Universidade do Minho
Escola de Engenharia

Perfil de Machine Learning: Fundamentos e Aplicações

Sistemas Baseados em Semelhança

1º/2º Ano, 1º Semestre

Ano letivo 2022/2023

Enunciado Prático nº 4

Tema: *Clustering*

Joana Mota PG45528

Tarefas:

T1. O *dataset* descarregado corresponde a dados de treino sobre características de vinhos, nomeadamente, a acidez, o açúcar, a densidade, o ph, a percentagem de álcool, a qualidade, entre outros.

T2. Nesta segunda tarefa realizou-se o tratamento de dados, começando por passar a *feature quality* para número, no entanto, antes disso, foi necessário utilizar o nodo *String Replacer*, pois cada registo continha "=5" e, assim, através deste nodo foi possível retirar o "=". Depois foi só usar o nodo *string to number*. De seguida, normalizou-se todos os atributos numéricos utilizando a transformação linear Min-max de forma a produzir um input normalizado entre 0 e 1, através do nodo *Normalizer*. Por fim, recorrendo ao nodo *Auto Binner*, criou-se 4 bins com a mesma frequência para a *feature citric acid*. E ainda, renomeou-se cada bin para Low, Medium, High e Very High.

T3. De forma a projetar os dados em apenas 2 dimensões utilizou-se o nodo PCA que permite definir o número de dimensões e utilizou-se um *color manager* e um *scatter plot* para visualizar as dimensões por cores obtidas pelo PCA. Também se utilizou um *3D scatter plot* para visualizar de forma tridimensional os dados.

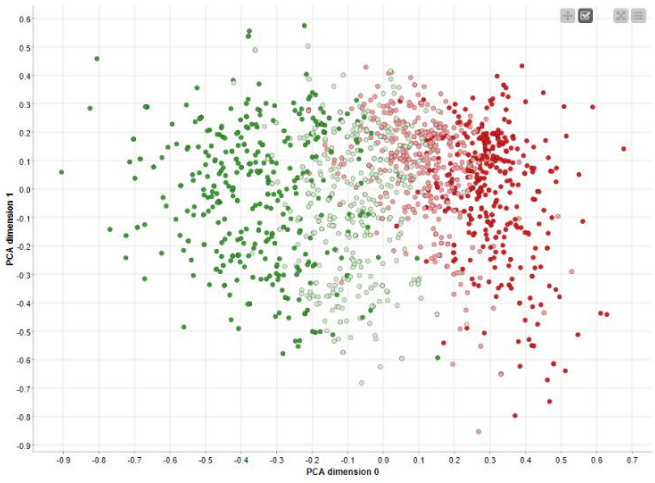

Figura 1. Scatter Plot PCA

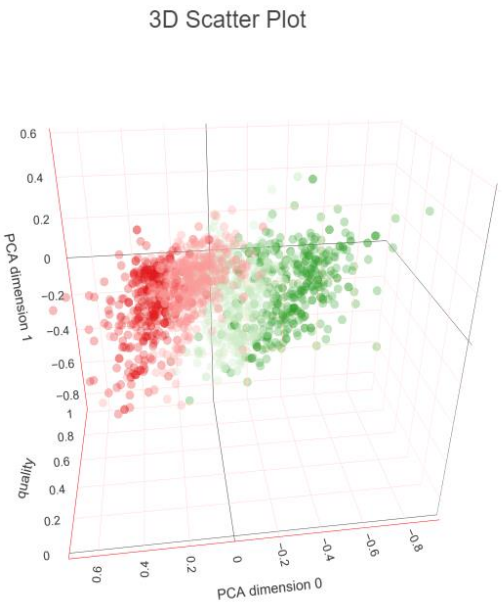

Figura 2. 3D Scatter Plot PCA

T4. Aplicando o nodo *K-means* segmentou-se os dados através de dois clusters, cluster 0 e cluster 1. De seguida, atribuiu-se a cor vemelho para o cluster 0 e o verde para o cluster 1, através do *color manager*. Aplicou-se o nodo *shaper manager* para poder dar formas aos clusters, sendo que para o cluster 0 foi atribuído um quadrado e para o cluster 1 foi atribuído um asterisco. De seguida, utilizou-se o nodo *scatter plot* e o nodo *scatter matrix* para analisar os clusters. Consegue-se

verificar, através do *scatter plot*, que o cluster 0 concentra-se mais na zona com menos ph e o cluster 1 com mais ph. No *scatter matrix*, através das *features residual sugar*, *density* e *alcohol*, verifica-se que o cluster 1 é o que possui menos açúcar e menos densidade, porém possui mais percentagem de álcool. É de salientar que o cluster 1 é relativo aos vinhos com qualidade designada de "high" e "very high". Posteriormente, descarregou-se o *dataset* dos dados relativos a teste, fez-se o tratamento dos mesmos, de igual forma dos dados de treino. Utilizou-se o nodo *cluster Assigner* de maneira a atribuir aos novos dados um cluster. Guardou-se os resultados em dois ficheiros de csv. num apenas guardou-se as *features* id e cluster e noutro todas as colunas.

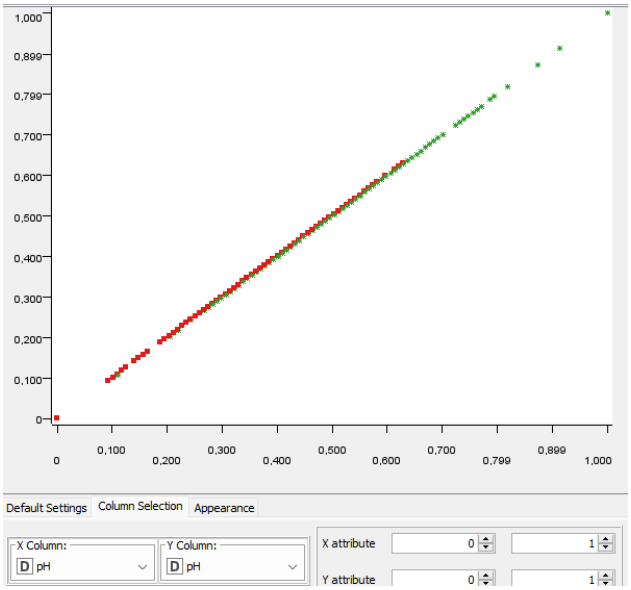

Figura 3. scatter plot

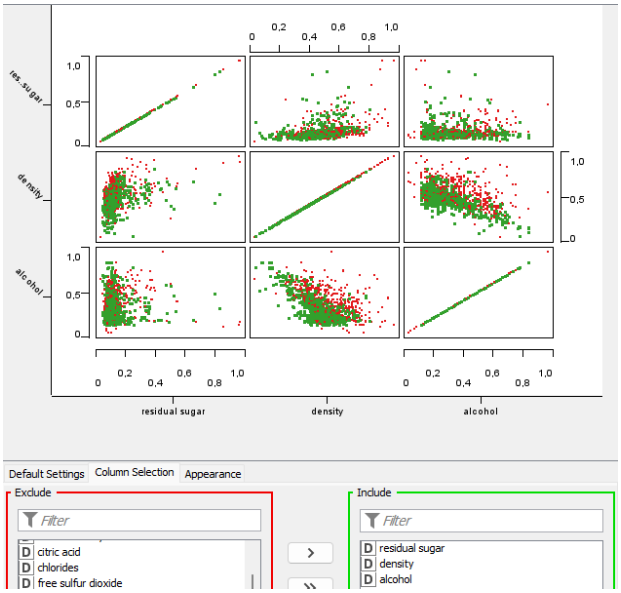

Figura 4. scatter matrix

T5. Para parametrizar o *workflow* de forma a definir o número de bins, de clusters e os títulos dos gráficos, utilizou-se variáveis de fluxo em certos nodos. Primeiramente, usou-se o nodo *Integer Widget* com uma variável chamada "bins_var" e ligou-se ao nodo *Auto Binner*. Nesta variável será possível definir o número de bins que se pretende. A seguir, para definir o número de clusters, utilizou-se o nodo *Integer Widget* com uma variável chamada "clusters_var" ligando ao nodo *K-means*. Depois, utilizou-se o nodo *String Widget* com a variável designada de "títulos_var" para alterar os títulos dos gráficos.

T6. Nesta parte, foi feito um componente designado "data viz" que agrupou os gráficos do PCA de modo a poder visualizar numa página só. Uma vez que os *scatter plot* e o *scatter matrix* eram a nível local não foi possível verificá-los numa só página, no entanto foram agrupados na mesma num componente designado "scatter".

T7. Foi utilizado outro método de segmentação, nomeadamente, o nodo *fuzzy C-Means* de forma a poder comparar. Assim, em relação ao *k-means*, os resultados são bastantes semelhantes, tirando algumas exceções. Aparentemente, a designação dos clusters é diferente de ambos os métodos, ou seja, o cluster 0 do método *k-means* corresponde ao cluster 1 do método *fuzzy C-means*, no entanto são os mesmos registos.

Row ID	es	D alcohol	D quality	S citric ac...	S Cluster
Row0		0.154	0.4	Low	cluster_1
Row1		0.215	0.4	Low	cluster_1
Row2		0.215	0.4	Low	cluster_1
Row3		0.215	0.6	Very High	cluster_0
Row4		0.154	0.4	Low	cluster_1
Row5		0.154	0.4	Low	cluster_1
Row6		0.154	0.4	Low	cluster_1
Row7		0.246	0.8	Low	cluster_1
Row8		0.169	0.8	Low	cluster_1
Row9		0.323	0.4	High	cluster_1
Row10		0.123	0.4	Low	cluster_1

Figura 5. Método K-Means

Row ID	D alcohol	D quality	S citric ac...	D cluster_0	D cluster_1	S Winner ...
Row0	0.154	0.4	Low			cluster_0
Row1	0.215	0.4	Low			cluster_0
Row2	0.215	0.4	Low			cluster_0
Row3	0.215	0.6	Very High			cluster_1
Row4	0.154	0.4	Low			cluster_0
Row5	0.154	0.4	Low			cluster_0
Row6	0.154	0.4	Low			cluster_0
Row7	0.246	0.8	Low			cluster_0
Row8	0.169	0.8	Low			cluster_0
Row9	0.323	0.4	High			cluster_0
Row10	0.123	0.4	Low			cluster_0

Figura 6. Método Fuzzy C-Means