Instituto Superior de Engenharia de Lisboa

Mineração de Dados em Larga Escala

Mestrado Engenharia Informática e Multimédia &
Mestrado Engenharia Informática e Computadores

Laboratório 2 - Data Representation and Dimensionality Reduction

Semestre de Verão, 2022/2023
19 de abril de 2023

Grupo 8:
Nome: Gonçalo Fonseca | Número: A50185
Nome: Pedro Diogo | Número: A47573

Índice

1 Laboratory Class Setup/Preparation - Data Inspection and Datasets Available on the Web

1.1 i)

No dataset "Diabetes", existem 6 features com valores inteiros (preg, plas, pres, skin, insu e age), e existem 2 features com valores reais (mass e pedi).
Por outro lado, no dataset "Influenza Outbreak", todas as 545features apenas contém valores inteiros de 0 e 1, pois este dataset verifica se cada tweet feito contém ou não um certo termo, dando o resultado de 1 ou de 0.

2 Data representation formats and data inspection and visualization tools

2.1 a)

Os ficheiros suportados pelo WEKA são: arff, arff.gz, bsi, csv, dat, data, json, json.gz, libsvm, m, names, xrff e xrff.gz. [1]

2.2 b)

2.2.1 i)

- **CSV**: a primeira linha é referente aos atributos, e o resto às várias instâncias, separado por virgulas;
- **JSON**: está em formato JSON. Começa com um array de atributos, e de seguida existem vários arrays de dados (de cada linha) com as várias instâncias. Ainda, inclui vários metadados em cada instância;
- **XRFF**: encontra-se em formato xml, onde este se divide em 1 bloco de atributos com a tag **attribute**, e vários blocos, em que cada apresenta as várias instâncias de cada linha, com a tag **value**;
- **XRFF.gz**: Este é igual ao anterior, mas está comprimido.

2.2.2 ii)

Abaixo, encontra-se a tabela com os resultados pedidos para cada dataset, sendo que o dataset "Influenza"foi usado o dataset de treino.

Tabela 1: Tabela com os resultados

	Diabetes	**Ionosphere**	**Influenza**
Nº Padrões	768	251	1095
Nº Features	8	34	2
Nº Classes	2	2	2
Nº Padrões por Classe	500 e 268	225 e 126	
Feature com maior nº valores distintos	pedi	a28	soon
Feature com maior variância	insu	a15	soon

No entanto, através do MEKA, não foi possível obter o número de padrões por classe pois não existem as labels nos ficheiros fornecidos. No entanto, usando os ficheiros do projeto que contêm as labels, e usando Python, descobriu-se que existem 1046 valores negativos (a 0) e 44 positivos (a 1).

2.3 c)

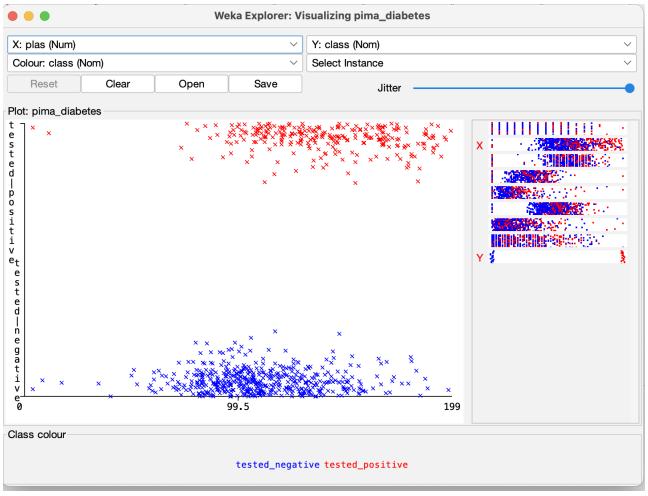

Figura 1: Feature 2 vs Class Label

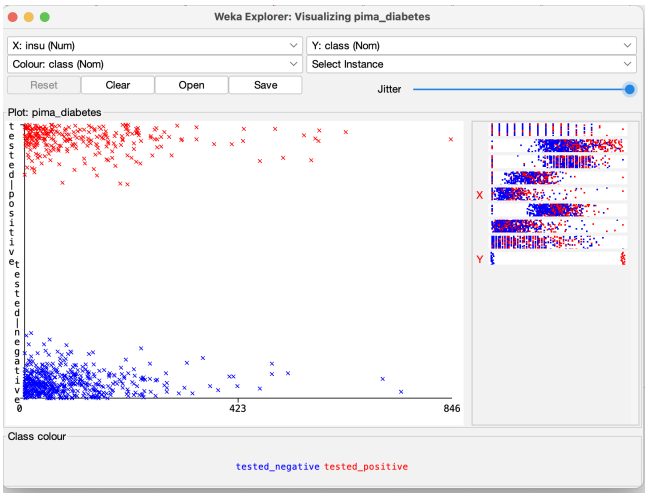

Figura 2: Feature 5 vs Class Label

Através dos *screenshots* acima, pode-se verificar que o melhor atributo é o atributo 5 (insu), devido a ser aquele que tem mais baixa variância.

3 Dimensionality reduction and evaluation

3.1 a)

Na figura abaixo, são apresentadas as técnicas de "Attribute Evaluator"existentes.

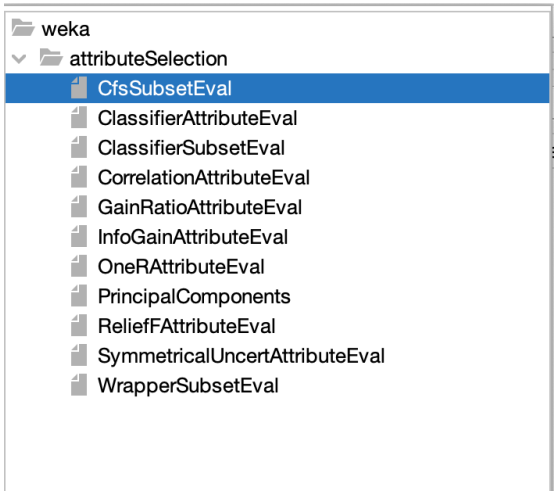

Figura 3: Técnicas de Attribute Evaluator

De seguida, são apresentadas as técnicas de "Search Method"existentes.

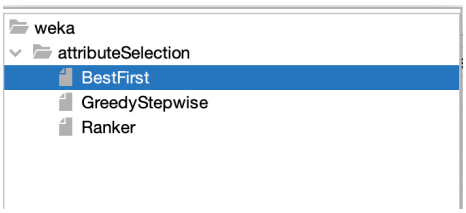

Figura 4: Técnicas de Search Method

Por fim, são apresentadas as técnicas de "Attribute Selection Mode"existentes.

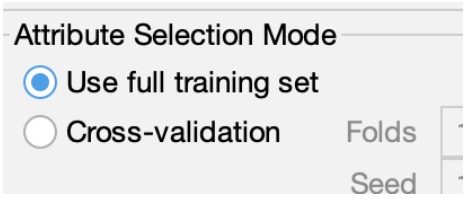

Figura 5: Técnicas de Attribute Selection Mode

3.2 b)

Usando o *dataset* ***Diabetes***, obteve-se o seguinte output:

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

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 37
        Merit of best subset found:    0.164

Attribute Subset Evaluator (supervised, Class (nominal): 9 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,6,7,8 : 4
                     plas
                     mass
                     pedi
                     age
```

Figura 6: Output usando o método CfsSubsetEval com a procura Best Search usando a opção Full training set

Os atributos selecionados foram quatro: plas, mass, pedi e age. Visto que se usou o training set por completo (opção "Use full training set"na aba "Attribute Selecion Mode"), o subset escolhido será sempre o mesmo.

3.3 c)

Para além das várias opções usadas anteriores, aplicou-se a opção de Cross-Validation com 10-folds. Com isto, obteve-se o seguinte output:

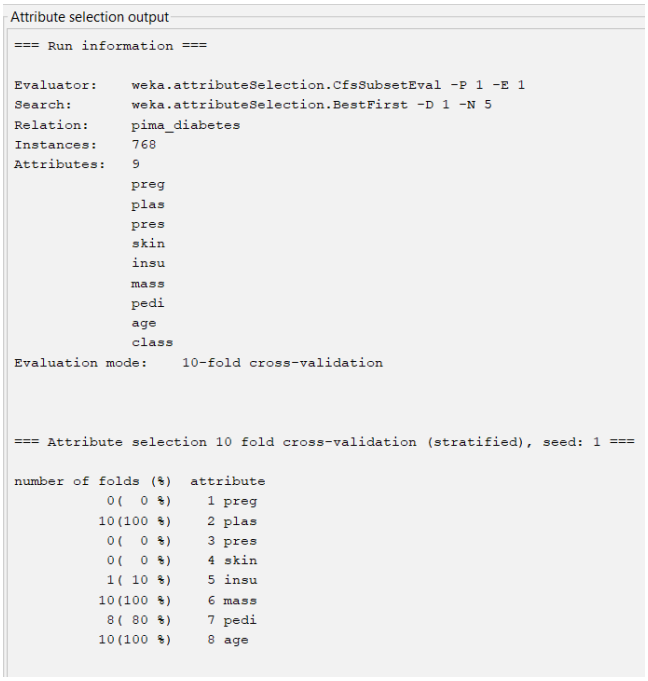

Figura 7: Output usando o método CfsSubsetEval com a procura Best Search usando a opção Full training set e Cross-Validation com 10-folds

Foram selecionadas 5 *features*, mais especificamente as *features* 2, 5, 6, 7 e 8. O resultado será sempre o mesmo, caso se mantenha o mesmo valor na "seed".

3.4 d)

O resultado do classificador j48 usando Cross-Validation com 10 folds, obteve-se uma acurácia de 73.82%. Olhando para a árvore construida, o atributo que se encontra na raiz da mesma é o **plas**. Logo, é o atributo mais relevante.
O output com a árvore e com o código Java está anexado a este relatório.

4 Orange environment and analysis of existing examples

4.1 b)

Na figura abaixo são apresentados os vários tipos de ficheiros possíveis a serem usados no Orange.

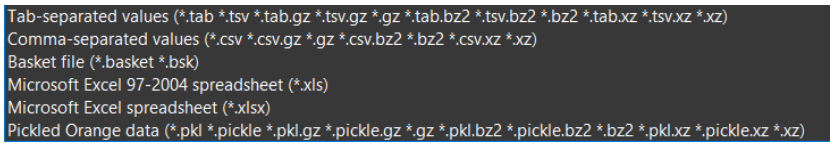

Figura 8: Tipos de ficheios usandos no Orange

De seguida, são apresentadas as análises estatísticas do dataset da Íris.

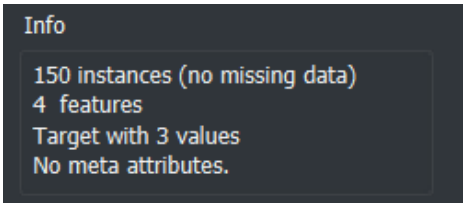

Figura 9: Análises Estatisticas - Dataset Íris

4.2 c)

O exemplo "Interactive Visualizations"existente no Orange serve para os utilizadores poderem obter uma melhor compreensão das relações entre as diferentes características das flores da íris e como podem ser utilizadas para classificar as diferentes espécies de flores da íris.
Em relação ao widget Data Info, este mostra várias informações do dataset como as apresentadas a seguir.

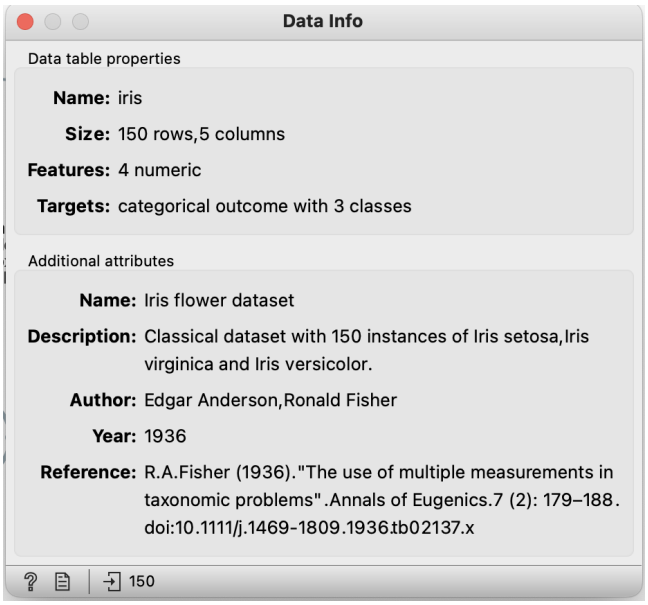

Figura 10: Informação sobre o dataset Íris

As features na figura seguinte são as features mais relevantes, segundo o critério IG e FCBF.

	#	Ga...io	FCBF
1	N petal length	0.544	1.542
2	N petal width	0.532	1.451
3	N sepal length	0.313	0.000
4	N sepal width	0.183	0.255

Figura 11: Features

5 Feature ranking and selection

5.1 a)

A feature mais relevante é a "petal length", tal como se verifica abaixo.

	#	Inf...ain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
1	N petal length	1.086	0.544	0.423	1179.034	98.946	0.367	1.542
2	N petal width	1.059	0.532	0.407	959.324	94.162	0.374	1.451
3	N sepal length	0.624	0.313	0.247	119.265	79.243	0.153	0.000
4	N sepal width	0.361	0.183	0.154	47.364	50.082	0.129	0.255

Figura 12: Feature 5 vs Class Label

5.2 b)

Tal como observador anteriormente, a feature mais relevante é o petal length que, analisando os Scatter Plots nas figuras abaixo realizados a cada uma das features, conclui-se que a "*petal length*"é a mais relevante por estar mais concentrado.

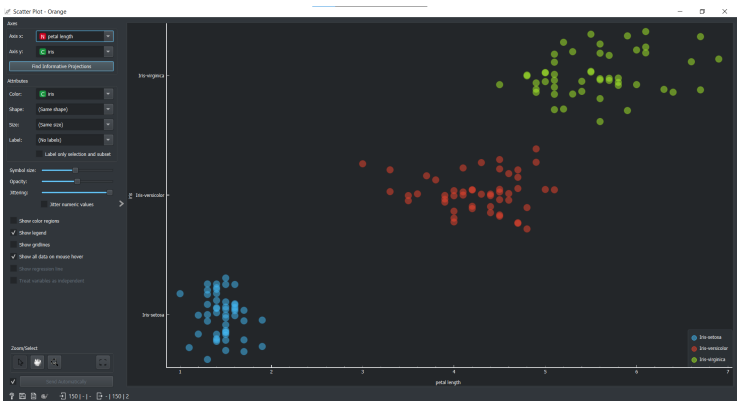

Figura 13: Scatter Plot - Petal Length

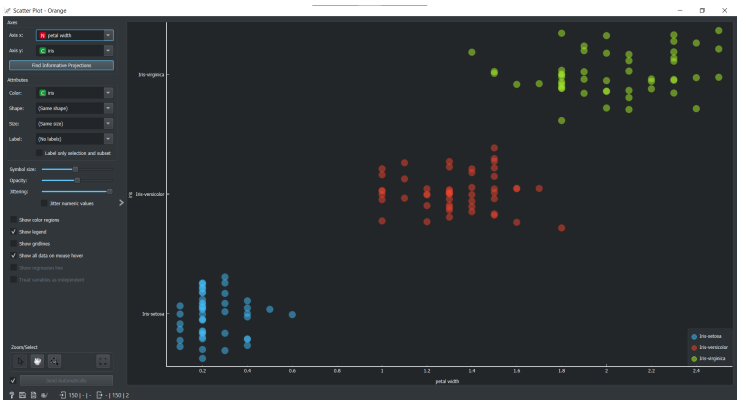

Figura 14: Scatter Plot - Petal Width

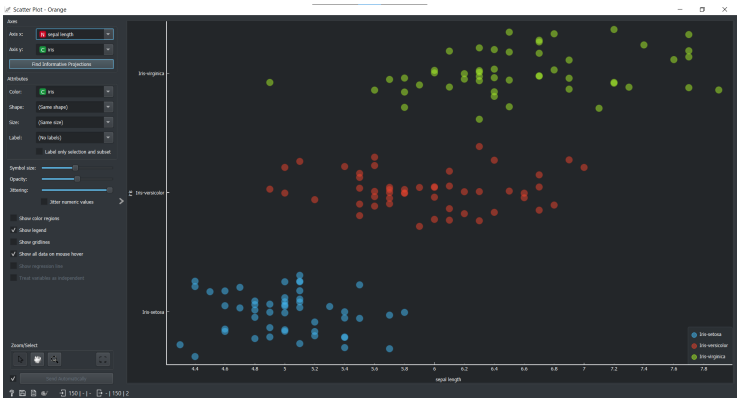

Figura 15: Feature 5 vs Class Label - Sepal Length

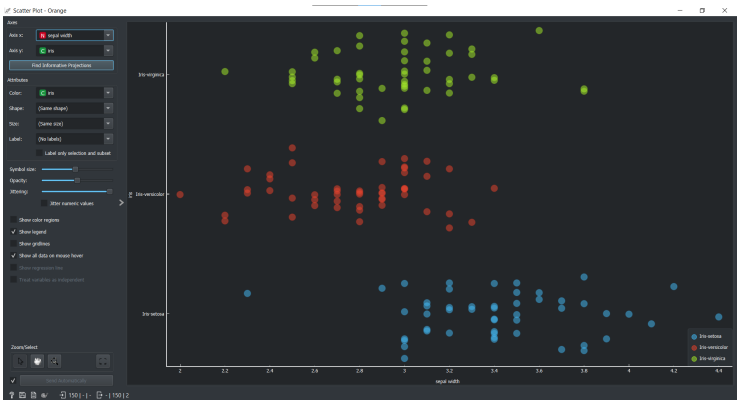

Figura 16: Feature 5 vs Class Label - Sepal Width

6 Feature reduction with principal component analysis and discretization

6.1 a)

As principais ações da demo são: configuração do widget de PCA, usar um plot (neste caso um Scatter Plot) para apresentar os dados, e posteriormente analisar os mesmos.
Para obter um valor adequado de dimensões reduzidas, colocou-se a variância esperada a 90%, em que se obteve 25 componentes.

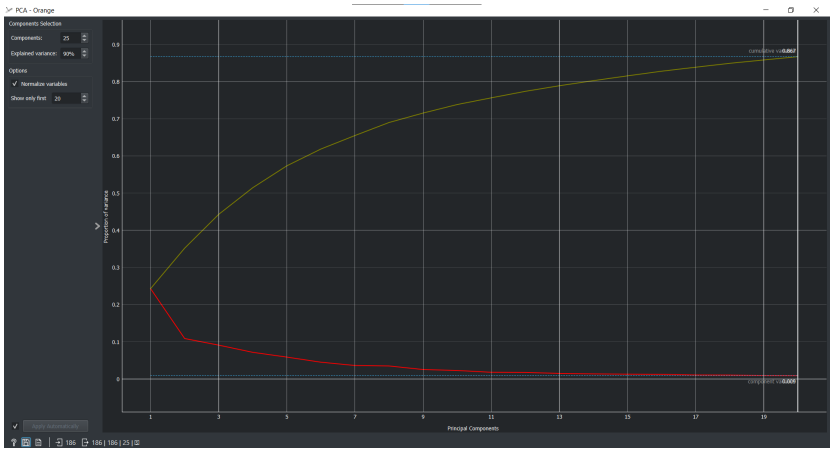

Figura 17: Feature 5 vs Class Label

6.2 b)

Para fazer a ação pretendida, usou-se o widget Discritize, que é apresentado na figura abaixo.

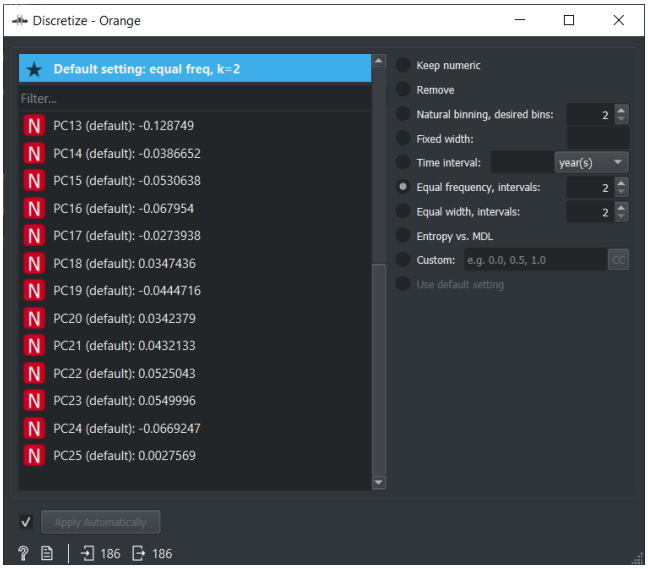

Figura 18: Widget Discritize

O resultado encontra-se no ficheiro guardado com o nome brown-selected.tab, que se encontra junto a este documento, e uma parte apresentado na figura abaixo.

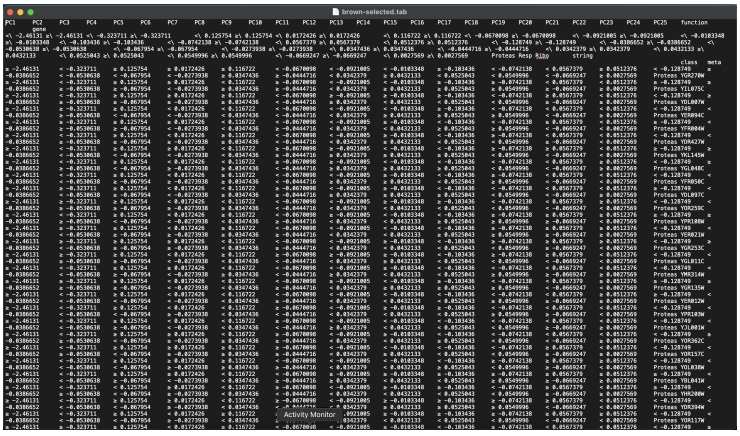

Figura 19: Ficheiro brown-selected.tab

7 R Studio - Feature Selection

Neste capítulo entraremos em detalhe de como podemos reduzir a dimensão de datasets através de *Feature Reduction*, *Feature Selection* e *Feature Discretization*.

7.1 Pré-processamento

Para carregar os datasets fornecidos (**diabetes.csv** e **train_data_1.csv**), utilizou-se a função read.csv do R com o parâmetro head = FALSE e skip = 1. Isto para evitar ler a primeira linha dos conjuntos de dados que por sua vez contêm o nome dos atributos.
Para além disso, no ficheiro diabetes.csv, foi necessário também remover a última coluna que contém as *labels* das várias instâncias.

7.2 a)

No dataset dos diabetes, podemos ver que existe um atributo que possui a maioria relevância, que é o "insu" quando usamos a variância e a diferença da média das medianas. As Figuras 20 e 21 mostram gráficos das variâncias ordenadas de forma decrescente para ambas as técnicas.

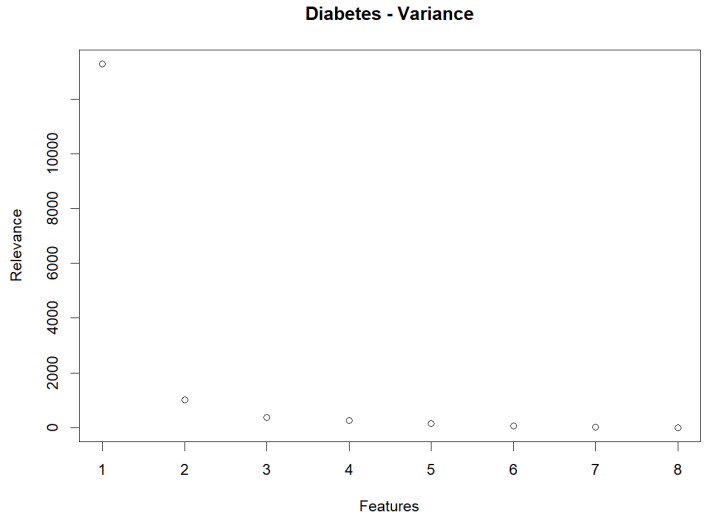

Figura 20: Diabetes - Variance

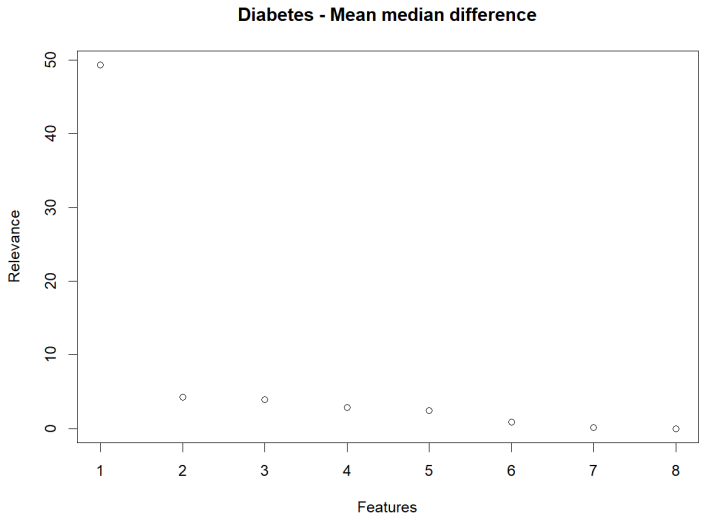

Figura 21: Diabetes - Mean median difference

Já no dataset influenza.csv podemos ver que a partir dos atributos começamos a estabilizar o valor da relevância, como se pode observar nas Figuras 22 e 23.

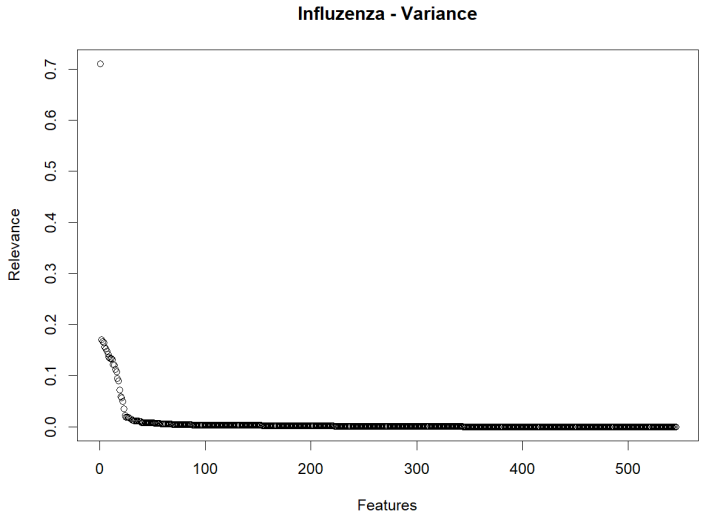

Figura 22: Influzenza - Variance

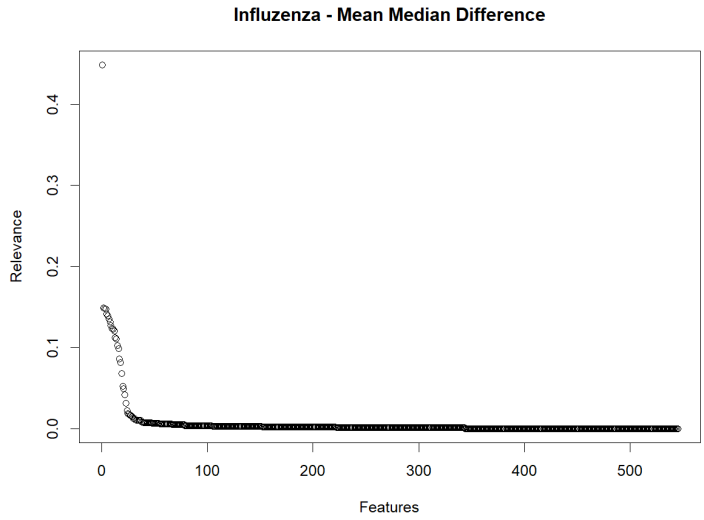

Figura 23: Influzenza - Mean Median Difference

7.3 b)

Para se calcular o número adequado de atributos, foram considerados como valores de *threshold*, 90%, 95% e 99% visto que queremos manter maior parte da relevância no nossos dados.
A Tabela 2 apresenta os resultados obtidos nas várias iterações.

Tabela 2 – Tabela com o número de features por threshold

	90%	**95%**	**99%**
Variance - Diabetes	1	2	4
MM_Diff - Diabetes	1	2	4
Variance - Influenza	91	167	297
MM_Diff - Influenza	104	181	302

8 Feature Reduction

8.1 a)

Para responder à questão, foram feitos 2 gráficos após a decomposição do PCA, sendo que foi feito 1 gráfico para cada dataset, apresentados abaixo.

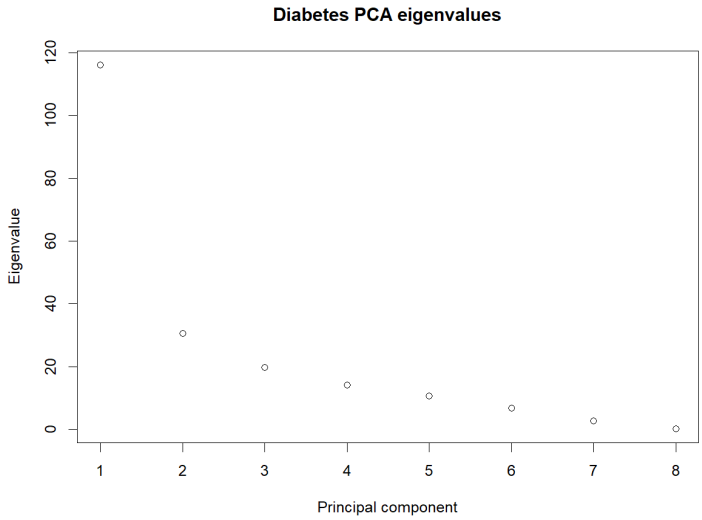

Figura 24: Diabetes PCA eigenvalues

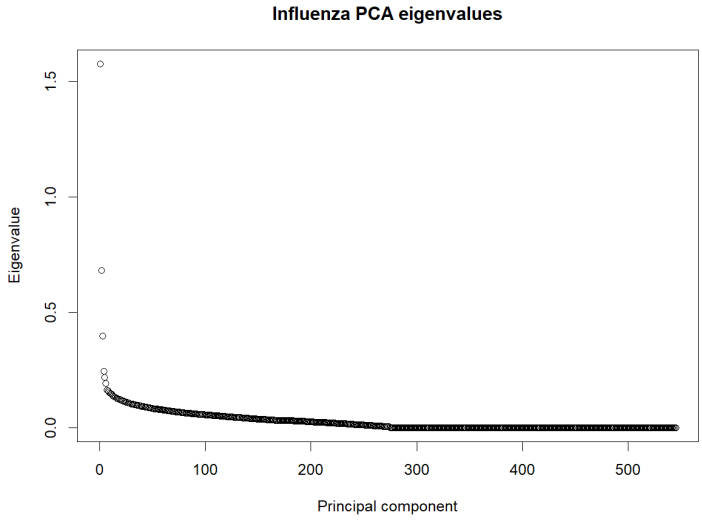

Figura 25: Influzenza PCA eigenvalues

Olhando para os gráficos gerados após ter sido feita a decomposição PCA, é possível observar que para o dataset de diabetes vemos que a partir de 4 ou 5 componentes principais, a reta começa a estabilizar. No caso do dataset influenza, vemos que nos 270 já temos uma reta que começa a ficar uniforme.

Logo m pode ser igual a 270.

8.2 b)

Tal como na alínea anterior, para responder à questão, foram feitos dois gráficos após a decomposição do SVD, sendo que foi feito um gráfico para cada dataset, apresentados abaixo.

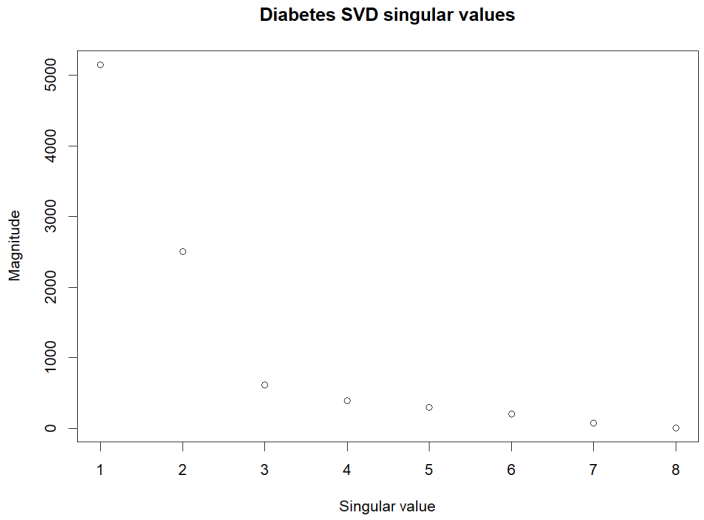

Figura 26: Diabetes SVD singular values

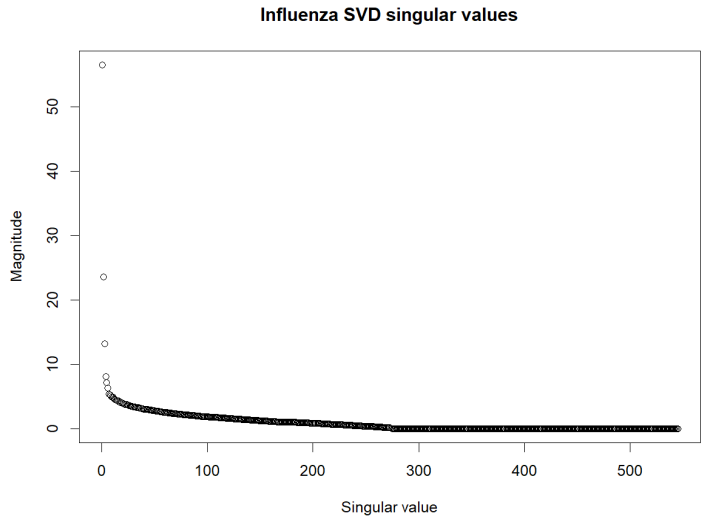

Figura 27: Influenza SVD singular values

Utilizando a decomposição SVD obtemos resultados muito semelhantes à alínea anterior. Podemos admitir então a mesma dimensão obtida anteriormente.

8.3 c)

Para efetuar a redução do dataset com base no PCA, temos de multiplicar a matriz de dados pela matriz que possui o valor das componentes principais obtidas anteriormente. Com base na alínea anterior definimos que para o dataset reduzido de diabetes o número de features é 5 e para o de influenza é 270.
O resultado obtido do SVD (matrizes d, u e v) representam a matriz original do dataset decomposta. Os valores diagonais presentes na matriz d são os singular values, que por sua vez representam a importância dos vários atributos. Com base nisso, fazendo a multiplicação de matrizes entre u d e a transposta de v obtemos um dataset reduzido. As dimensões selecionadas foram as mesmas do PCA.

9 Feature Discretization

O dataset escolhido para fazer *feature discretizaion*, foi o "diabetes". Recorrendo a duas billbiotecas (arules e arulesCBA) obtiveram-se os seguintes resultados:

Tabela 3 – Equal frequency binning

Atributo	preg	plas	pres	skin	insu	mass	pedi	age
# intervalos	3	3	3	3	2	3	3	3

Tabela 4 – MDLP

Atributo	preg	plas	pres	skin	insu	mass	pedi	age
# intervalos	2	4	1	1	3	2	2	2

Em suma, no atrvés do método MDLP houveram duas *features* (pres e skin) cujo intervalo ficou de menos infinito até mais infinito. Isto indica que o algoritmo não foi capaz de formular intervalos para esses atributos. Já o EFB deu o mesmo número de intervalos para todas os atributos exceto o "insu".

Referências

[1] Weka - File Formats - [Em linha] [Consult. 29 mar. 2023]. Disponível em https://www.tutorialspoint.com/weka/weka_file_formats.htm.