Team Nr. 7

Data Science Project

Nr 81478 Name: Joana Godinho
Nr 81811 Name: João Pedro Crespo
Nr 81888 Name: Pedro Caldeira

MEIC-A SAD 2017

INDEX

1 INTRODUCTION

Nesta primeira fase, vão ser analisados resultados de técnicas de aprendizagem não supervisionada, mais especificamente *association rules discovery* e *clustering*, aplicadas a dois *datasets, crabs.csv e noshows.csv*.

O *dataset crabs* é constituído por 200 entradas com 8 atributos. O atributo *sp* representa a espécie do caranguejo (atributo nominal). O atributo *sex* representa o sexo do caranguejo (atributo nominal). Para cada entrada existe um índice (atributo inteiro) que identifica cada caranguejo dentro da sua espécie e sexo. Os outros 5 atributos dizem respeito a características morfológicas do caranguejo (atributos numéricos em mm): O tamanho do lóbulo frontal (FL), largura traseira (RW), tamanho da carapaça (CL), largura da carapaça (CW) e profundidade do corpo (BD).

O dataset noshows é constituído por 110527 entradas com 14 atributos. Cada entrada representa características de um paciente referente a cada consulta marcada. O atributo *PacientID* e *AppointmentID* são atributos numéricos e representam, respetivamente, a identificação do paciente e da consulta. O atributo *Gender* representa o sexo do paciente (atributo nominal). O atributo *ScheduledDay e AppointmentDay* (atributos do tipo *datetime*) representam, respetivamente, a data de marcação de consulta e a data marcada. Age representa a idade do paciente (atributo inteiro). *Neighbourhood*, contem informação sobre a região do paciente (atributo nominal). O atributo *noshow* (atributo nominal) diz se o paciente foi à consulta. *Handcap* exibe o número de deficiências (atributo numérico). Os outros atributos são lógicos e representam se o paciente tem bolsa família (*Scholarship*), hipertensão, diabetes, alcoolismo e se receberam o SMS.

2 PRE PROCESSING

Primeiramente, removeram-se os atributos classe de cada *dataset: species* do *crabs* e *Noshow* no *noshows*.

2.1 Problem 1

Regras de Associação: Inicialmente, para o teste de controlo, simplesmente arredondaram-se os valores dos atributos FL, CL, CW, RW e BD de forma a obter valores discretos, possibilitando assim, a recolha de regras de associação com o método ***Apriori***. Posteriormente, para os testes comparativos, utilizou-se discretização de valores por intervalo ou por frequência.
Clustering: Substituiu-se o atributo nominal *gender* pelo atributo binário *isMale,* dado que a variável nominal pode apenas tomar dois valores (M ou F).

2.2 PROBLEM 2

Regras de Associação: Para os atributos *ScheduledDay* e *AppointmentDay* separaram-se as horas das datas. Obtiveram-se assim dois novos atributos, *ScheduledHour e AppointmentHour,* no entanto *o AppointmentHour* foi removido, pois tomava um único valor. Consequentemente, os atributos *ScheduledDay e AppointmentDay* passaram a ter apenas informação referente ao dia, informação esta que foi traduzida em valores de 1 a 366, consoante o dia do ano. Adoptaram-se ainda outras **taxonomias**, nomeadamente, a semana do ano (1-52), o mês (1-12) e o dia da semana (Seg-Dom). Para além destas, o atributo idade foi separado em seis intervalos, referentes às fases de vida de uma pessoa (infância, adolescência, jovem adulto, adulto, idoso).

Clustering: Primeiro removeram-se os ***outliers***, nomeadamente, uma entrada que tinha idade negativa e outras 63 entradas cujo *ScheduledDay* referia-se a datas do ano 2015. Assim como no Problema 1, o atributo *gender* foi substituido pelo atributo binário *isMale*. Para além disso foram ignorados os atributos *AppointmentID, que se* tratava de uma *primary key*, e o *PatientID*, uma vez que era quase diferente para cada entrada. Foi ainda feita normalização e PCA para capturar melhor a variância dos dados, de forma a verificar se os resultados seriam melhores.

3 EXPLORATION

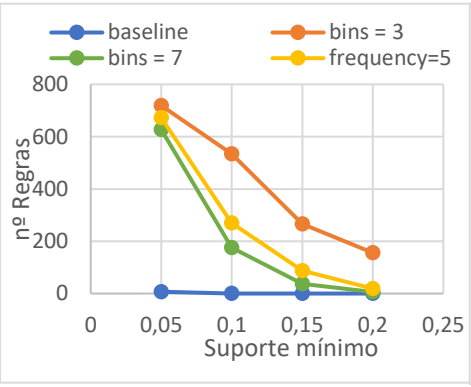
3.1 Problem 1

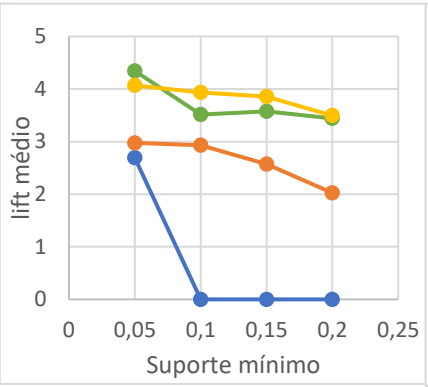
3.1.1 Methods and Parametrization

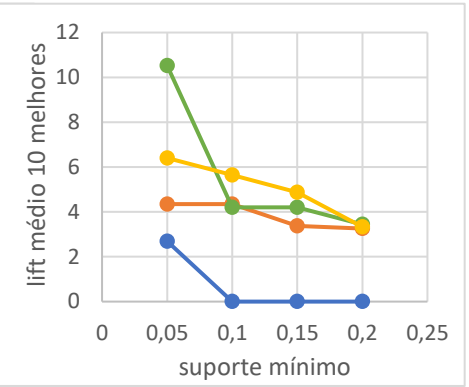
Regras de Associação: Método ***Apriori*** em que se manteve a confiança a 75% e variou-se o suporte mínimo entre 5% e 20%. Relativamente às discretizações mencionadas no pré-processamento, agruparam-se os valores das variáveis em 3 e 7 intervalos (discretização por intervalos) e por 5 grupos de caranguejos com igual frequência.

Clustering: Método ***k-means*** com número de *clusters* a variar de 2 a 9. Para a avaliação dos *clusters* foram utilizados os índices de ***Dunn*** e ***Davies***, tendo em conta a combinação ***centroids-centroid*** por ser menos sensível a extremos. Testou-se ainda o método ***Expectation-Maximization***.

3.1.2 Results





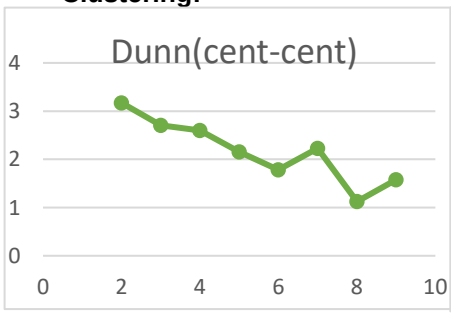


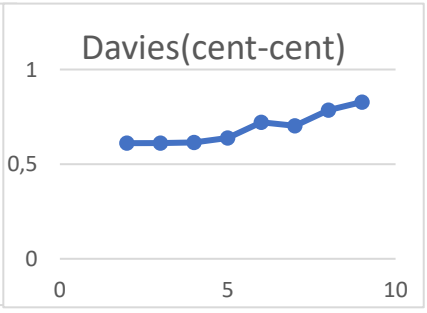
Como se verifica através da análise dos gráficos, à medida que **aumentamos** o suporte mínimo, o número de regras e o *lift* médio das mesmas fica mais **reduzido**. Podemos então concluir que com uma discretização com 7 intervalos e uma discretização com frequência de 5 são os casos em que se obtém *lift* médio melhor. No entanto, a partir do nível de suporte mínimo de 10%, a discretização por frequência de 5 apresentou regras com maior qualidade.

Regras obtidas:

Regra	*Suporte*	*Confiança*	*Lift*
CL = [15,26) => CW = [17.0,29.7)	0.190	1.000	5.000
FL = [7.0,12.3)} => {CL = [15,26)	0.170	0.919	4.834
FL = [7.0,12.3) => CW =[17.0,29.7)	0.170	0.919	4.595
CL = [15,26)} => {BD = [6.0,11.3)	0.185	0.977	3.819
FL=[20,23]} => {BD=[18,22]	0.145	1.000	5.405

Clustering:





Com base nos gráficos podemos concluir que o melhor *clustering* é com 2 *clusters*, dado que para k = 2 o índice de *Dunn* é máximo e para o índice de *Davies* é mínimo.
Pode-se ainda concluir que o índice de *Dunn* tende a diminuir com o número de clusters e o índice de *Davies* tende a aumentar.

O algoritmo *Expectation-Maximization* confirma o resultado do *k-means*: obtiveram-se 2 *clusters* com **igual volume, forma e orientação**.

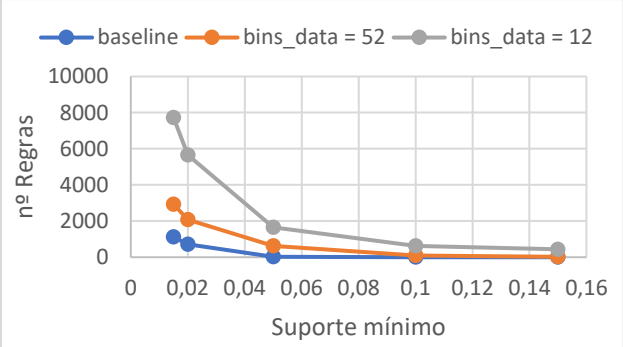
3.2 Problem 2

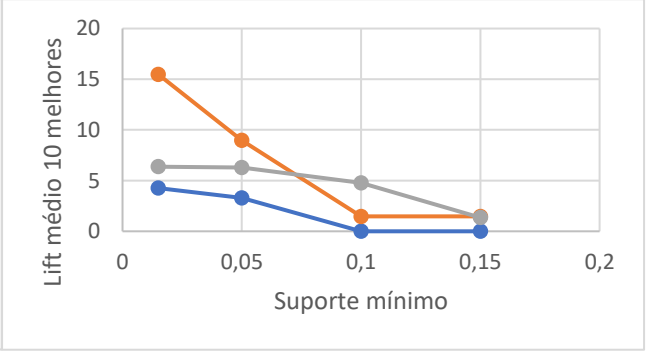
3.2.1 Methods and Parametrization

Regras de Associação: Método ***Apriori***, tal como no Problema 1, a confiança mínima escolhida para as regras foi de 75%. Contudo, variou-se o suporte mínimo entre 0,015 e 0,15 devido ao **maior número de instâncias**. Todas as regras obtidas pelo algoritmo foram filtradas de forma a que as que possuíssem como cabeça a ausência de um sintoma (Hipertensão, Diabetes, Handcap e Alcoolismo) ou a ausência de Bolsa Escolar fossem omitidas, dado que a probabilidade destes acontecimentos é demasiado elevada (acima de 80%).

Clustering: Método ***k-means***, tal como no problema 1, com o número de *clusters* a variar entre 2 e 30. A avaliação dos *clusters* foi também feita através dos índices de *Dunn* e de *Davies* através da combinação de medida ***intercluster*** – medida ***intracluster: centroids*** *(distância entre dois centroides)* – ***centroid*** *(dobro da distância média ao cluster)*, devido à sua robustez. Testou-se ainda o método *Expectation-Maximization*, no entanto, **não foi possível terminar** devido à grande quantidade de instâncias.

3.2.2 Results

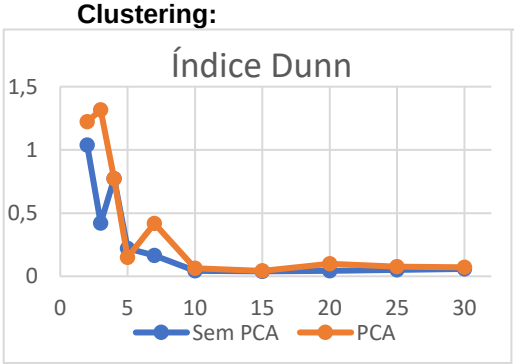


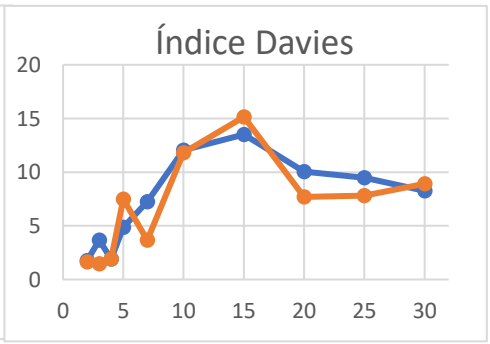


A **taxonomia** dos dias da semana não apresentou diferenças nos resultados relevantes, daí só serem exibidos os resultados das restantes taxonomias.
Tal como no *dataset* do problema 1, à medida que se **aumenta** o suporte mínimo, o número de regras vai **diminuindo**. Para a taxonomia das semanas do ano, obtiveram-se regras com *lift* superior, até um suporte mínimo de 0.05. A partir daí a taxonomia dos meses do ano foi a que apresentou melhores resultados.

Regra	*Suporte*	*Confiança*	*Lift*
ScheduledDay=23 week => AppointmentDay=23 week	0.0518	1.00	8.135
Diabetes=1 => Hipertension=1	0.0586	0.816	4.146

Clustering:





Tal como se pode ver nos gráficos, o melhor *clustering* é feito com 2 e 3 clusters quando não se efectua PCA e quando se efectua, respectivamente.

Pode ainda aferir-se que a aplicação do PCA levou a um melhor resultado, pois apresentou os melhores valores dos índices de *Dunn e Davies* (*Dunn* maior e *Davies* menor).

4 CRITICAL ANALYSIS

No *dataset* dos *crabs*, com base nos resultados obtidos, para a discretização por 7 intervalos e um suporte mínimo de 5%, o *lift* médio das dez melhores regras é quase o triplo do *lift* médio de todas as regras obtidas, o que significa que com estes parâmetros aparecem regras que se destacam mais. No entanto, um suporte de 5% (10 instâncias) torna-se pequeno para um *dataset* da dimensão do *crabs*. A partir de um suporte mínimo de 10%, para as discretizações por 7 intervalos e por frequência 5, as regras obtidas tendem a convergir para um dado *lift*. Esta convergência é corroborada pelo facto do *lift* das dez melhores regras se aproximar do *lift* médio, o que reflecte a diminuição da variância do mesmo. Assim, conclui-se que as regras descobertas não são muito relevantes para o **domínio do problema** e caso existam será para um suporte muito **pequeno**.

Pelas regras obtidas podemos ver que determinadas características morfológicas dos caranguejos estão correlacionadas. Tal era expectável pois é esperado que um caranguejo com FL alto tenha um CW e um CL altos também.

No *clustering*, os resultados para dois *clusters* vão de encontro com as regras obtidas pois num dos *clusters* encontravam-se caranguejos com CL,CW e FL maiores e no outro caranguejos com estas medidas menores, o que corrobora as **correlações encontradas** nas regras de associação.

No *dataset* do *noshows*, com base nos gráficos, podemos concluir que a discretização por meses do ano não leva a regras relevantes, uma vez que o *lift* das dez melhores regras é praticamente constante, à medida que se aumenta o suporte mínimo. Ao invés, a taxonomia por semanas do ano apresenta uma maior variação do *lift*. Para além disso, esta apresenta um *lift* médio elevado para um suporte adequado ao *dataset*, 5% (~5500 instâncias). Logo, é a partir desta discretização que é **mais provável obter regras possivelmente relevantes**.
Contudo, apenas se encontrou uma regra com bons resultados e relevante para o domínio do problema: Diabetes=1 => Hipertension=1, pois revela que pacientes que tenham diabetes estão mais predispostos a ter hipertensão.

O facto das regras que concluíam a ausência de um sintoma serem descartadas, **não apresenta grandes implicações nos resultados**, pois a informação que estas regras trariam seria pouco pertinente para o domínio do problema, ao contrário daquelas que concluíam a presença de um sintoma.

No *clustering* do *noshows*, apesar da utilização do PCA no pré-processamento, o resultado obtido (3 clusters), não permite extrair ilações sobre o *dataset*. Tal pode dever-se ao facto de se ter removido o atributo *Neighbourhood* para se utilizar o *k-means*, dado que este apenas pode ser utilizado com atributos quantitativos e binários. No entanto, a remoção pareceu ser a melhor decisão pois a atribuição de valores numéricos a cada valor do atributo *Neighbourhood*, implicaria que havia uma relação de ordem entre estes, o que não traduziria a realidade.

5 CONCLUSIONS

Numa perspectiva geral, podemos concluir que os resultados para o *noshows* tanto para as regras de associação como para o *clustering* **não foram tão favoráveis** como no *crabs* devido à **heterogeneidade**, **dimensão** e, consequente **dificuldade** do *dataset*. Ademais, em nenhum dos *datasets*, utilizando quer as regras de associação quer o *clustering*, se obteve resultados muito interessantes para o domínio do respetivo problema. Contudo, os resultados das regras de associação demonstraram-se mais úteis para extrair alguma informação sobre os *datasets*.