Entretanto, os atributos “dia”, “mês”, “ano”, “hora” e “feriado” foram relevados pois influenciavam no resultado por conta do cálculo da distância realizada por esse algoritmo.

Nesse método, também foram realizadas diversas interações mas, nesse caso, o número de clusters foi alterado entre as iterações para a análise da diferença da execução com mais grupos.

IMAGEM 5 - IMAGEM DA MUDANÇA DOS PARÂMETROS (AGRUPAMENTO)



FONTE: WEKA

5.3 Métodos utilizados na análise dos resultados

Para a análise dos resultados, foi feita uma pesquisa da área para uma análise com maior propriedade dos resultados. Além disso, no método de associação analisou-se os itens pela confiança e pelo suporte. O grupo buscou regras que tivessem, a princípio, baixo suporte e alta confiança pois, a partir disso, têm-se regras que podem ser consideradas interessantes, visto que essas regras apesar de ter pouca ocorrência, mas na maioria das vezes em que aparece está presente na base.

No caso do algoritmo de agrupamento, foi feita uma análise dos grupos, sem entretanto, o cálculo de possíveis coeficientes, como por exemplo, o de Silhueta. Buscou-se uma análise gráfica, porém o grupo não conseguiu localizar o local que imprimia os gráficos do agrupamento.

6 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

6.1 Método de Associação

Em relação a utilização do método de associação, buscou-se, através das medidas de confiança e suporte achar regras que mostrassem algum tipo de relação entre os atributos da base de dados. Como visto, o atributo objetivo, nesse caso, foi a ‘demanda’. Entretanto neste ponto teve-se um problema pois, no Weka, não foi possível escolher o consequente da regra.

Com isso, a análise foi feita a partir de dados que apresentavam o atributo ‘demanda’ no antecedente da regra. Apesar disso, alguns resultados encontrados podem ser considerados interessantes, como os apresentados a seguir:

IMAGEM 6 - SAÍDA DO ALGORITMO DE ASSOCIAÇÃO

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 25

Size of set of large itemsets L(2): 64

Size of set of large itemsets L(3): 43

Size of set of large itemsets L(4): 14

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

Best rules found:

 1. umidade=LOW 2358 ==> precipitacao=NULL 2005    <conf:(0.85)> lift:(1.1) lev:(0.02) [189] conv:(1.53)
 2. demanda=MEDIUM umidade=LOW 1545 ==> precipitacao=NULL 1310    <conf:(0.85)> lift:(1.1) lev:(0.01) [120] conv:(1.51)
 3. demanda=LOW umidade=MEDIUM 1875 ==> temperatura=MILD 1548    <conf:(0.83)> lift:(1.18) lev:(0.02) [232] conv:(1.71)
 4. estacao=PRIMAVERA umidade=MEDIUM 2916 ==> temperatura=MILD 2406    <conf:(0.83)> lift:(1.18) lev:(0.03) [360] conv:(1.7)
 5. estacao=INVERNO temperatura=MILD 3102 ==> precipitacao=NULL 2552    <conf:(0.82)> lift:(1.07) lev:(0.01) [163] conv:(1.3)
 6. demanda=HIGH 2010 ==> precipitacao=NULL 1652    <conf:(0.82)> lift:(1.07) lev:(0.01) [104] conv:(1.29)
 7. diaDaSemama=SEXTA 1728 ==> precipitacao=NULL 1413    <conf:(0.82)> lift:(1.06) lev:(0.01) [82] conv:(1.26)
 8. estacao=PRIMAVERA umidade=MEDIUM precipitacao=NULL 2171 ==> temperatura=MILD 1775    <conf:(0.82)> lift:(1.17) lev:(0.02) [252] conv:(1.63)
 9. temperatura=HOT 1992 ==> precipitacao=NULL 1626    <conf:(0.82)> lift:(1.06) lev:(0.01) [92] conv:(1.25)
10. estacao=VERAO 2327 ==> precipitacao=NULL 1892    <conf:(0.81)> lift:(1.06) lev:(0.01) [100] conv:(1.23)
```

Fonte: Weka

Nesse caso, ao se rodar o algoritmo em busca de regras de associação, com, nesse caso, regras com suporte entre 10 e 30% e confiança acima de 80%, obtivemos as seguintes regras:

A primeira diz que quando a demanda é média, e a umidade é baixa, temos que a precipitação é nula (confiança = 0.85). Com isso, podemos notar que,em muitos casos em que, na base a demanda encontra-se média e a umidade baixa, temos que a precipitação é nula. Explica-se esse fato ao ponto de que quando a umidade é baixa e a precipitação é nula, torna-se muito mais suscetível fazer as tarefas domésticas, o que justifica essa medição da demanda.

A segunda regra diz que quando a demanda é baixa e a umidade é média, a temperatura é amena (confiança = 0.83). Assim como na anterior, nota-se que na maior parte dos casos em que a umidade é média e a demanda é baixa, a

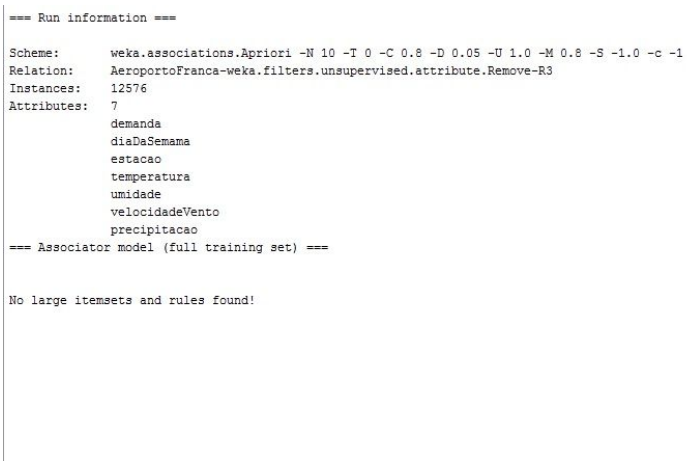
temperatura é amena. Uma hipótese para essa regra seria de que, por conta da umidade média, que não é elevada e, portanto, não passa uma sensação de desconforto quando está um pouco mais elevada, e por conta da temperatura ser amena, as pessoas tendem a sair mais de casa e, portanto, diminui a demanda de água na região, visto que é uma área majoritariamente residencial.

Por fim, a última regra se refere ao fato de que quando a demanda é alta, a precipitação é nula, ou seja, tem-se uma alta necessidade de água quando a não chove (confiança = 0.82). Isso pode ser explicado ao passo de que, por conta da precipitação ausente, serviços domésticos são de mais fácil resolução o que demanda uma maior utilização da água.

Todos essas regras citadas, foram descobertas a partir de um baixo suporte e uma alta confiança, ou seja, são regras muito interessantes, pois, apesar de aparecer poucas vezes na base de dados, sabe-se que na maior parte em que aparecem, a regra é cumprida.

Em contrapartida, regras de associação com alto suporte (acima de 0.8) e alta confiança (acima de 0.8) não foram encontradas durante a execução do algoritmo no Weka.

Imagem 7 - SAÍDA DO ALGORITMO DE ASSOCIAÇÃO



```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.8 -S -1.0 -c -1
Relation:     AeroportoFranca-weka.filters.unsupervised.attribute.Remove-R3
Instances:    12576
Attributes:   7
              demanda
              diaDaSemama
              estacao
              temperatura
              umidade
              velocidadeVento
              precipitacao
=== Associator model (full training set) ===


No large itemsets and rules found!
```

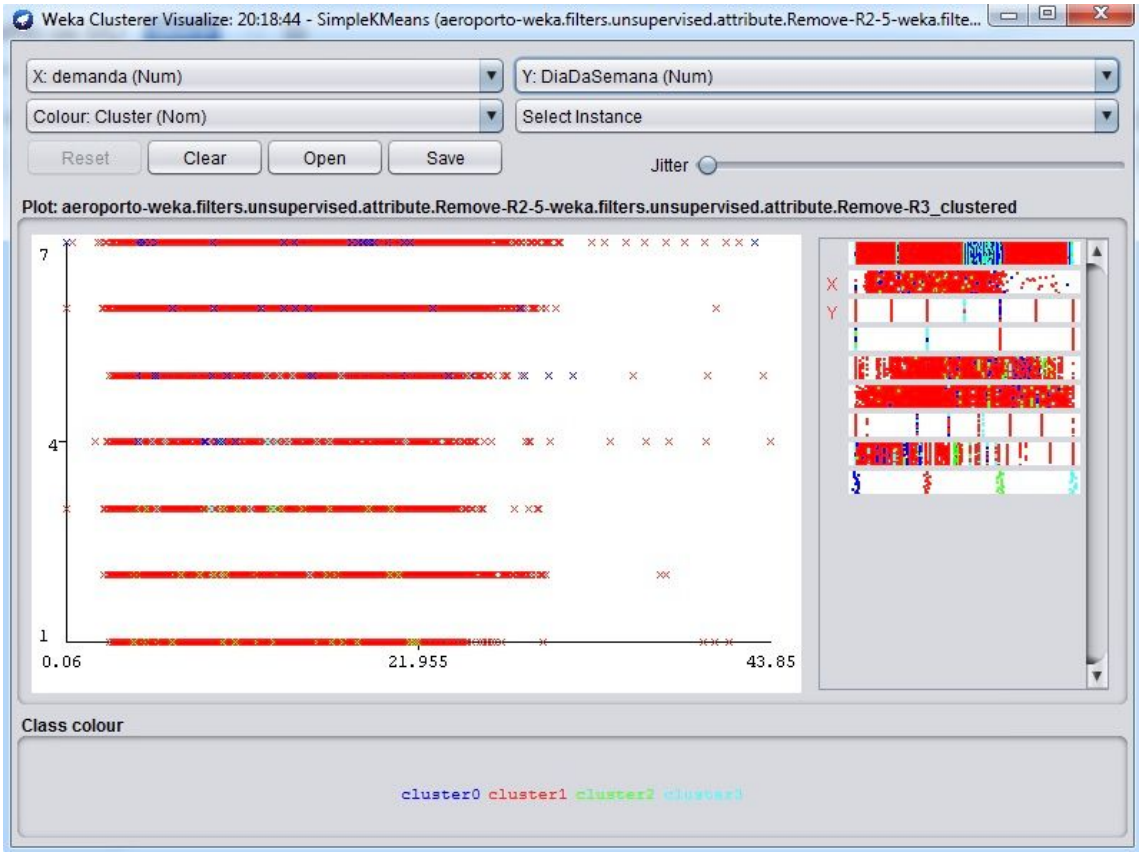
Fonte: Weka

6.2 Método do Agrupamento

Em relação a utilização do algoritmo de agrupamento, como já citado, foi feita uma análise gráfica levando em conta, ‘demanda’ como o atributo principal e desconsiderando ‘dia’, ‘mês’, ‘ano’ e ‘feriado’ da base. Além disso, utilizou-se a base original, ou seja, a que possuía a maior parte dos atributos numéricos.

Outro ponto é que foi feita vários testes com quantidade de clusters variados porém os resultados foram parecidos, e, nos gráficos que serão apresentados, o número de clusters é de 4. Aliado a isso, o objetivo do grupo foi a explicação e a detecção dos atributos que apresentavam alta demanda. A seguir os gráficos encontrados com a demanda:

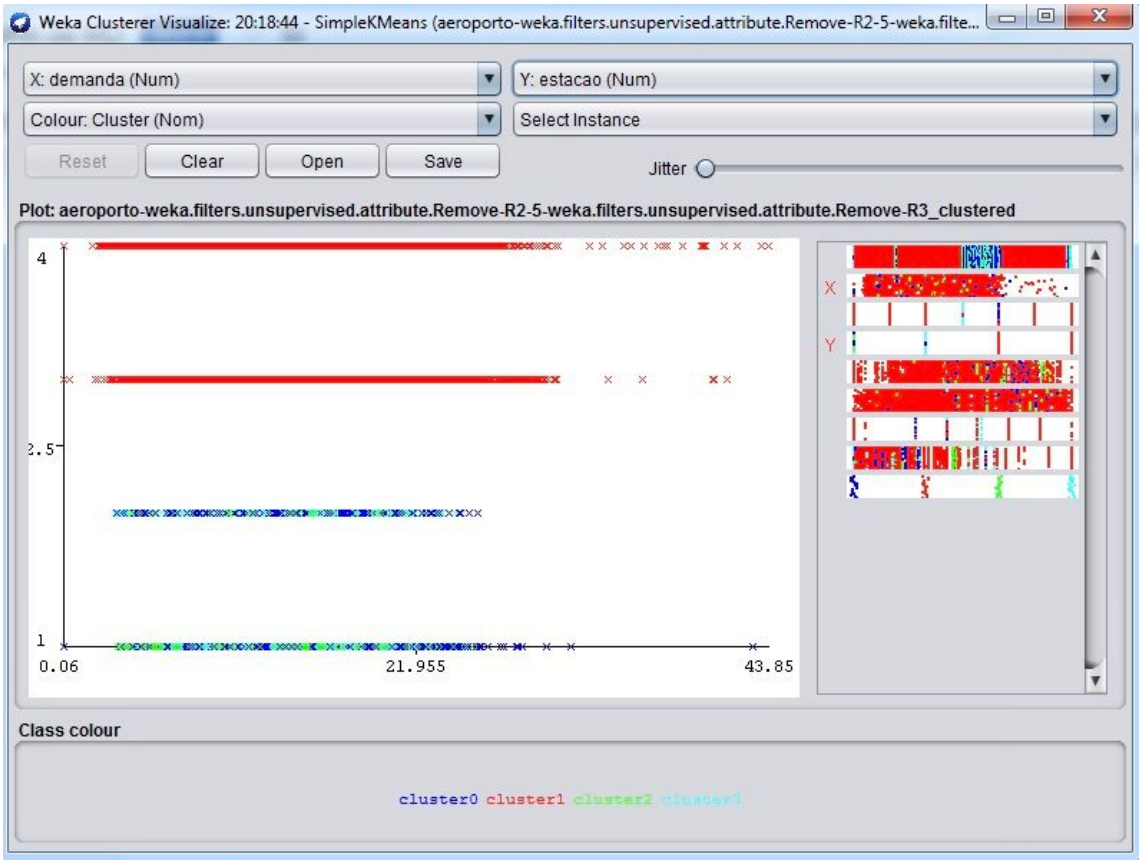
IMAGEM 6 - GRÁFICO DEMANDA X DIA DA SEMANA



FONTE: WEKA

Analisando esse gráfico (DEMANDA X DIA DA SEMANA) percebeu-se que,diferentemente do que imagina-se, no domingo, que é o dia da semana que, normalmente, as pessoas ficam mais em casa por conta da folga do serviço, é um dos dias que a demanda, no caso, apresenta os seus menores índices. Em contrapartida, no sábado, assim como espera-se, os níveis mais altos foram registrados, e, nesse caso, uma explicação plausível seria de que, por conta de se tratar de um dos dias de folga, as pessoas tendem a realizar a maior parte dos serviços domésticos como, por exemplo, lavar as roupas da semana.

IMAGEM 7 - GRÁFICO DEMANDA X ESTAÇÃO DO ANO

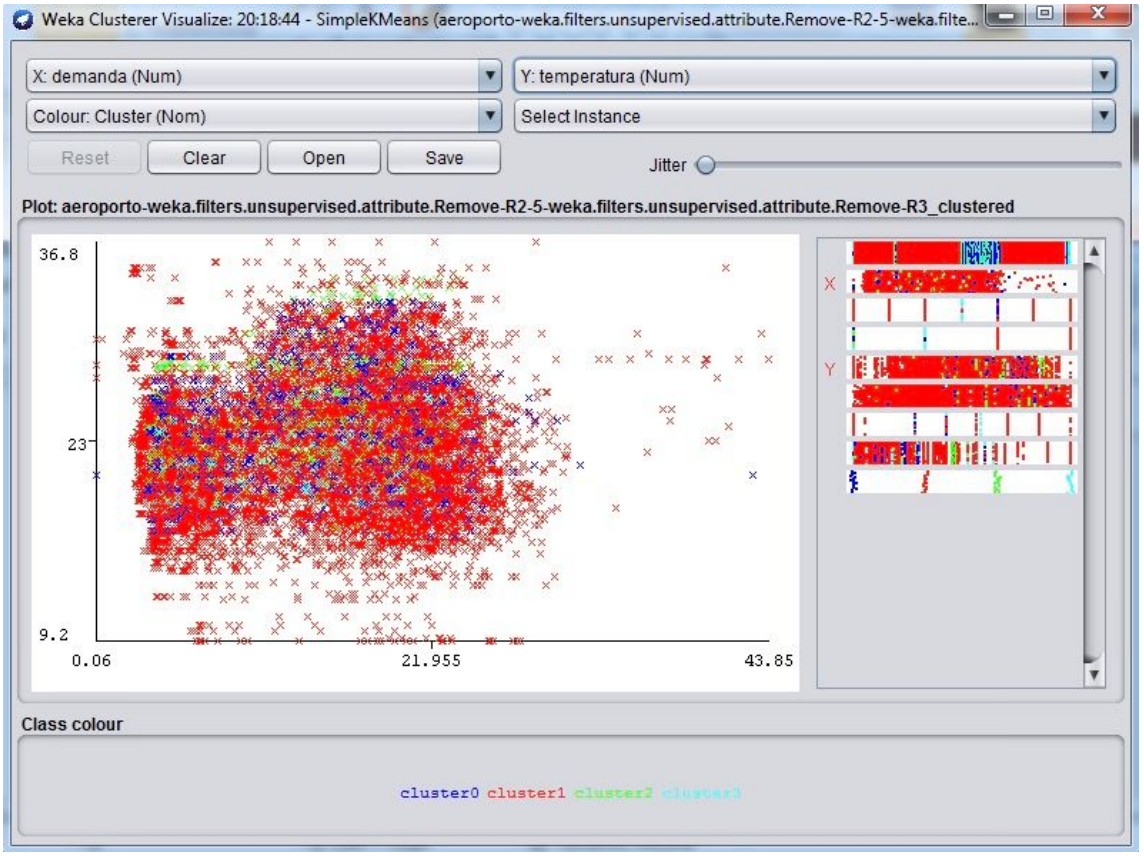


FONTE: WEKA

Percebe-se, que, a maior quantidade de demanda hídrica na região foi no período tanto da primavera quanto do inverno, o que para o grupo foi surpreendente pois esperava-se que o maior consumo de água fosse

registrado no período do verão, visto que, por conta do calor, as pessoas tendem-se mais a tomar banhos refrescantes, por exemplo. Porém, um ponto que pode-se explicar essa dominância da primavera e do inverno é que, por conta do calor, as pessoas tendem a sair mais de suas residências para possíveis passeios, piqueniques e etc.

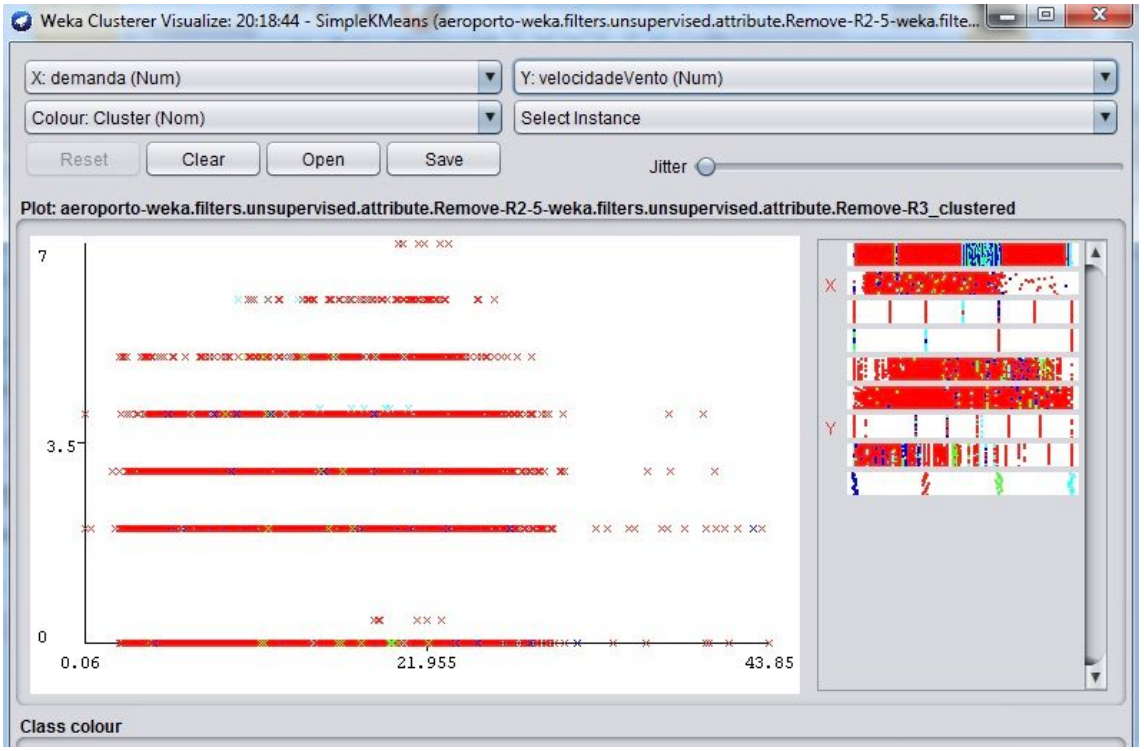
IMAGEM 8 - GRÁFICO DEMANDA X TEMPERATURA



FONTE: WEKA

Assim como esperado, os dias que apresentam uma maior temperatura tendem a ter os maiores índices da demanda de água visto que atividades domésticas tornam-se mais viáveis, entretanto, nota-se que em todas as porções de temperatura há um domínio da demanda baixa/média.

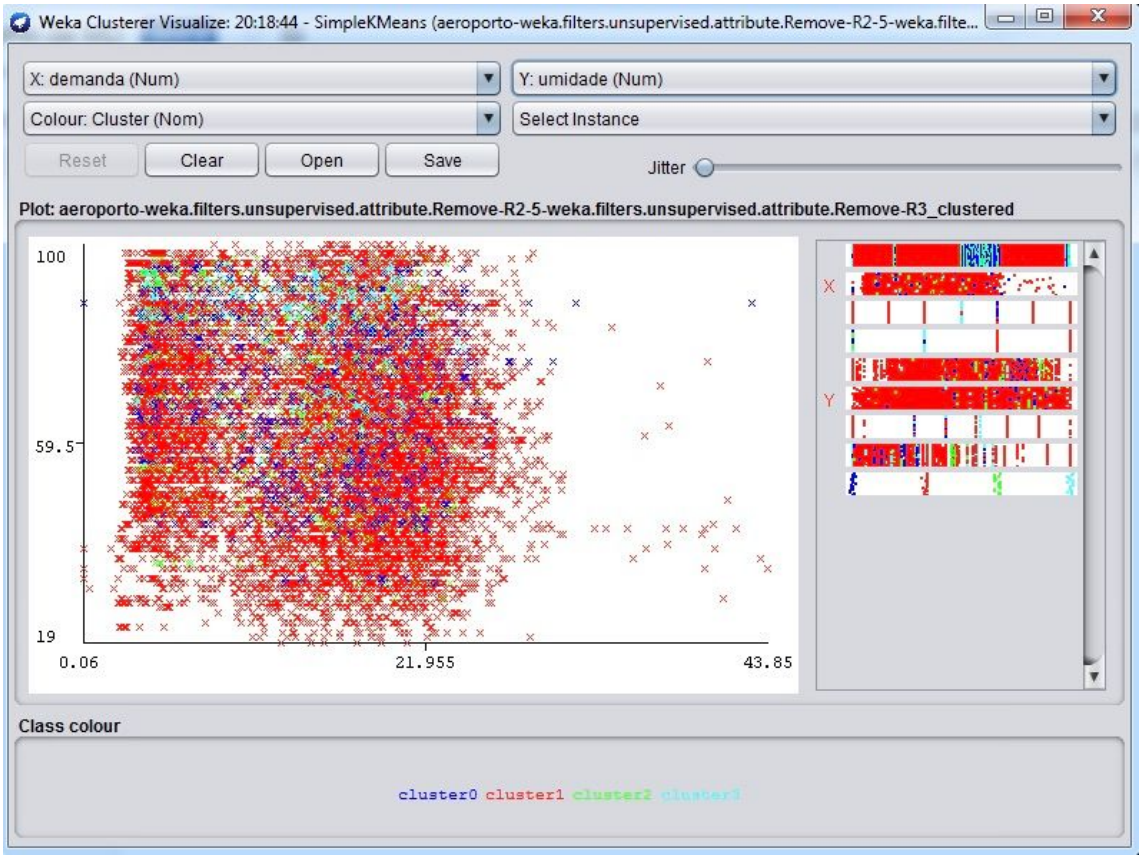
IMAGEM 9 - GRÁFICO DEMANDA X VELOCIDADE DO VENTO



FONTE: WEKA

Os níveis mais altos de demanda, nesse caso, foram registrados nos dias que apresentavam menores velocidade do vento, isso em partes, pode-se explicar pelo fato de que, com uma menor força das incidência dos ventos, tende-se a, assim como na temperatura, realizar uma maior quantidade de trabalhos domésticos, como, por exemplo, nesse caso, limpar o seu carro, ou a fachada da sua casa.

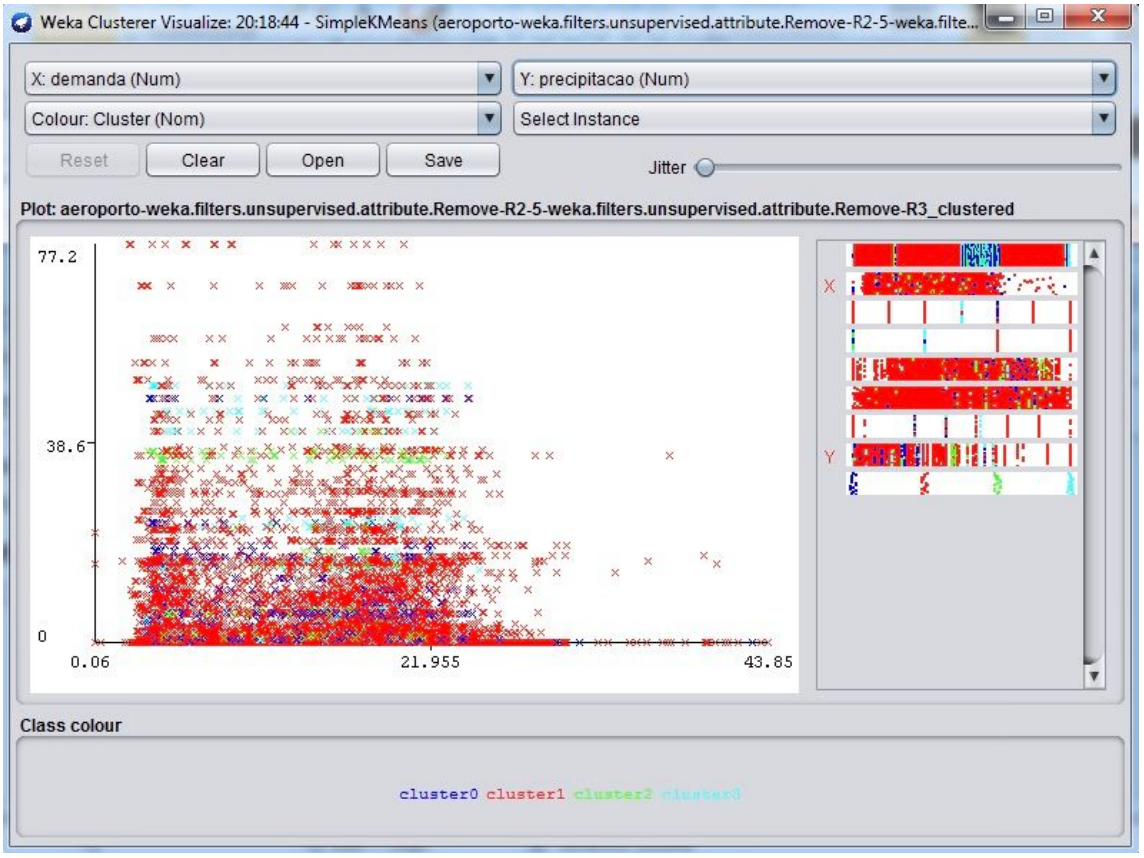
IMAGEM 10 - DEMANDA X UMIDADE



FONTE: WEKA

A concentração dos níveis mais altos da demanda encontram-se em níveis de umidade mais baixos, isso, aliado a explicação da temperatura e da velocidade do vento, demonstram que em dias com menor umidade as pessoas tendem a preferir realizar atividades domésticas, por conta da facilidade em secar os objetos e, não se tem a sensação da alta umidade, que em muitos casos causa mal estar.

IMAGEM 11 - DEMANDA X PRECIPITAÇÃO



FONTE: WEKA

Nesse último caso, percebe-se que, assim como todas as outras medidas climáticas, quando a condição climática é mais propícia para a execução de atividades domésticas, são os pontos em que a demanda atinge seus maiores índices, ou seja, quando não chove, as pessoas preferem realizar suas atividades em relação a dias chuvosos.

Nota-se entretanto, que a ‘demanda’ alta ainda em todos os casos é encontrada em grande escala ainda visto que, no algoritmo de associação, estipulou-se que demandas acima de 21.21 seriam consideradas altas.

7 CONCLUSÃO

A partir dessas informações é possível notar que a demanda encontra-se intrinsecamente ligada nas condições climáticas da região, o que foi demonstrado tanto no algoritmo de associação quanto no de agrupamento. Especificamente, no algoritmo de associação houve uma grande correlação entre precipitação e demanda em relação aos demais atributos que se encontravam também na base.

Outro ponto, é que apesar de ser o algoritmo secundário, o método de agrupamento foi de mais fácil análise e compreensão quando comparado ao de associação, visto que no de agrupamento, diversos gráficos foram apresentados e, os clusters ficaram bem visíveis. E, além disso, possibilitou a comparação da demanda com cada um dos demais atributos.

Um outro ponto, foi que os atributos que se referiam a data, não foram utilizados visto que o encaixe destes é bem complicado na tabela transacional enquanto, na tabela numérica causavam interferência na distância usada para calcular os clusters.

REFERÊNCIAS BIBLIOGRÁFICAS

https://www.aquare.la/o-que-e-data-mining-mineracao-de-dados/
e o pdf da ana estela
https://exame.abril.com.br/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/
https://www.infonova.com.br/artigo/o-que-e-mineracao-de-dados/
https://novaescola.org.br/conteudo/1142/a-agua-e-um-recurso-natural-esgotavel 09/06