Universidade do Minho
Mestrado em Matemática e Computação

Sistemas Baseados em Similaridade - TP2

Simão Pedro Batista Caridade Quintela - PG52257

Outubro
2023

Universidade do Minho
Mestrado em Matemática e Computação

Relatório

Relatório realizado no âmbito do TP4 da UC Sistemas Baseados em Similiaridade do Mestrado em Matemática e Computação.

Outubro
2023

Conteúdo

1 Contextualização

Para a realização do TP4 foi-nos proposto a aplicação de métodos de **clustering** sobre um dataset de **vinhos**. Para isso foi-nos fornecido um dataset de treino e um de teste.

Para além de aplicar técnicas de clustering, este trabalho tem também como objetivo a aplicação de técnicas de exploração e tratamento de dados, bem como a parametrização do workflow desenvolvido.

2 Tarefas

2.1 Tarefa 1

Enunciado: Carregar, no Knime, os datasets descarregados e explorar os dados.

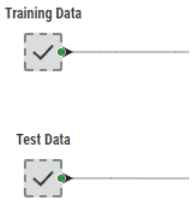

Figura 1: Workflows da tarefa 1

Inicialmente criei dois workflows para lidar com dados de treino e de teste de forma a que o workflow fosse mais legível.

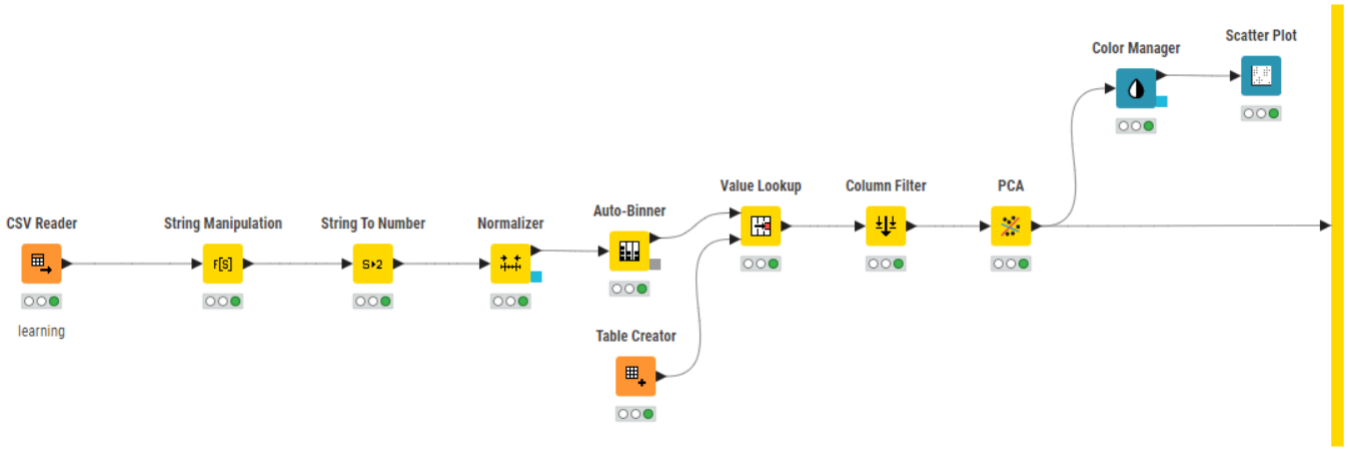

Figura 2: Workflow Training Data

2.2 Tarefa 2

a) Fazer cast do atributo “quality” para inteiro;

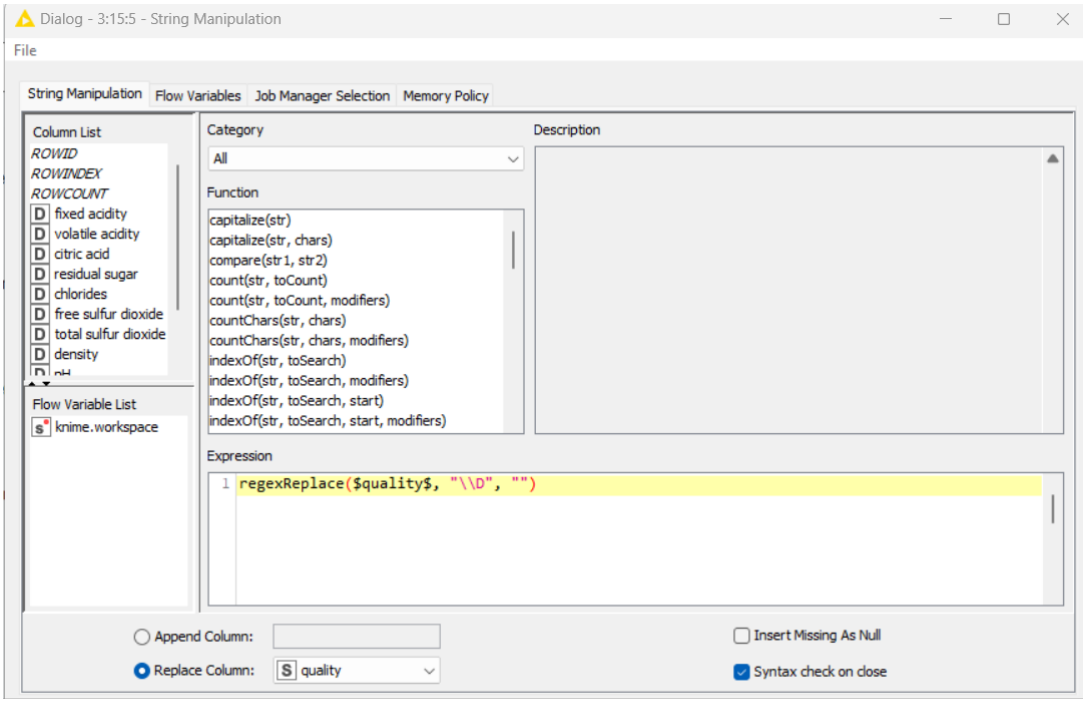

Figura 3: Utilização de regex para manipulação de strings

A coluna quality inicialmente era uma string com o seguinte formato: ”= x”, em que x é um número. Para isso utilizei o nodo **string manipulation** com regex, para selecionar apenas os dígitos da string. Captada a string, utilizei o nodo **string to number** para converter para inteiro.

b) Normalizar todos os atributos numéricos utilizando a transformação linear Min-max de forma a produzir um input normalizado entre 0 e 1

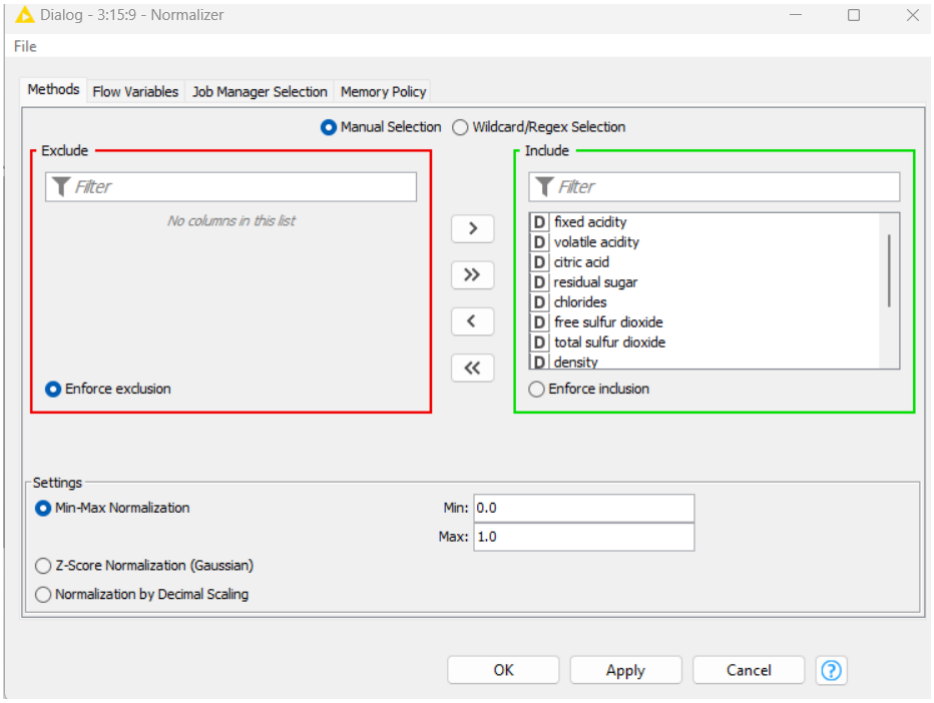

Figura 4: Configuração do nodo normalizer

Para realizar a normalização dos atributos numéricos utilizei o nodo **normalizer** com as configurações mostradas na imagem.

c) Criar 4 bins de igual frequência para a feature “citric acid”, substituindo a feature original;

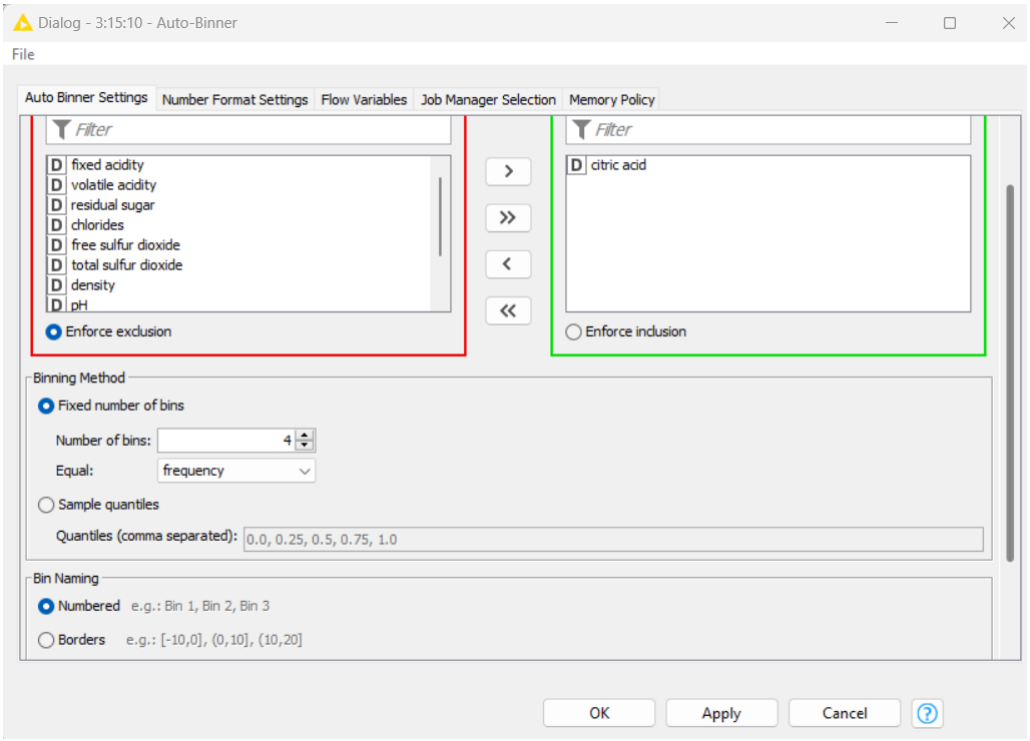

Figura 5: Configuração do auto-binner

Para criar 4 bins com as configurações pedidas utilizei o nodo **Auto-Binner** com as configurações mostradas na figura

d) Renomear cada bin de forma a que o primeiro corresponda a Low, o segundo a Medium, o terceiro a High e o quarto a Very High.

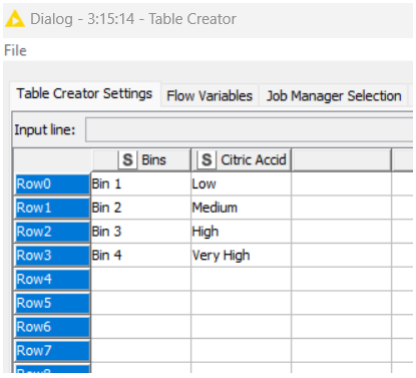

Figura 6: Configuração do nodo Table Creator

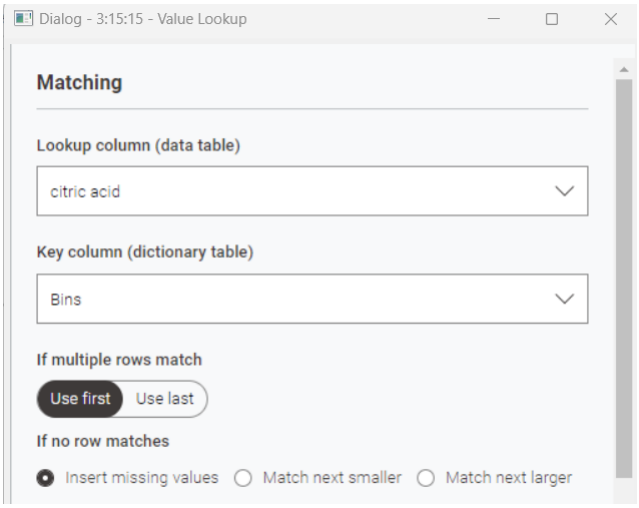

Figura 7: Configuração do nodo Value Lookup

Para realizar o proposto nesta alínea utilizei os nodos **Table Creator** e **Value Lookup**. Com o table creator criei uma tabela com 2 colunas e 4 linhas, nas quais as linhas da coluna Bins são os nomes dos Bins criados na alínea anterior, e os valores da coluna Citric Accid são os valores Low, Medium, High e Very High. Posto isto, utilizei o nodo Value Lookup para dar match às linhas com o nome do bin correspondente e quando o match acontece, o valor é transportado para o valor correspondente ao Bin na coluna Citric Accid.

2.3 Tarefa 3

a) Uma Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões.

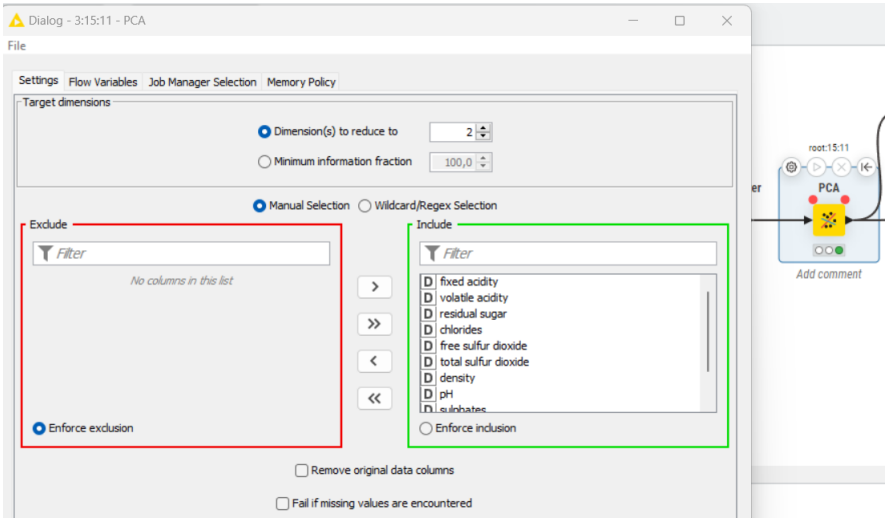

Figura 8: Configuração do PCA

Para projetar os dados em duas dimensões utilizei o algoritmo PCA(Principal Components Analysis). A configuração é a mostrada na figura 8.

b) Utilizar um scatter plot para visualização dos resultados obtidos pelo PCA.

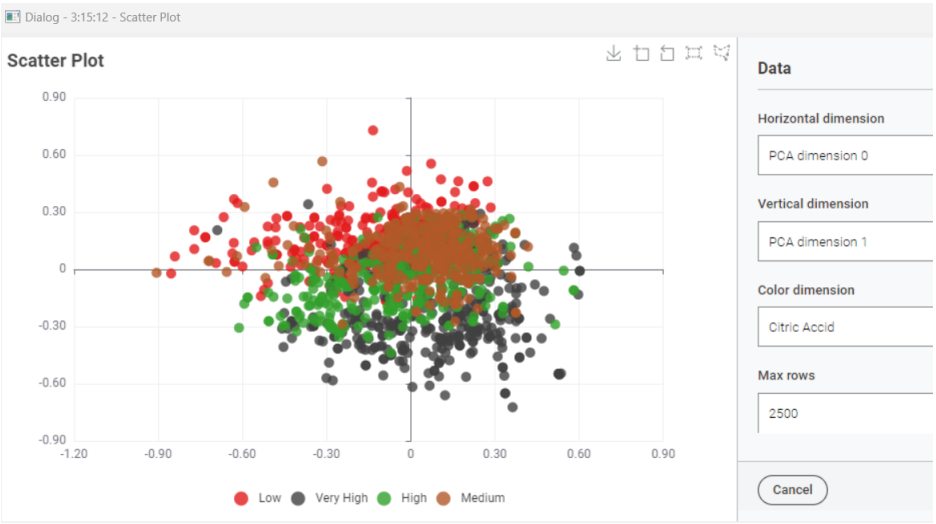

Figura 9: Configuração do PCA

Na figura 9 vemos o resultado da aplicação do PCA. Para isso utilizei um nodo **Color Manager** e, posteriormente, um nodo **Scatter Plot**.

2.4 Tarefa 4

a) Segmentar o dataset usando o algoritmo **K-Means**.

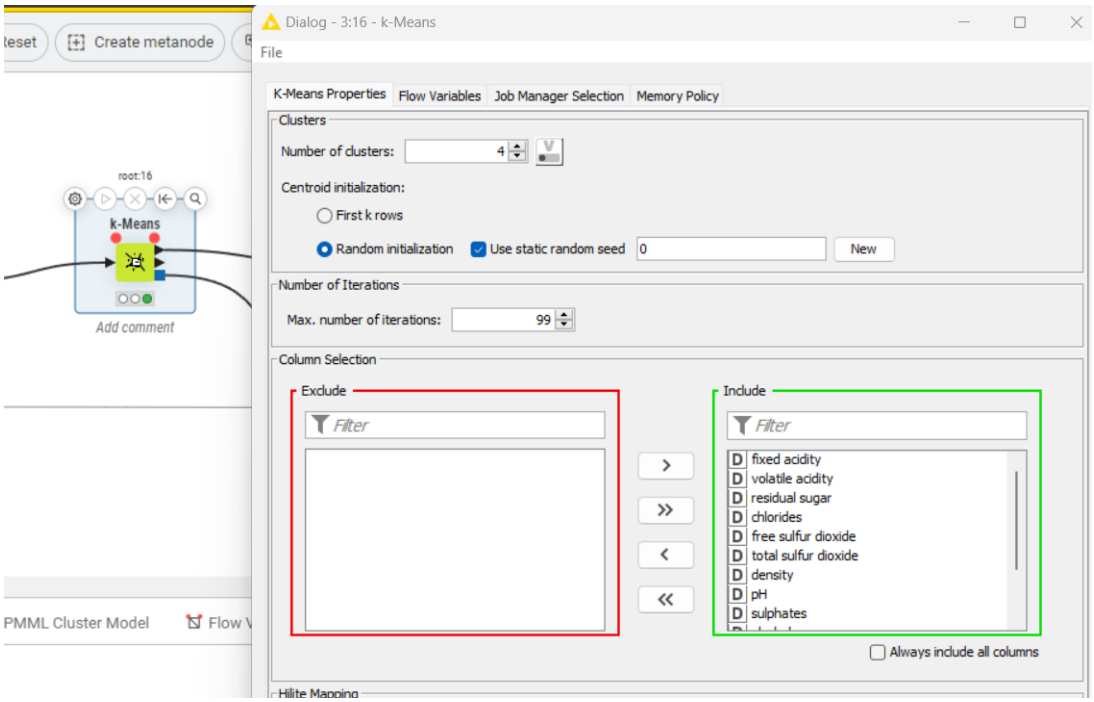

Figura 10: Configuração do nodo K-Means

Na figura 10 vemos a configuração do nodo K-Means, utilizado para segmentar o dataset. Para isso utilizei 99 iterações para criar 4 clusters.

b) Atribuir diferentes cores por qualidade do vinho e diferentes formas aos clusters.

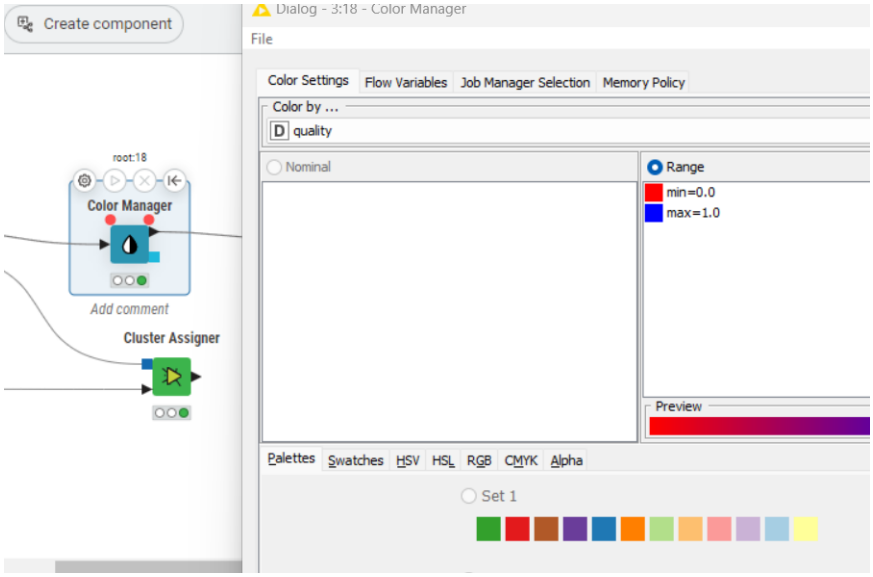

Figura 11: Configuração do nodo Color Manager

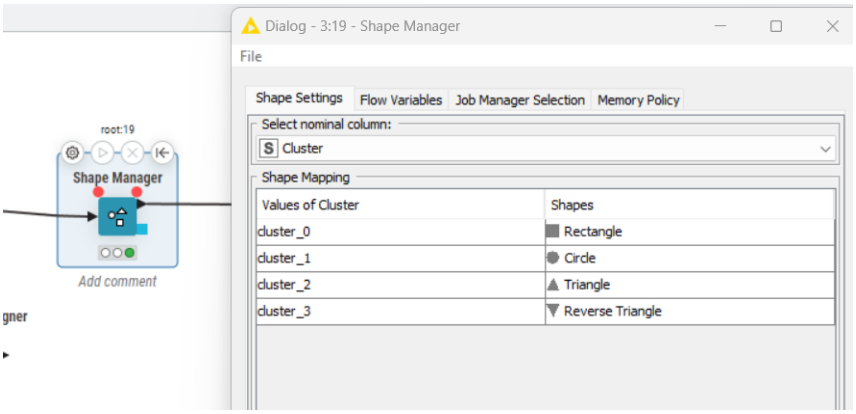

Figura 12: Configuração do nodo Shape Manager

Para atribuir cores por qualidade de vinho utilizei o nodo Color Manager e apliquei um gradiente de cores, conforme o valor da qualidade. Esta decisão baseia-se no facto dos valores de qualidade serem números entre 0 e 1. Na figura 12 vemos a configuração do nodo Shape Manager, utilizado para alterar a forma dos clusters em representações geométricas.

c) Criar scatter plots e scatter matrixes que permitam ter uma noção gráfica, em duas dimensões, dos atributos e dos clusters criados;

Figura 13: Nodo Scatter Plot

Figura 14: Nodo Scatter Plot Matrix

d) Ler e tratar os dados de teste de forma a que, com base no modelo desenvolvido nos passos anteriores, seja atribuído um cluster a cada registo deste ficheiro

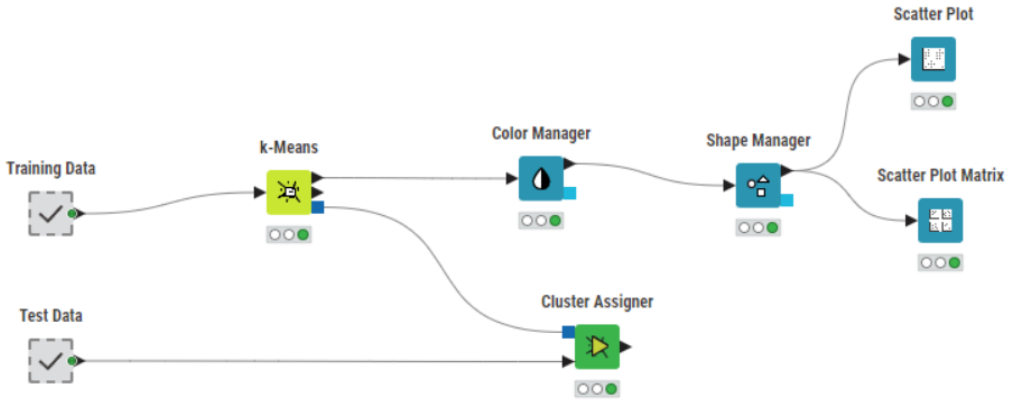

Figura 15: Workflow com dados de teste

e) Guardar o resultado da atribuição num ficheiro csv

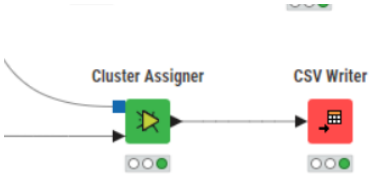

Figura 16: Workflow com dados de teste

Para atribuir a cada registo um cluster utilizei o nodo **Cluster Assigner** e, posteriormente, utilizei o nodo **CSV Writer** para escrever a respetiva atribuição num ficheiro do tipo csv.

2.5 Tarefa 4

Enunciado: Parametrizar o workflow, utilizando variáveis de fluxo para definir o número de bins, o número de clusters e os títulos dos gráficos criados

Flow Variables

Count: 4

Owner ID	Data Type	Variable Name	Value
3:22	StringType	Scatter Plot Cluster	Scatter Plot Cluster node
	StringType	knime.workspace	C:\University\MMCC\1ano\1sem
3:16	IntType	NumberClusters	4
3:15:10	IntType	NumberBins	4

Figura 17: Variáveis de fluxo criadas

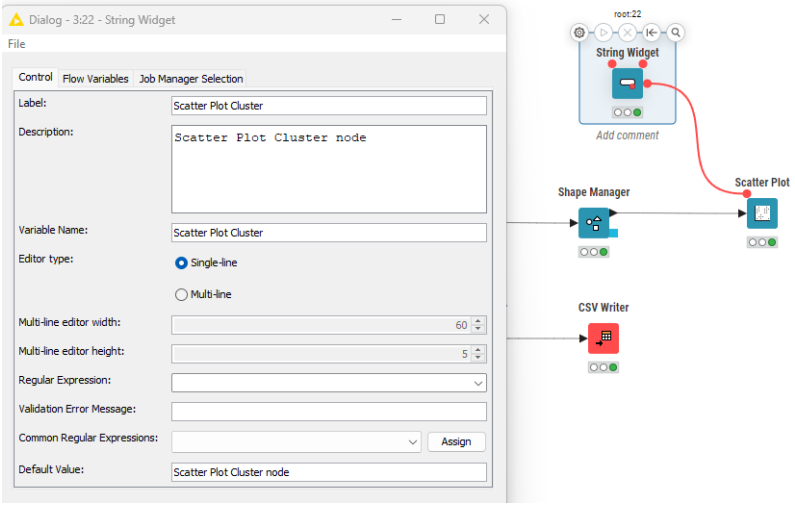

Figura 18: Criação da variável usando o String Widget

Na figura 17 vemos as variáveis que criei para a realização desta tarefa. Inicialmente, criei a variável **Scatter Plot Cluster** (presente na figura 18) que serve para dar título a nodos relacionados com a visualização de clusters. Para isso utilizei o nodo **String Widget**.

Prosseguindo, criei duas variáveis **NumberClusters** e **NumberBins** às quais atribui o valor 4, que corresponde ao número de clusters e de bins criados no trabalho. Com as variáveis criadas posso agora atribuir estas variáveis a outros nodos dando dinamismo ao workflow.

2.6 Tarefa 6

Enunciado: Produzir o workflow de maneira a que seja possível visualizar, numa única página, todos os componentes visuais implementados

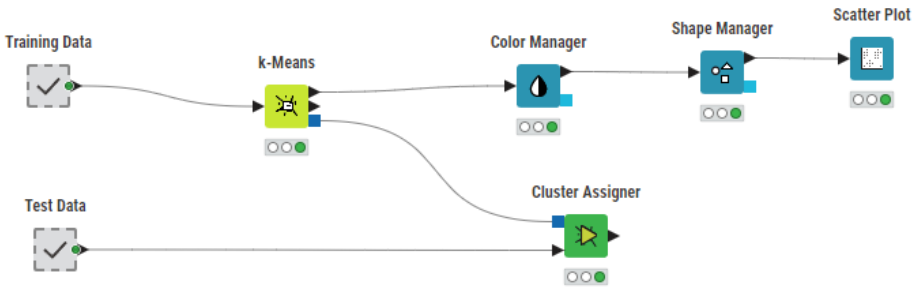

Figura 19: Worflow completo

Para organizar o workflow de forma a que fosse possível visualizar, numa única página, os componentes visuais implementados utilizei metanodos. A utilização dos mesmos faz com que o workflow seja mais facilmente manipulável e mais legível.

2.7 Tarefa 7

Enunciado: Experimentar, avaliar e comparar outros métodos de segmentação.

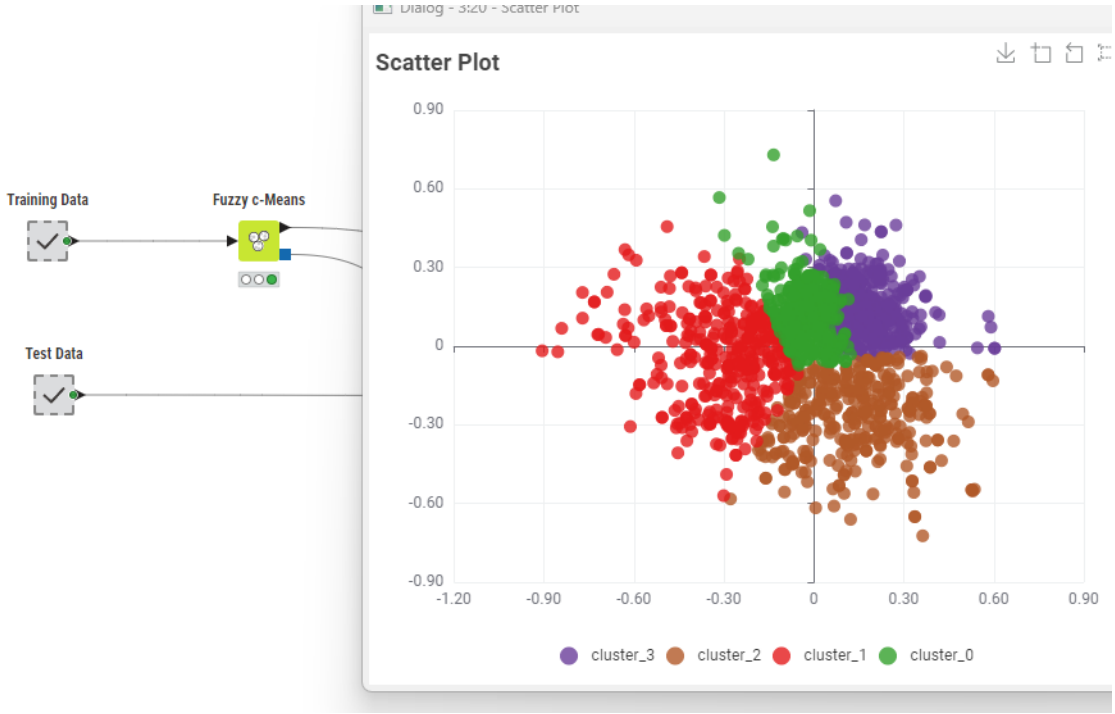

Figura 20: Output usando Fuzzy c-Means

Utilizando o algoritmo Fuzzy c-Means que deixa um ponto pertencer parcialmente a vários clusters, verifica-se uma alteração da distribuição dos pontos na zona central e no 1º quadrante.

Figura 21: Output usando Hirearchical Clustering

Utilizando o algoritmo Hirearchical Clustering o output apresentado foi o mostrado na figura 19. Notamos claramente uma diferença em relação ao algoritmo Fuzzy c-Means, o centro do referencial não tem nenhum cluster associado. Em vez disso, os clusters encontram-se divididos pelos restantes quadrantes.

3 Conclusão

Para concluir, esta tarefa foi importante porque foram apresentados vários conceitos e algoritmos acerca de clustering, um conceito importantíssimo na área. Para além disso, continuamos a explorar o Knime e a descobrir novos nodos que não tinham sido utilizados antes. Estudamos também o benefício de usar Flow Variables e como isso nos pode ajudar a criar um workflow dinâmico.