Sistemas Baseados em Similaridade

Enunciado Prático Individual 4

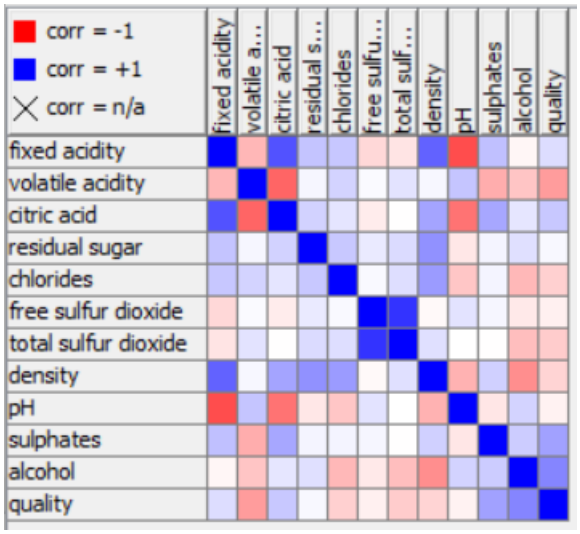
José Virgílio Silva Loureiro (PG52252)

T1. Nesta tarefa carreguei no Knime os dois *datasets* usando o nodo "CSV Reader" para recolher os dados referentes à qualidade do vinho.

Para a exploração dos dados criei um metanode com o nome Exploração de Dados, onde primeiramente utilizo o nodo "Data Explorer", de onde podemos tirar as seguintes conclusões:

- Não existem missing values;
- O atributo que apresenta um desvio-padrão maior é o "total sulfur dioxide".
- Os atributos que apresentam maior dispersão em torno da média (Variância) são o "total sulfur dioxide "e a "free sulfur dioxide".
- O atributo que apresenta maior Skewness e maior Kurtosis é a "chlorides".

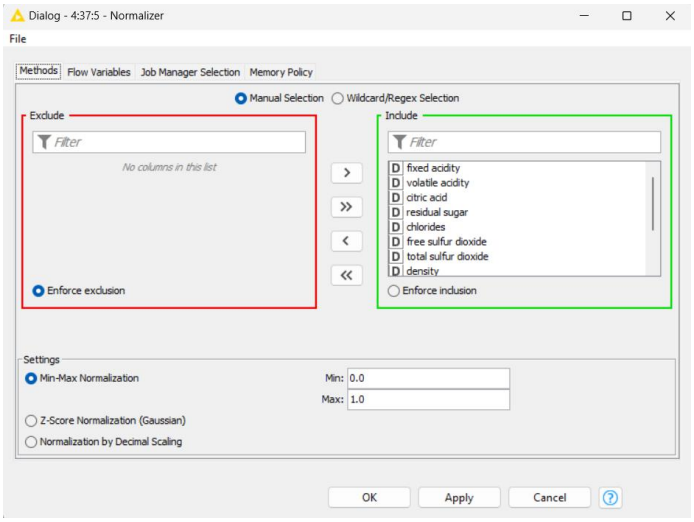
Seguidamente com o node "Rank correlation" consigo observar que existe grande correlação entre o "total sulfur dioxide" com o "free sulfur dioxide", consigo ver também que existe uma correlação modesta entre o "álcool" e a "quality" e ainda que existe uma correlação negativa significativa entre o "pH" e a "fixed acidity".



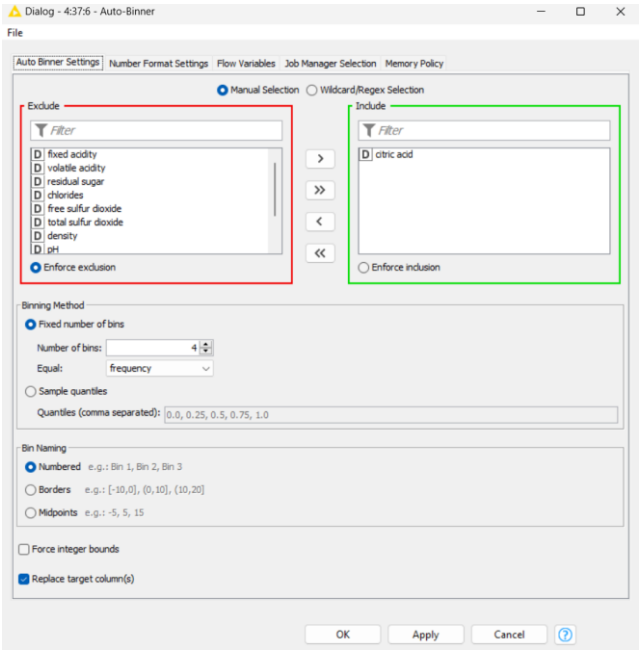
T2. A) Nesta tarefa manipulei o atributo quality do dataset com o node "String manipulation" para eliminar o "=" que estava presente na coluna e estava a mais, depois usei o node "String to number" para tornar esse atributo numérico.

quality Number (dou...
5
5
5
6
5
5
5
7

T2.B) Nesta tarefa utilizei o node ”Normalizer” para normalizar todos os atributos numéricos do dataset, dentro deste nodo utilizei a configuração da transformação linear Min-Max entre 0 e 1 tal como pedido.

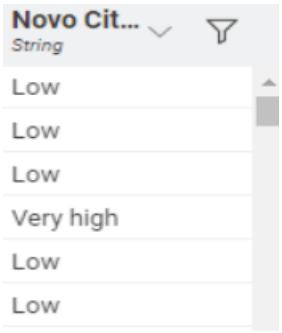


T2.C) e D) Nesta tarefa utilizei o node “Auto binner” para criar 4 bins de igual frequência sobre o valor do atributo ”citric acid”, apliquei também a opção de substituir a nova coluna de bins pela coluna antiga da feature “citric acid”.

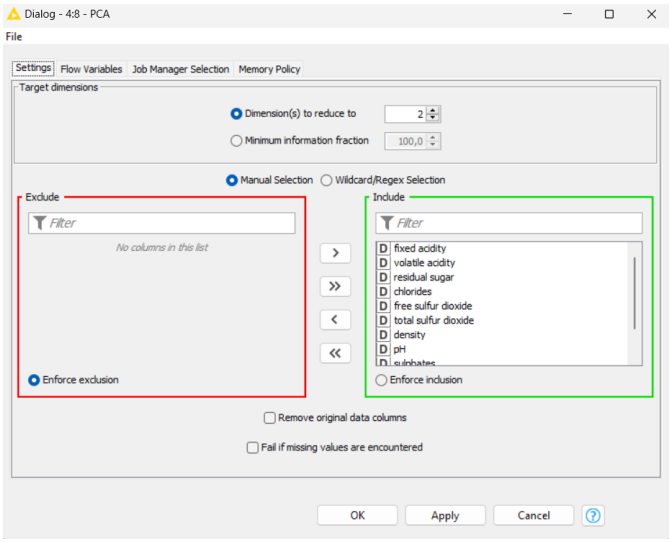


Depois utilizei o node “Table Creator”, onde criei a tabela ao lado, com o objetivo de renomear cada Bin, o primeiro corresponde a Low, o segundo a Medium, o terceiro a High e o quarto a Very High. Depois usei também utilizei o node “Value Lockup” para substituir os respetivos Bins.

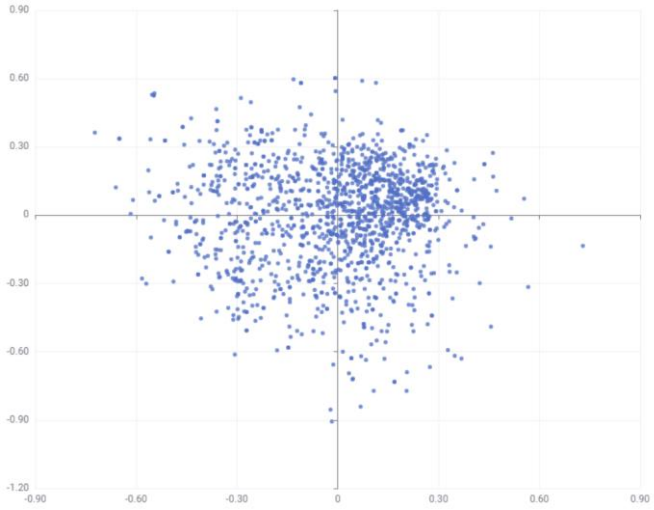
S Citric acid	S Novo Cit...
Bin 1	Low
Bin 2	Medium
Bin 3	High
Bin 4	Very high

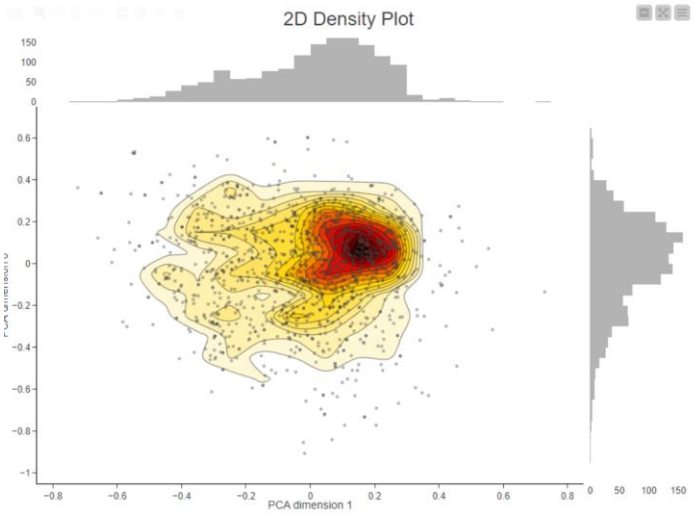


T3.A) e B) Nesta tarefa comecei por utilizar o node “PCA” para fazer uma redução de dimensão dos dados para apenas 2 dimensões tal como pedido.

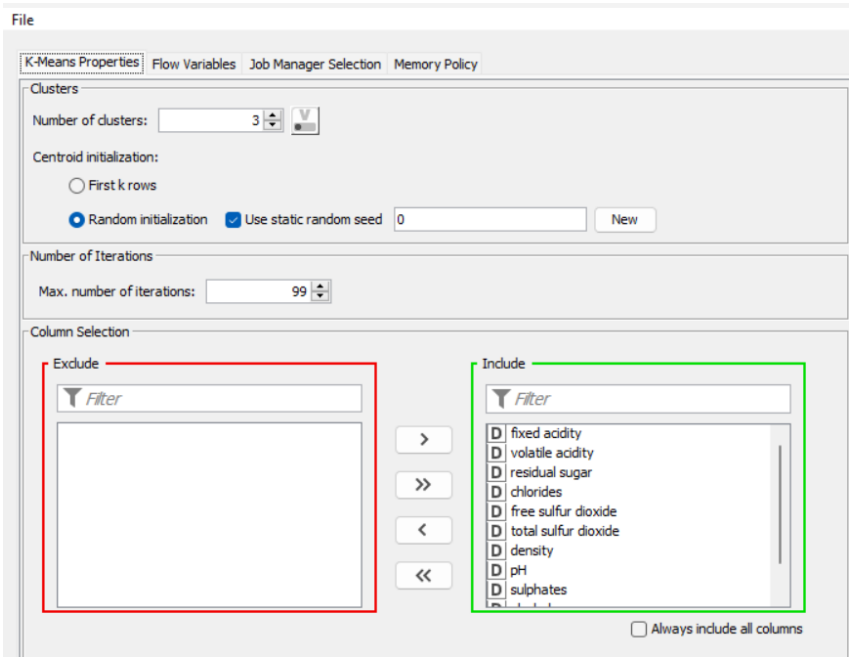


Depois, utilizando os nodes “Scatter Plot” e “2D Density Plot(Legacy)” obtive as visualizações dos dados obtidos pelo PCA.

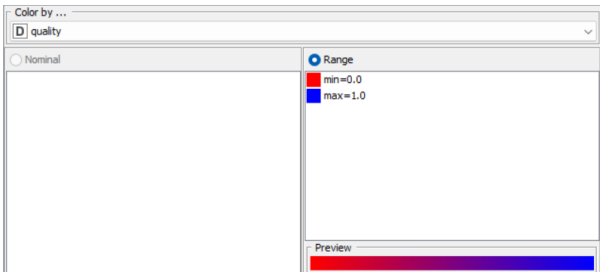


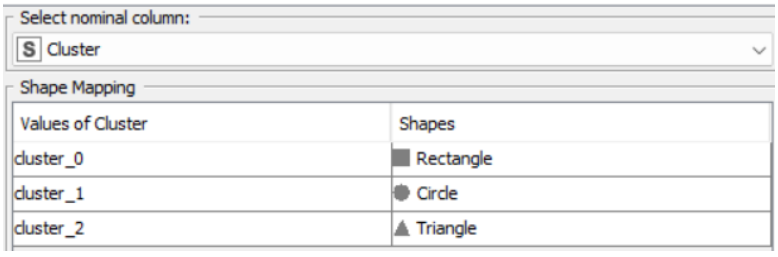


T4. A) Nesta tarefa apliquei o node “K-means” para segmentar o dataset em 3 clusters.

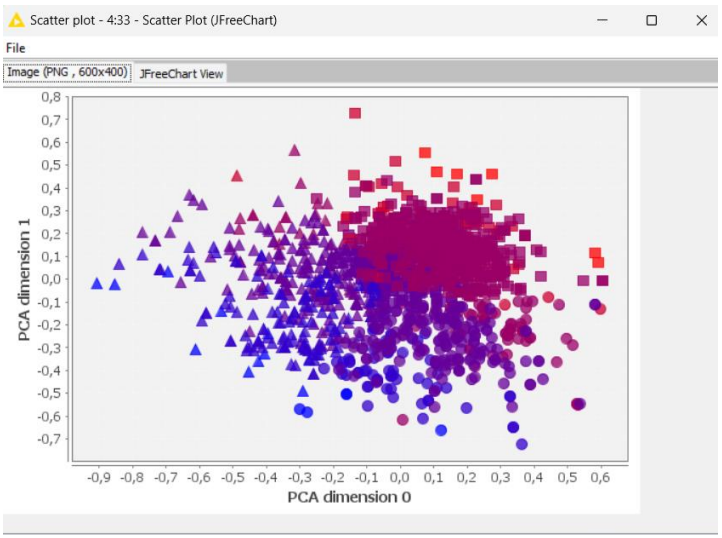


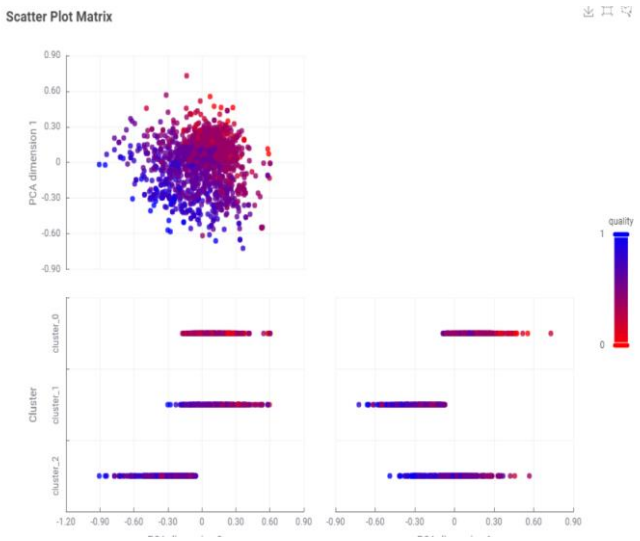
TA. B) e C) Nesta tarefa comecei por utilizar os nodes “Color Manager” e “Shape Manager” para atribuir diferentes cores por qualidade de vinho e diferentes formas a cada cluster do dataset:





Depois utilizando os nodes “Scatter plot” e “Scatter Plot(JFreeChart)” respetivamente obtive os seguintes gráficos:

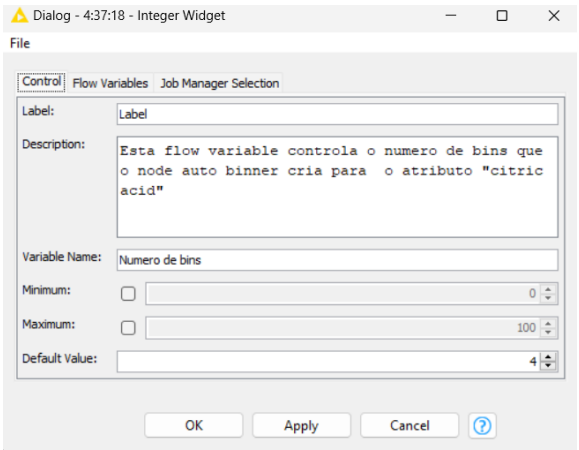




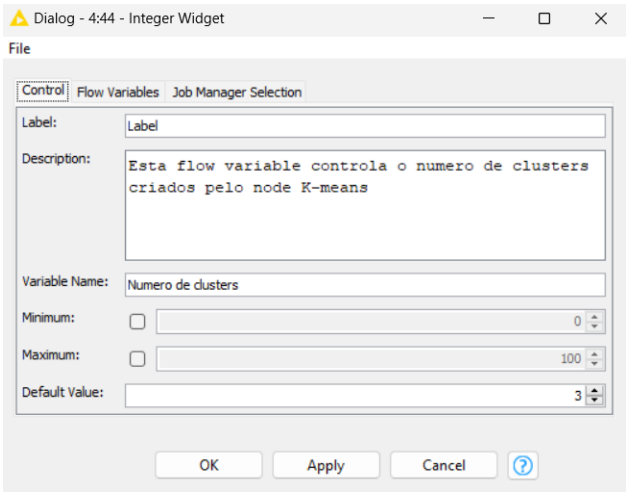
Onde consigo ver por exemplo que o cluster 0 tem elementos que em geral tem a feature da qualidade mais próxima de 0.

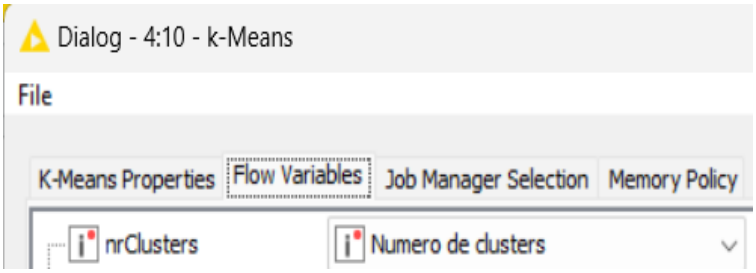
T4.D) e E) Para estas tarefas comecei por usar o node ”Cluster Assigner” o qual atribuiu a cada elemento dos dados de teste um cluster baseado nos dados de aprendizagem, depois com o node “CSV write” fiz um ficheiro que aos dados de teste acrescenta-se uma coluna com os clusters respetivos que o node anterior atribuiu, o ficheiro encontra-se anexado.

T5. Nesta tarefa começei por criar uma flow variable para definir o numero de bins criados com o “Auto-binner”, para isso usei o node “Integer Widget” e usei a seguinte configuração:

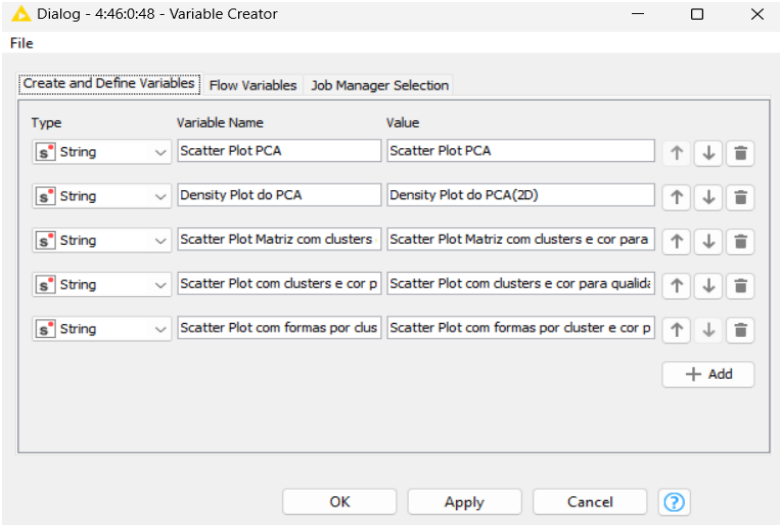


Depois criei uma flow variable para definir o numero de clusters criados com o “K-means”, para isso usei o node “Integer Widget” e usei a seguinte configuração:

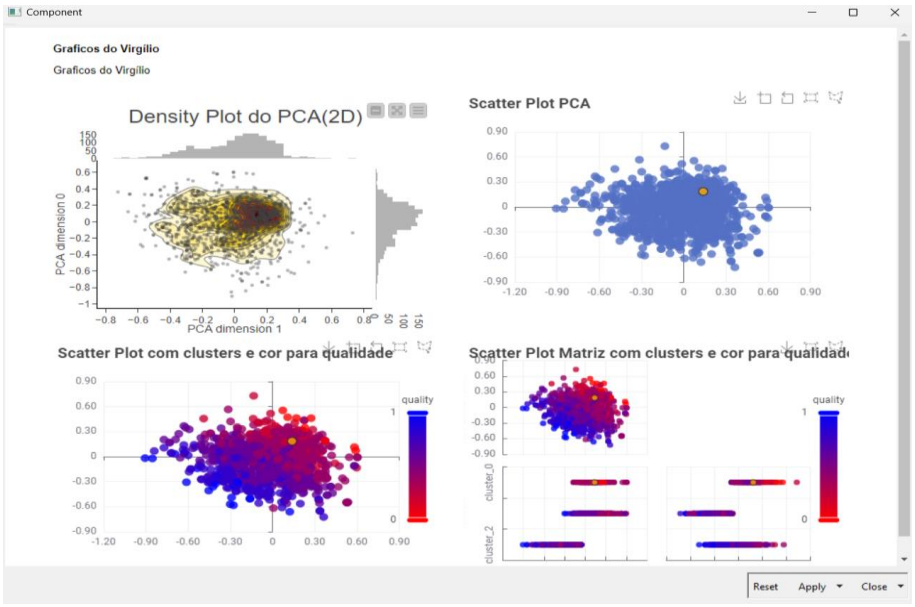




Depois criei um conjunto de flow variables para definir os nomes dos gráficos que fiz até aqui, para isso usei o node “Variable Creator”, liguei o node “Variable Creator” a cada um dos nodes que produziu os plots e usei a seguinte configuração:



T6. Nesta tarefa juntei todos os nodes que produziram graficos até aqui num só componente que os mostrara juntos:



T7. Nesta tarefa começei por testar resultados com o node “K-medoids”, para isso tive de começar por calcular as distancias euclidianas entre os diferentes pontos dispostas em 2 dimensoes, dei depois cores para o atributo de qualidade e separadamente para o atributo de cluster à semelhança do que tinha feito com o k-means e obtive as seguintes plots:

O que me fez concluir que a segmentação de dados é como esperado muito semelhante à realizada pelo node “k-means”.

De seguida experimentei tambem o node “Fuzzy-Means” onde separei os dados novamente por 3 clusters, tendo desta vez depois de dadas cores e visualizado agora obtido segmentções bem diferentes: