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 7

Tema: Tuning de modelos de clustering, medidas de qualidade e criação de datasets usando APIs públicas

Joana Mota PG45528

Tarefas:

T1. De forma a devolver a todas as cidades portuguesas, foi utilizada a plataforma Open AQ que fornecia uma API que disponibilizava informação sobre a qualidade do ar. Foi utilizado um GET Request para devolver todas as cidades portuguesas.

T2. De seguida, passou-se o JSON para tabela utilizando o nodo *JSON to table* e removeu-se todas as colunas, exceto dos resultados através do nodo *column filter*. Depois foi utilizado o nodo *transpose* e o *nodo JSON to table.*

T3. Pela Open AQ, obteu-se os dados mais recentes do nível de ozono para cada cidade através do nodo *string manipulation* passando depois o *JSON* para tabela. De seguida, filtrou-se as colunas, deixando ficar as mais relevantes. Modificou-se o nome da coluna "value" para ozono, uma vez que corresponde aos valores do nível deste. Por último, passou-se a coluna de ozono para inteiro.

T4. Aplicou-se o nodo *sorter* para ordenar os registos de cada cidade por nível de ozono e utilizou-se o nodo *bar chart* para visualizar o nível de ozono por cidade. Verifica-se que a cidade Viana de Castelo é a que possui o maior nível de ozono, já a cidade de Braga é a que apresenta o nível de ozono mais baixo. De seguida, foi feito exatamente o mesmo processo para as partículas de ar, aplicando o nodo *bar chart* para visualizar o nível de partículas de ar por cidade.

Verifica-se que a Ilha do Faial é a que apresenta maior nível de partículas de ar e a cidade que não apresenta partículas no ar é a cidade de Viana de Castelo.

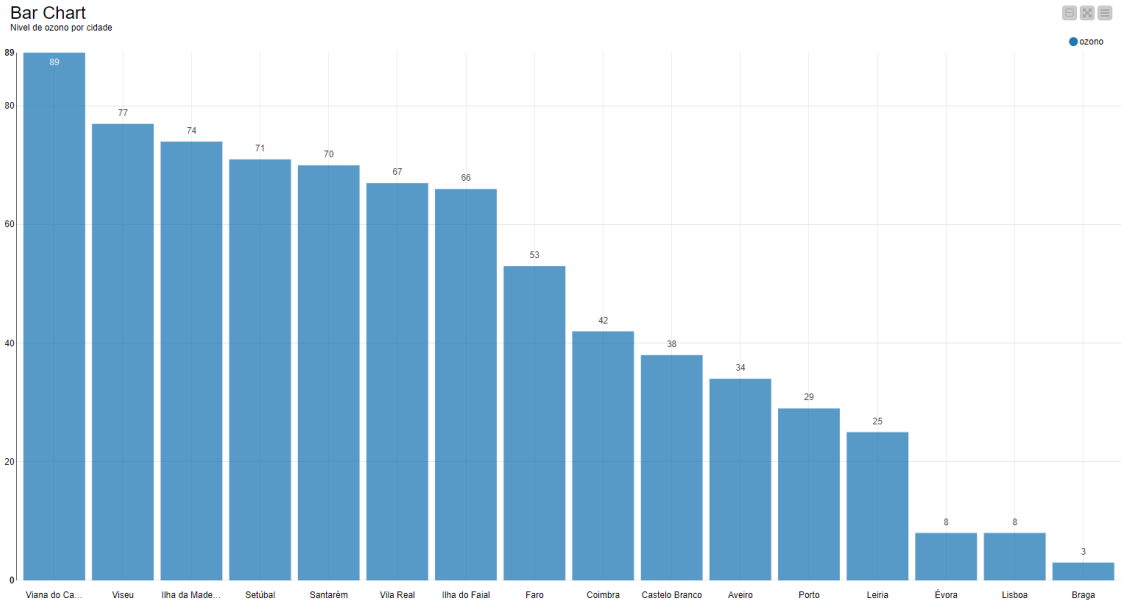

Figura 1 - Nível de ozono por cidade

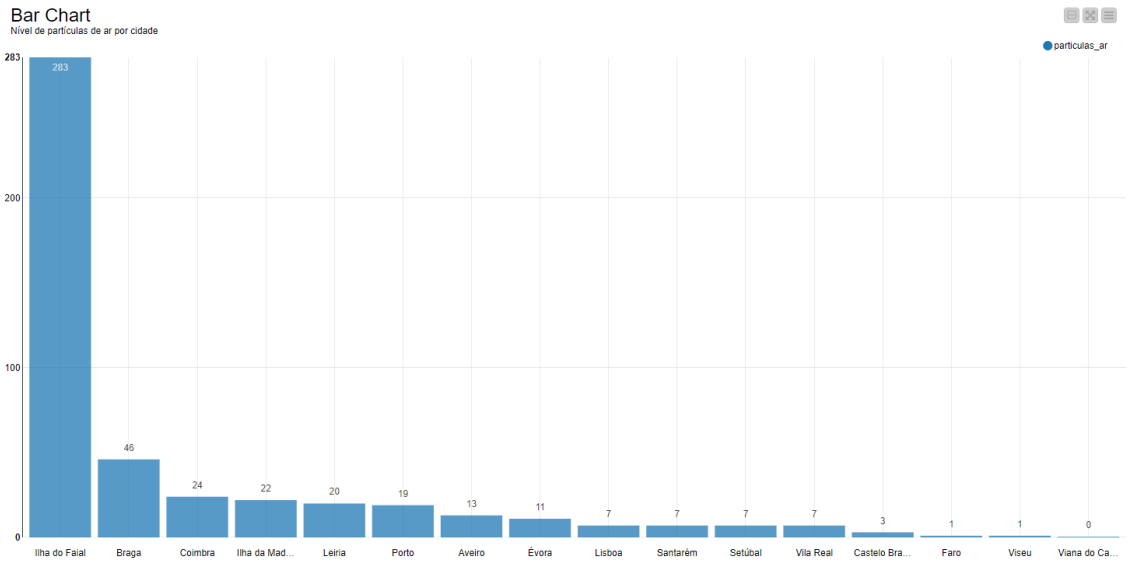

Figura 2 - Nível de partículas de ar por cidade

T5. Através da plataforma de Weather API, devolveu-se os dados disponíveis das temperaturas mínimas e máximas. Utilizou o *JSON to table* e, seguidamente, filtrou-se as colunas com as mais relevantes. Agrupou-se as temperaturas pela média, de forma a obter a média da temperatura mínima e máxima de cada cidade. Visualizou-se os dados através de um *bar chart* e verifica-se que a ilha da madeira apresenta as maiores temperaturas mínimas e máximas em termos médios. Já Vila Real é que a apresenta as temperaturas mais baixas quer seja mínima ou máxima.

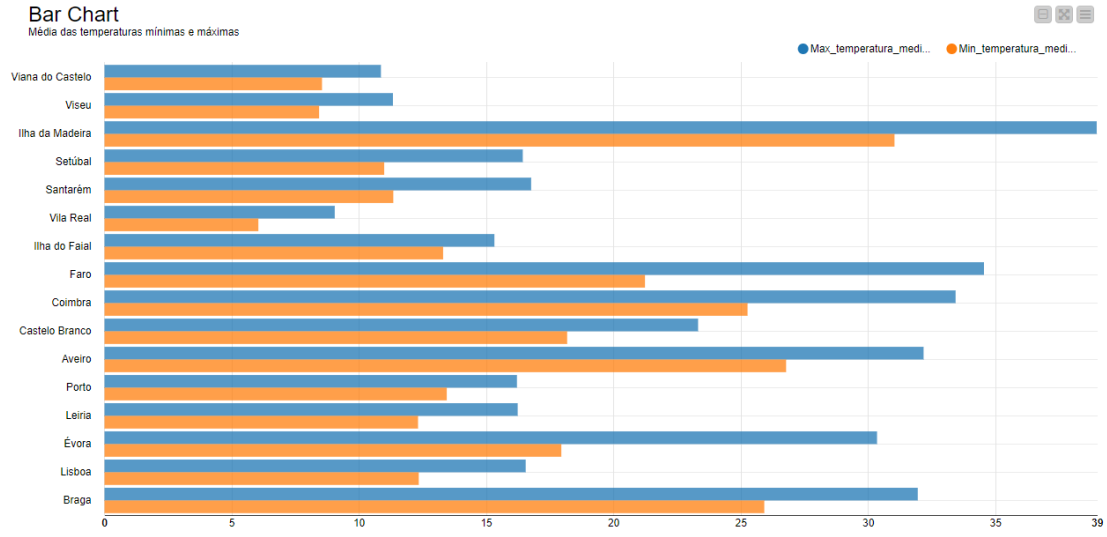

Figura 3 - Média das temperaturas mínimas e máximas por cidade

T6. Selecionou-se o dataset *wine quality* e fez-se o tratamento deste.

T7. Foi aplicado o nodo scatter matrix para visualizar diagramas de dispersão, bem como um PCA para projetar os dados em duas dimensões. No entanto, através de um scatter plot não se consegue concluir nada, uma vez que só aparece dados de uma dimensão.

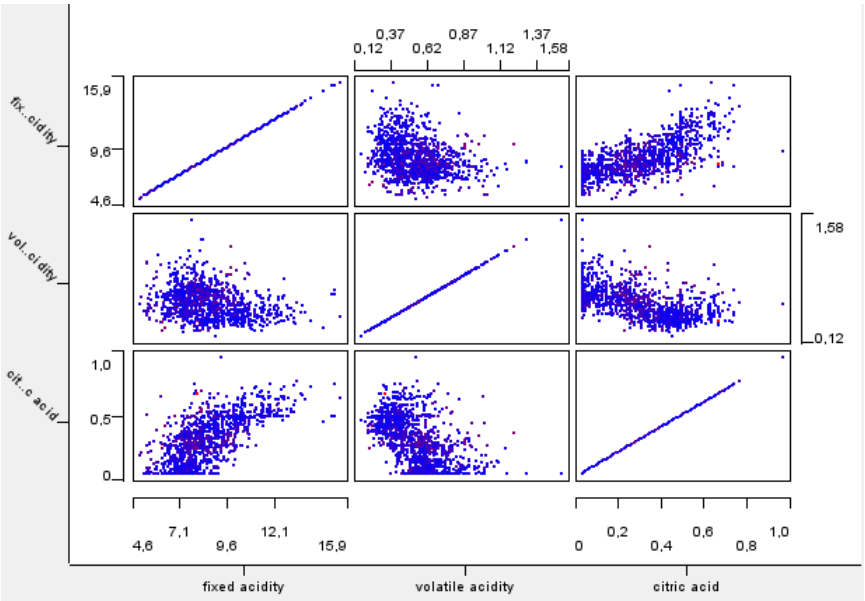

Figura 4 - Diagramas de dispersão

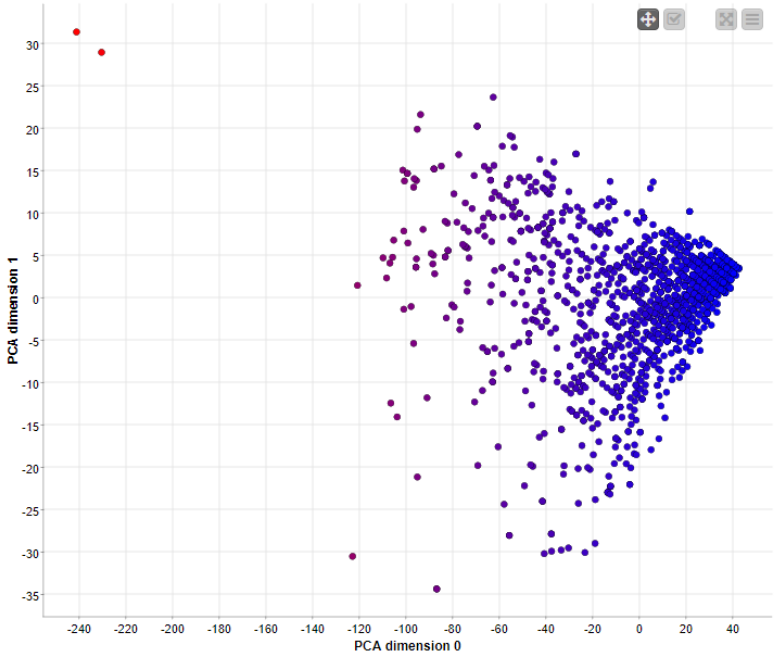

Figura 5 - Diagrama de dispersão com duas dimensões

T8. Utilizando K-medoids, aplicou-se o método do cotovelo, onde foi possível identificar que 3 é o número ótimo de clusters para o modelo, como se verifica no gráfico seguinte. Verifica-se, também, que o valor com maior diferença de MAE (erro) é o que apresenta 3 clusters.

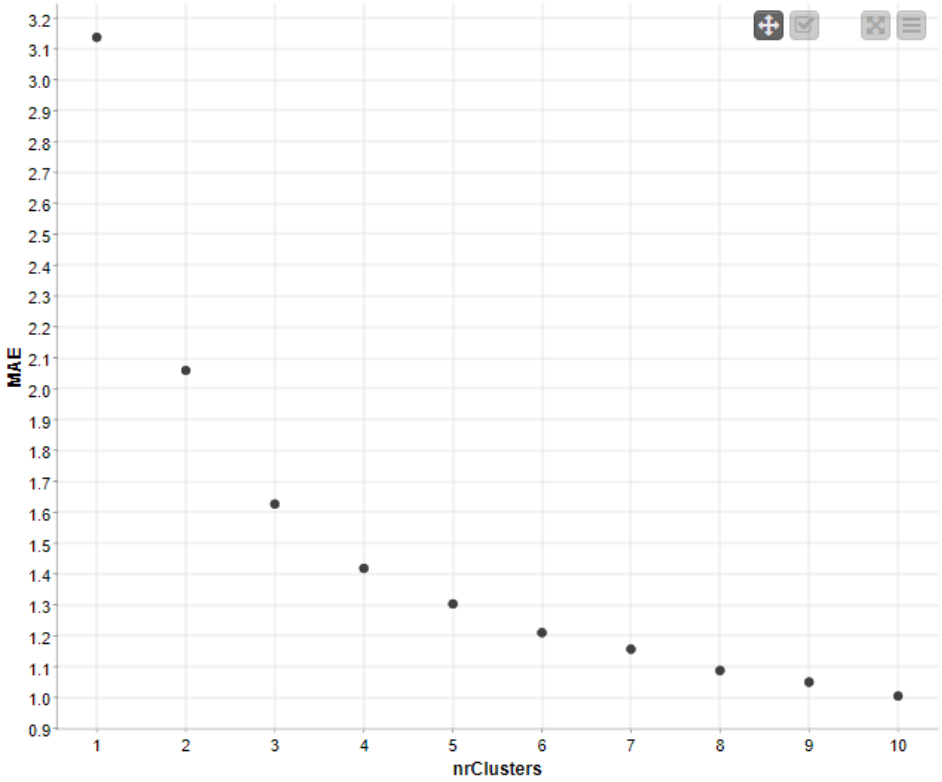

Figura 6 - Número de clusters pelo método de K-medoids

Row ID	D MAE	I nrClusters	I Iteration	D MAE_dif
Row0#1	2.059	2	1	1.078
Row0#2	1.626	3	2	0.433
Row0#3	1.418	4	3	0.208
Row0#4	1.303	5	4	0.115
Row0#5	1.21	6	5	0.093
Row0#7	1.087	8	7	0.069
Row0#6	1.157	7	6	0.053
Row0#9	1.005	10	9	0.045
Row0#8	1.05	9	8	0.038
Row0#0	3.137	1	0	0

Figura 7 - MAE

T9. Aplicou-se um componente designado de "gráficos" com o objetivo de visualizar os gráficos realizados pelo método de cotovelo, apesar de só apresentar um gráfico. Para além disto, aplicou-se o nodo *integer widget* que permite criar uma variável para aplicar ao método clustering. Esta variável é designada por "clusters_var" e será possível definir o número de clusters.

T10. Utilizando K-Means, aplicou-se o método do cotovelo, onde foi possível identificar que 3 é o número ótimo de clusters para o modelo, como se verifica no gráfico seguinte. E através da diferença de MAE (erro), verifica-se que aquele que tem maior valor de diferença é o que apresenta 3 clusters.

Comparando com t8, ambos têm como número ótimo 3 clusters, assim é indiferente o método de clustering a aplicar no modelo.

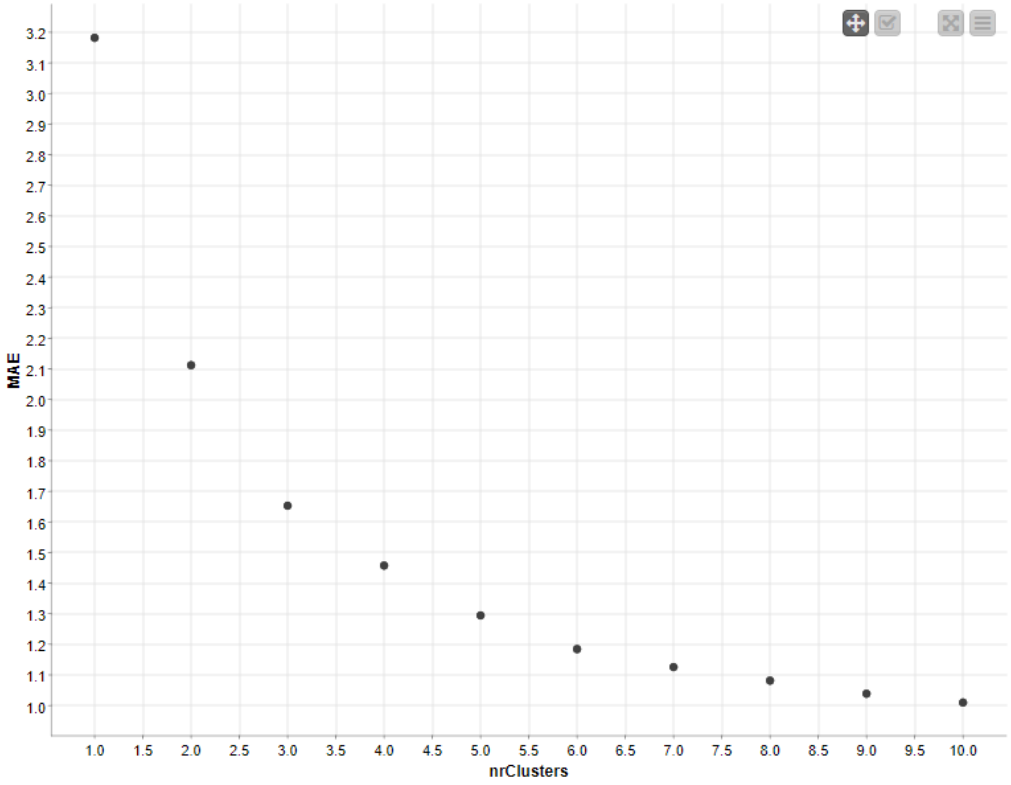

Figura 8 - Número de Clusters pelo método de K-Means

Table "default" - Rows: 10 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	D MAE	I nrClusters	I Iteration	D MAE_dif
Row0#1	2.112	2	1	1.07
Row0#2	1.653	3	2	0.459
Row0#3	1.457	4	3	0.196
Row0#4	1.294	5	4	0.163
Row0#5	1.184	6	5	0.11
Row0#6	1.126	7	6	0.059
Row0#7	1.081	8	7	0.044
Row0#8	1.039	9	8	0.043
Row0#9	1.01	10	9	0.029
Row0#0	3.182	1	0	0

Figura 9 - MAE