

Universidade do Minho

Departamento de Informática

Mestrado em Engenharia Informática

Perfil de Machine Learning Fundamentos e Aplicações

Sistemas Baseados em Similaridade

1º Ano, 1º Semestre

Ano letivo 2020/2021

Enunciado Prático nº 7

12 de dezembro de 2020

Joel Costa Carvalho

PG42837

Índice

[Tarefas - 5 -](#_Toc58700053)

[T1 – Analisar a Open AQ Platform, a qual fornece uma API que disponibiliza dados sobre a qualidade do ar (ver https://docs.openaq.org/). Efetuar um GET request que devolva todas as cidades portuguesas atualmente disponibilizadas pela Open AQ; - 5 -](#_Toc58700054)

[T2 - Analisar e tratar a resposta JSON obtida em T1 - 6 -](#_Toc58700055)

[a) Passar o *JSON* para tabela, sendo que na configuração do nodo devem garantir (1.) a expansão dos *arraysJSON* para colunas e (2.) que o parâmetro *children expansion* vá, exatamente, até ao nível 2; - 6 -](#_Toc58700056)

[b) Remover todas as colunas exceto as dos resultados; - 6 -](#_Toc58700057)

[*c)* Obter a transposta e aplicar novamente o nodo *JSON* *to Table* - 7 -](#_Toc58700058)

[T3 - Para cada cidade portuguesa disponibilizada pela Open AQ: - 7 -](#_Toc58700059)

[a) Obter os dados mais recentes sobre o nível de ozono, limitando o número de resultados obtidos a apenas 1 (ver https://docs.openaq.org/#api-Latest); - 7 -](#_Toc58700060)

[b) Passar o JSON a tabela e filtrar todas as colunas exceto a *city*, *parameter*, *value* e *lastUpdated*. Fazer o *cast* da coluna *value* de *double* para inteiro; - 8 -](#_Toc58700061)

[T4 - Ordenar os registos de cada cidade por nível de ozono e implementar técnicas de visualização de dados numa vista web. Obter também dados referentes a outros parâmetros ambientais; - 9 -](#_Toc58700062)

[T5 - Obter e visualizar dados de outras plataformas como a OpenWeatherMaps ou a OpenUV, por exemplo. - 12 -](#_Toc58700063)

[T6 - Consultar, selecionar e tratar um dataset sobre o qual deverá ser, numa tarefa posterior, aplicado um método de clustering; - 14 -](#_Toc58700064)

[T7 - De forma a tentar identificar, visualmente, clusters no dataset devem ser utilizados diagramas de dispersão. Deve também ser aplicada uma Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões. O que se conclui? - 14 -](#_Toc58700065)

[T8 - Aplicar o método do cotovelo de forma a identificar o número ótimo de clusters, utilizando k-Medoids como método de clustering. Utilizar a métrica Mean Average Error (MAE) como medida de qualidade; - 15 -](#_Toc58700066)

[T9 - Desenvolver uma vista que permita a um utilizador analisar os gráficos gerados pelo método do cotovelo. Através de widgets, permitir que um utilizador defina o número de clusters a ser utilizado pelo método de clustering; - 16 -](#_Toc58700067)

[T10 - Experimentar o método de clustering k-Means e comparar a sua performance com os resultados obtidos em T8. - 17 -](#_Toc58700068)

Índice de figuras

[Figura 1 - Efetuar pedido - 5 -](#_Toc58699766)

[Figura 2– Resultado - 5 -](#_Toc58699767)

[Figura 3 - Visão Geral - 6 -](#_Toc58699768)

[Figura 4 - Configurações gerais - 6 -](#_Toc58699769)

[Figura 5 - Resultado de JSON para tabela - 6 -](#_Toc58699770)

[Figura 6 - Remover colunas exceto os resultados - 6 -](#_Toc58699771)

[Figura 7 - Aplicar transposta e transformar em tabela - 7 -](#_Toc58699772)

[Figura 8 - Construção do URL - 7 -](#_Toc58699773)

[Figura 9 - Output final do URL trabalhado - 7 -](#_Toc58699774)

[Figura 10 - Efetuar pedido por cada cidade - 8 -](#_Toc58699775)

[Figura 11 - Resultado dos pedidos GET - 8 -](#_Toc58699776)

[Figura 12 - T3. Visão Geral - 9 -](#_Toc58699777)

[Figura 13 - Componente Tarefa 4 - 9 -](#_Toc58699778)

[Figura 14 - o3 por cidade - 10 -](#_Toc58699779)

[Figura 15 - Percentagem de o3 por cidade - 10 -](#_Toc58699780)

[Figura 16 - Visão Geral da exploração de outros parâmetros - 11 -](#_Toc58699781)

[Figura 17 - Box Plot. Monoxido de carbono - 11 -](#_Toc58699782)

[Figura 18 - Partículas com diâmetro menor que 10 - 12 -](#_Toc58699783)

[Figura 19 - T5. Visão geral - 12 -](#_Toc58699784)

[Figura 20 - Temperaturas em Vila Verde entre 10/12 e 12/12 - 13 -](#_Toc58699785)

[Figura 21 - Exploração de dados - 13 -](#_Toc58699786)

[Figura 22 - Plot Temperatura/Velocidade do Vento - 14 -](#_Toc58699787)

[Figura 23 - Análise do dataset vinhos e aplicação do método k-Medoids - 14 -](#_Toc58699788)

[Figura 24– Visão geral - 15 -](#_Toc58699789)

[Figura 25 - PCA - 15 -](#_Toc58699790)

[Figura 26 - Ciclo com k-Medoids para calcular MAE - 15 -](#_Toc58699791)

[Figura 27– Cálculo da Mean Average Error (MAE) - 16 -](#_Toc58699792)

[Figura 28 - Elbow graficamente - 16 -](#_Toc58699793)

[Figura 29 - Widget conectado ao loop - 16 -](#_Toc58699794)

[Figura 30 - Componente com widget - 17 -](#_Toc58699795)

[Figura 31 - MAE com clustering k-Means - 17 -](#_Toc58699796)

[Figura 32 - k-Meloids vs k-Means - 17 -](#_Toc58699797)

[Figura 33 - Plots k-Medoids vs k-Means - 18 -](#_Toc58699798)

**Tarefas**

**T1 – Analisar a *Open AQ Platform*, a qual fornece uma *API* que disponibiliza dados sobre a qualidade do ar (ver https://docs.openaq.org/). Efetuar um *GET request* que devolva todas as cidades portuguesas atualmente disponibilizadas pela *Open AQ*;**

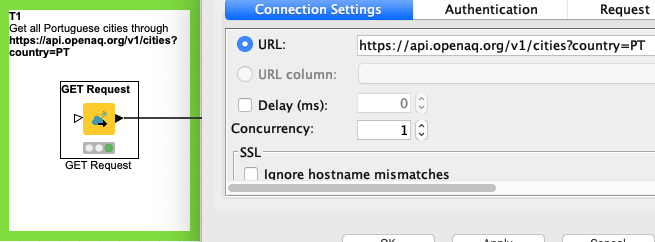


Figura - Efetuar pedido

Uma imagem com texto

Descrição gerada automaticamente

Figura – Resultado

**T2 - Analisar e tratar a resposta *JSON* obtida em T1**

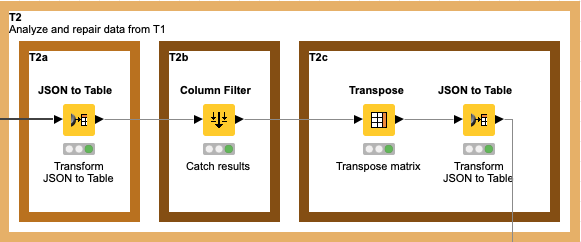


Figura - Visão Geral

1. **Passar o *JSON* para tabela, sendo que na configuração do nodo devem garantir (1.) a expansão dos *arraysJSON* para colunas e (2.) que o parâmetro *children expansion* vá, exatamente, até ao nível 2;**

Uma imagem com texto

Descrição gerada automaticamente

Figura - Configurações gerais

Uma imagem com texto

Descrição gerada automaticamente

Figura - Resultado de JSON para tabela

1. **Remover todas as colunas exceto as dos resultados;**

Uma imagem com texto

Descrição gerada automaticamente

Figura - Remover colunas exceto os resultados

1. **Obter a transposta e aplicar novamente o nodo *JSON* *to Table***

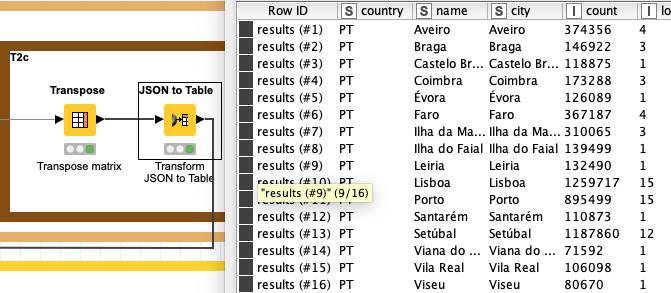


Figura - Aplicar transposta e transformar em tabela

**T3 - Para cada cidade portuguesa disponibilizada pela *Open AQ*:**

1. **Obter os dados mais recentes sobre o nível de ozono, limitando o número de resultados obtidos a apenas 1 (ver** [**https://docs.openaq.org/#api-Latest**](https://docs.openaq.org/#api-Latest)**);**

Para realizar este pedido foi necessário aplicar 2 transformações ao nome das cidades. A primeira consistiu em substituir os espaços por ‘%20’, de modo a conseguir realizar o pedido GET via *browser* e por fim a criação de uma coluna *‘url’* para facilitar o processo de iteração.

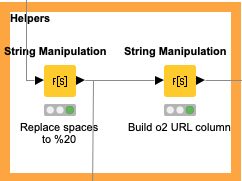
****

Figura - Construção do URL

Uma imagem com mesa

Descrição gerada automaticamente

Figura - Output final do URL trabalhado

Posto isto, substitui a variável URL pela variável de fluxo/coluna ‘url’.

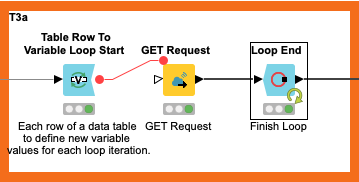


Figura - Efetuar pedido por cada cidade

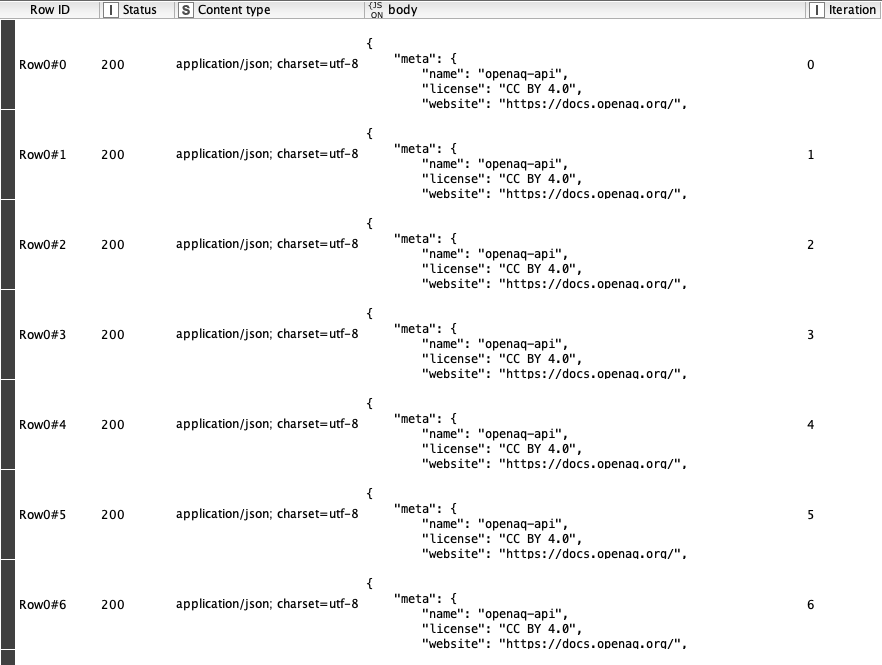


Figura - Resultado dos pedidos GET

1. **Passar o JSON a tabela e filtrar todas as colunas exceto a *city*, *parameter*, *value* e *lastUpdated*. Fazer o *cast* da coluna *value* de *double* para inteiro;**

Os nodos *‘JSON to Table’* são similares à tarefa 2, relativamente aos restantes servem para filtrar e executar o *cast requerido.*

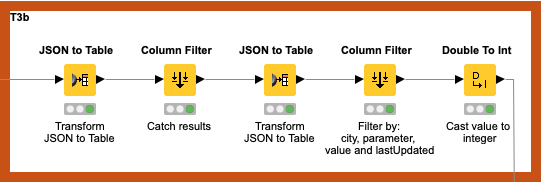


Figura - T3. Visão Geral

**T4 - Ordenar os registos de cada cidade por nível de ozono e implementar técnicas de visualização de dados numa vista web. Obter também dados referentes a outros parâmetros ambientais;**

Todo o fluxo foi guardado dentro de um componente. Para além da ordenação e visualização dos dados do ozono pelas cidades portuguesas foi analisado e explorado também os dados do monóxido de carbono e partículas com diâmetro menor que 10.

Observando a Figura 16., posso concluir que:

* + 1. Mediana igual a 253;
    2. Faro é *outlier;*
    3. Aveiro com 186, contém o valor mais baixo com a emissão de monóxido de carbono.

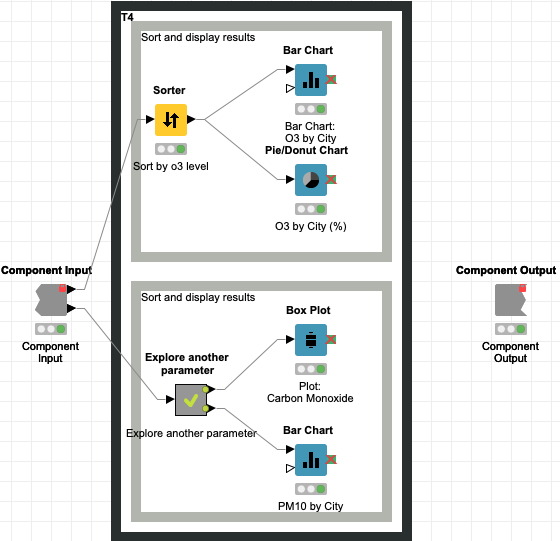


Figura - Componente Tarefa 4

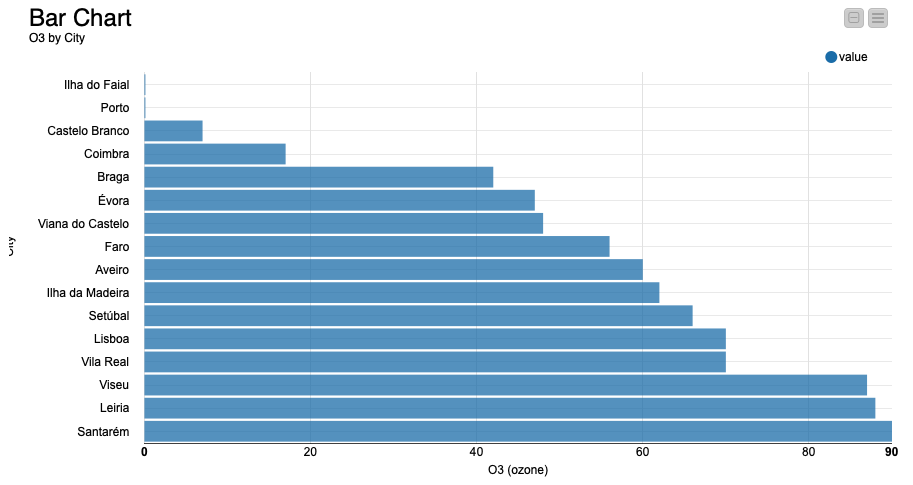


Figura - o3 por cidade

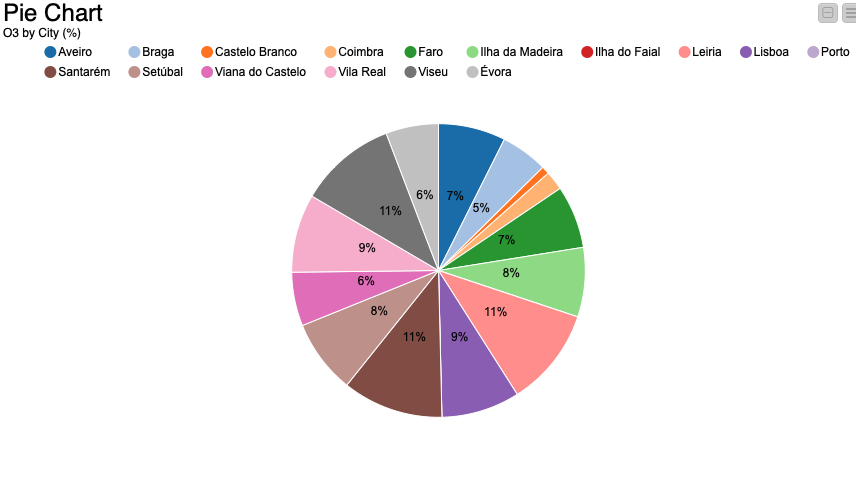


Figura - Percentagem de o3 por cidade

Obtenção de dados referentes a outros parâmetros ambientais:

* + 1. Co – *Carbon Monoxide*
    2. Pm10 – *Particulate Matter 10*

Uma imagem com mesa

Descrição gerada automaticamente

Figura - Visão Geral da exploração de outros parâmetros

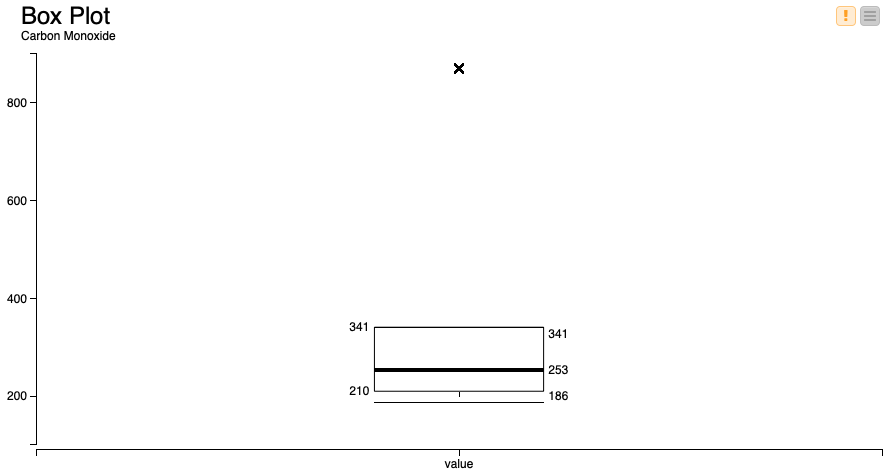


Figura - Box Plot. Monoxido de carbono

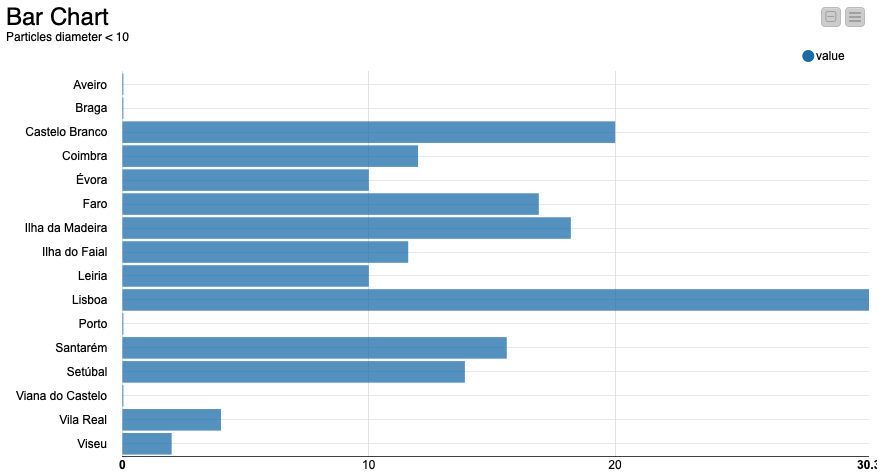


Figura - Partículas com diâmetro menor que 10

**T5 - Obter e visualizar dados de outras plataformas como a *OpenWeatherMaps* ou a *OpenUV*, por exemplo.**

A plataforma explorada foi *OpenWeatherMaps,* esta API permite obter uma diversidade de dados relativos à meteorologia de uma ou várias localidades.

A localidade analisada foi Vila Verde correspondente ao ID número 2732326.

Esta tarefa, arrecadada num componente, efetua um pedido *GET* e a informação obtida é agregada a um ficheiro csv. O objetivo consiste em guardar informação de diferentes dias, de modo, a obter resultados interessantes.

Consequentemente, para manter o ciclo minimamente automatizado, a cada *GET* a coluna *header* é adicionada ao ficheiro, daí ser necessário removê-la na leitura do mesmo. Para isso, foi aplicado o modulo ao número de cada linha, de modo a obter a linhas ímpares e concludentemente apaga-las.

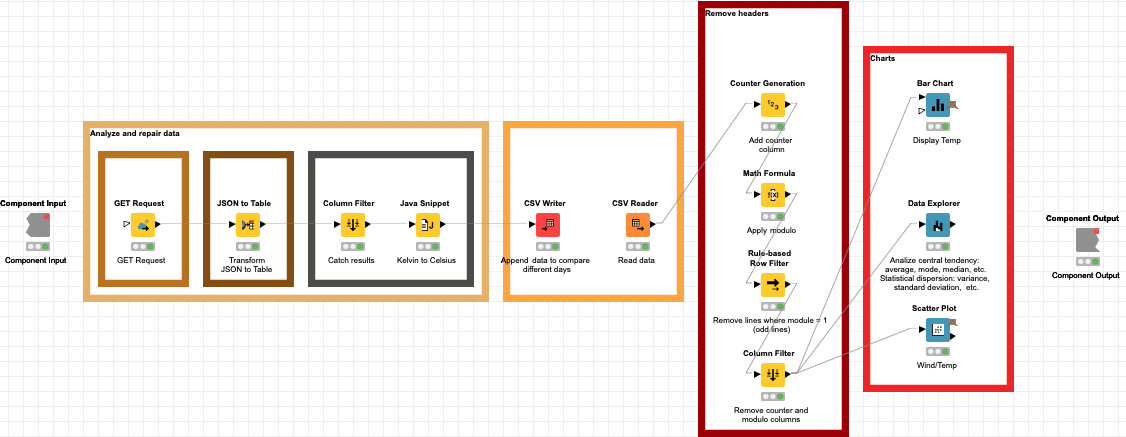


Figura - T5. Visão geral

Média das temperaturas ao longo dos dias 10/12-12/12 de 2020, em Vila Verde:

1. Sentida: 11,7º C
2. Mínima: 10.7º C
3. Máxima: 15º C

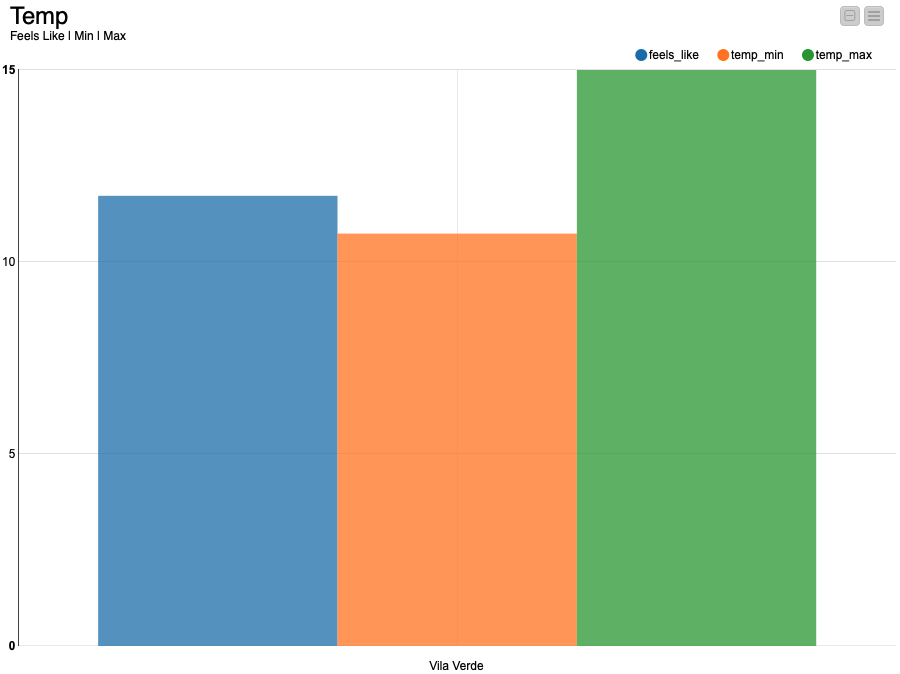


Figura - Temperaturas em Vila Verde entre 10/12 e 12/12

Uma imagem com mesa

Descrição gerada automaticamente

Figura - Exploração de dados

Analisando a figura abaixo, concluí que quanto maior a velocidade do vento mais elevada é a temperatura ambiente.

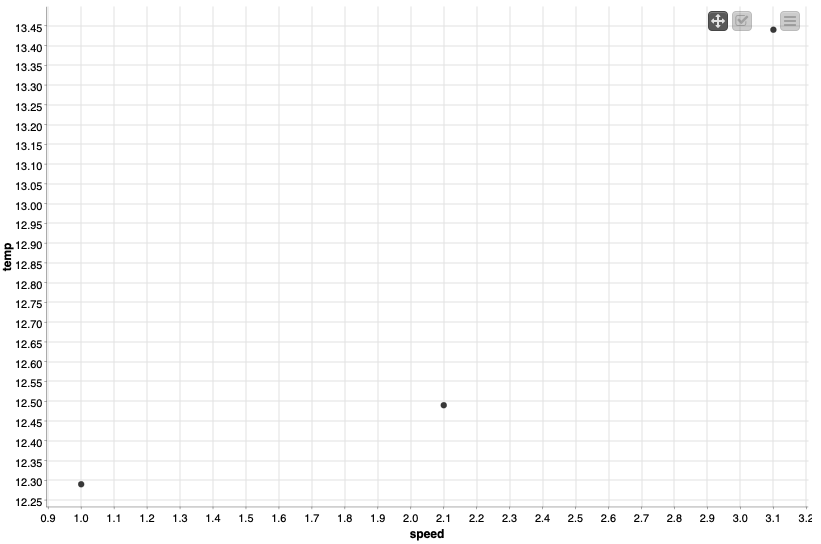


Figura - Plot Temperatura/Velocidade do Vento

**T6 - Consultar, selecionar e tratar um *dataset* sobre o qual deverá ser, numa tarefa posterior, aplicado um método de *clustering*;**

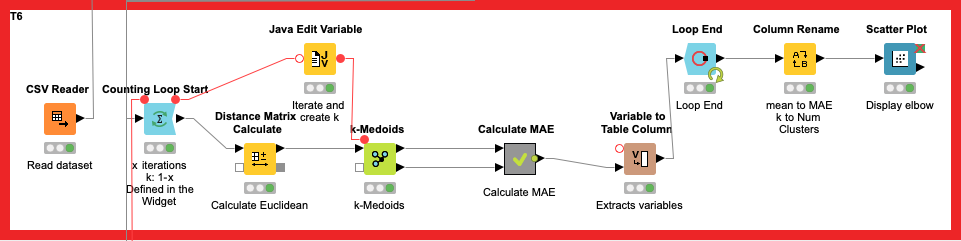
****

Figura - Análise do dataset vinhos e aplicação do método k-Medoids

**T7 - De forma a tentar identificar, visualmente, clusters no *dataset* devem ser utilizados diagramas de dispersão. Deve também ser aplicada uma Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões. O que se conclui?**

Para uma melhor análise dos dados, na feature quality o símbolo ‘=’ foi removido e todos os valores inteiros foram normalizados e de forma a diminuir a dimensionalidade e obter melhores valores, foi aplicado um PCA (Figura 25.).

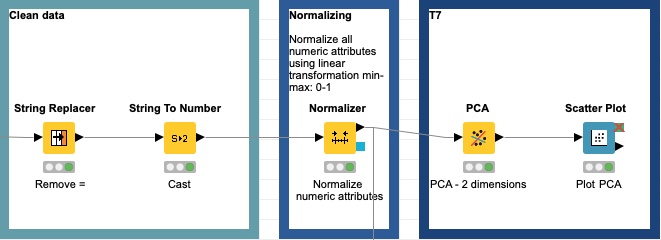
****

Figura – Visão geral

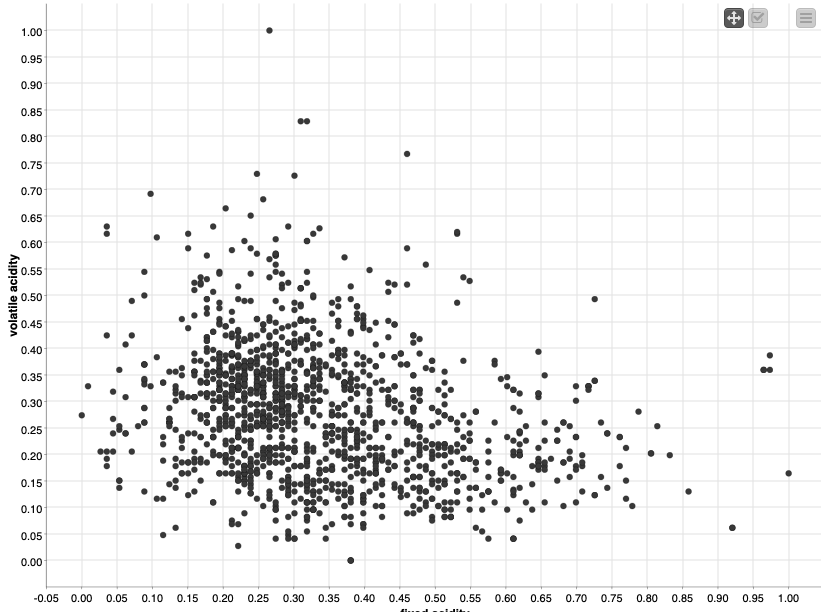
****

Figura - PCA

**T8 - Aplicar o método do cotovelo de forma a identificar o número ótimo de *clusters*, utilizando *k-Medoids* como método de *clustering*. Utilizar a métrica *Mean Average Error* (MAE) como medida de qualidade;**

Número ótimo de *clusters*: 2 (Figura 28.)

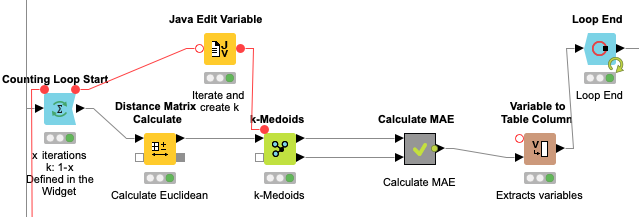
****

Figura - Ciclo com k-Medoids para calcular MAE

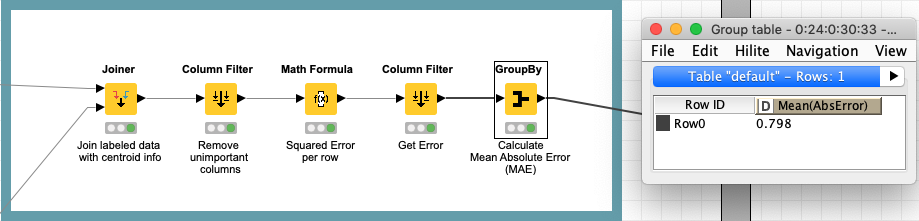
****

Figura – Cálculo da Mean Average Error (MAE)

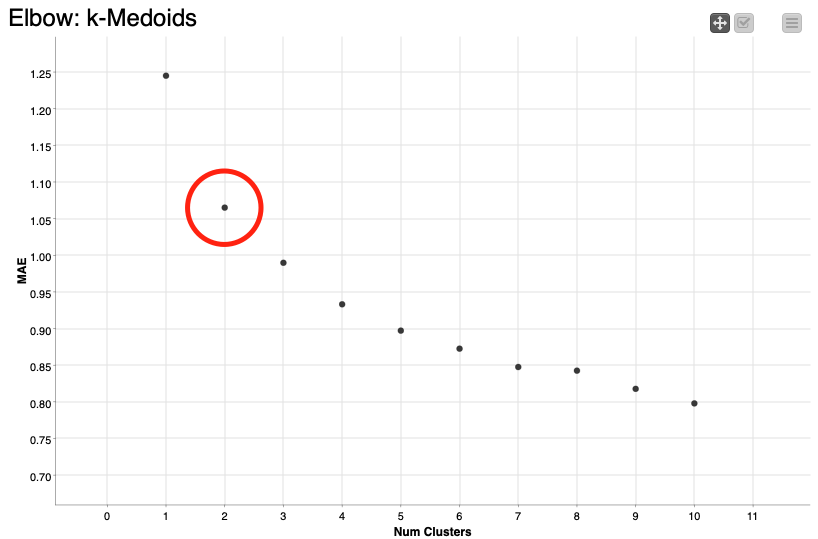
****

Figura - Elbow graficamente

**T9 - Desenvolver uma vista que permita a um utilizador analisar os gráficos gerados pelo método do cotovelo. Através de *widgets*, permitir que um utilizador defina o número de clusters a ser utilizado pelo método de *clustering*;**

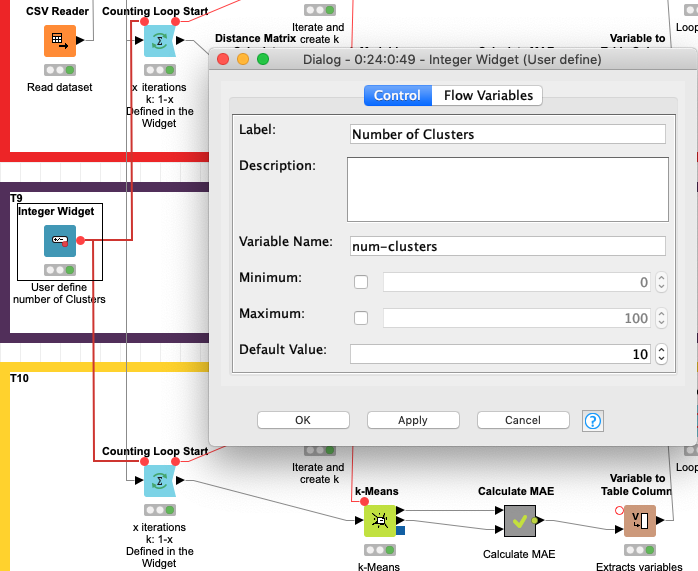
****

Figura - Widget conectado ao loop

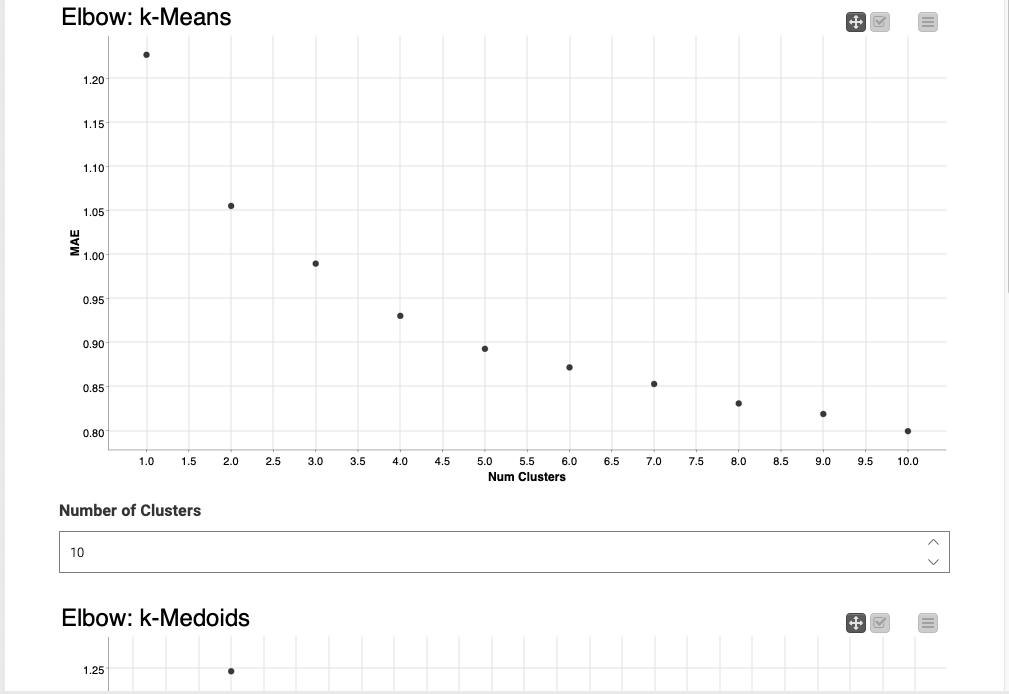
****

Figura - Componente com widget

**T10 - Experimentar o método de *clustering* *k-Means* e comparar a sua performance com os resultados obtidos em T8.**

Número ótimo de *clusters*: 2 (Figura 33.)

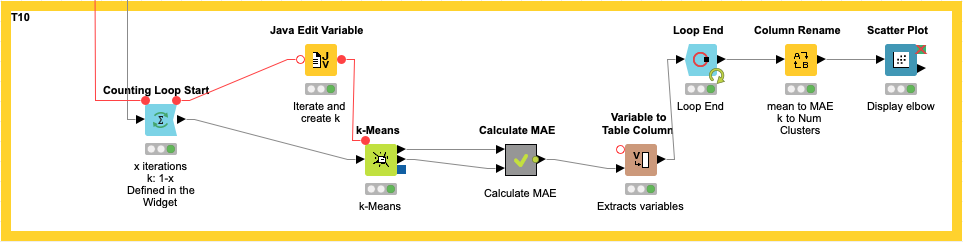
****

Figura - MAE com clustering k-Means

Uma imagem com mesa

Descrição gerada automaticamente

Figura 32 - k-Meloids vs k-Means

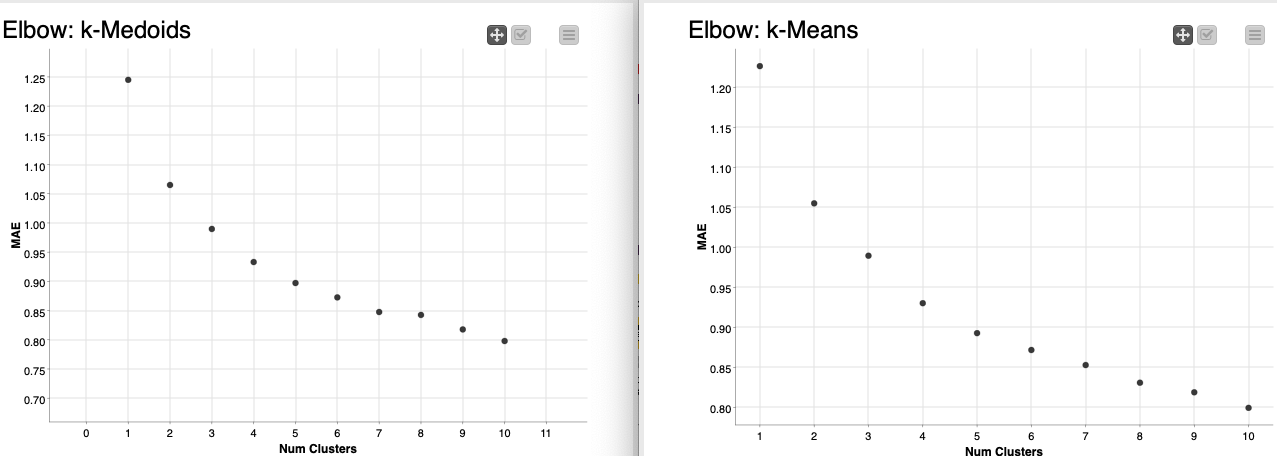


Figura 33 - Plots k-Medoids vs k-Means

|  |  |  |
| --- | --- | --- |
|  | *k-Medoids* | *k-Means* |
| Nº ótimo de *clusters* | 2 | 2 |
| Tempo de execução (K=10) | Demorado | Rápido |
| Maior diferença entre *clusters* | 1.245 – 1.065 = 0.18 | 1.227 – 1.055 = 0.172 |
| MAE (K=10) | Similar | |