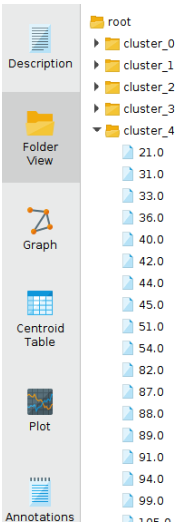
FICHA 8

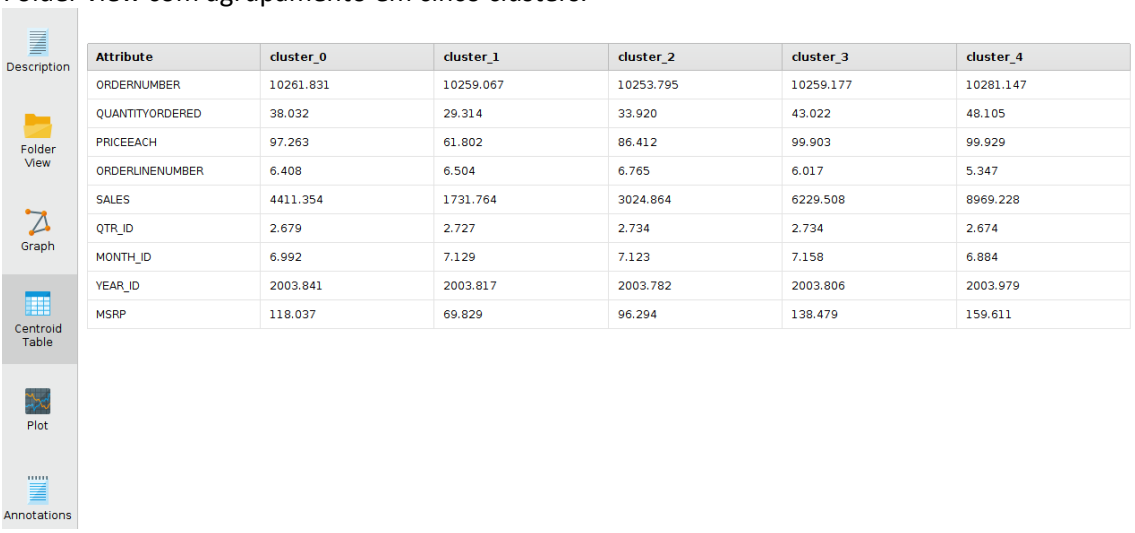
DECISION TREES

1. O que significa k-Means clustering?
   R: Em mineração de dados, k-Means é um método de clustering que tem como objetivo particionar n observações em k grupos onde cada observação pertence ao grupo mais próximo da média.
2. Como se identificam os clusters e qual o processo que o rapidMiner usa para definir e colocar as observações num determinado cluster?
   R: O RapidMiner utiliza a média de todos os pontos para estabelecer um ponto de partida inicial. Posteriormente, em consecutivas medições, tenta diminuir a distância desse ponto aos restantes, criando novas retas/planos que dividem os dados em clusters.
3. O que revela a Centroid Table ao utilizador? Como se interpretam os valores nessa tabela?
   R: Apresenta todos os pontos finais do processo de clustering, de acordo com os parâmetros disponíveis para análise. Revela, assim, a média em cada um dos clusters para os distintos dados.
4. Depois do exercício introdutório pensar num problema que possa ser resolvido agrupando observações em clusters. Procurar na internet um dataset que possa ser utilizado e aplicado um modelo de k-Means.
   a. Garantir que os dados estão no formato CSV e importar os mesmos para o RapidMiner;
      R: O dataset para este exercício foi obtido em https://www.kaggle.com/kyanyoga/sample-sales-data/version/1?fbclid=IwAR2ozm9e4wpIxl61kfSjrzUpxf18LV8zJuR6LvOVqNOJfDcyYRxOSixC-nk
      É referente a um conjunto de informações acerca de vendas de produtos e processos associados.
   b. Fase de preparação dos dados. Pode incluir componentes de inconsistência de dados, missing values, ou alteração do tipo de dados;
      R: Retiraram-se algumas colunas, relacionadas com o nome das cidades, estados e ruas. Nestas havia alguns *missing values*, que não se verificam nas restantes.
   c. Ligar um operador de k-means clustering ao dataset e alterar os parâmetros de acordo com a necessidade (sobretudo o k, para adequar ao problema em questão);
      R: O algoritmo foi aplicado utilizando um agrupamento em cinco e doze clusters, para tentar ajustar aos meses do ano.
   d. Avaliar a Centroid Table, Folder View, e outras ferramentas de avaliação;

R:

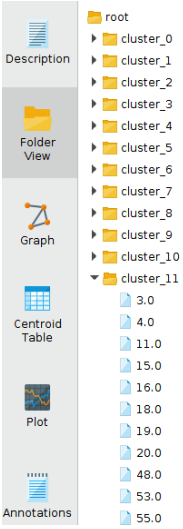


Folder view com agrupamento em cinco clusters.

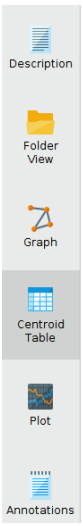


Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
ORDERNUMBER	10261.831	10259.067	10253.795	10259.177	10281.147
QUANTITYORDERED	38.032	29.314	33.920	43.022	48.105
PRICEEACH	97.263	61.802	86.412	99.903	99.929
ORDERLINENUMBER	6.408	6.504	6.765	6.017	5.347
SALES	4411.354	1731.764	3024.864	6229.508	8969.228
QTR_ID	2.679	2.727	2.734	2.734	2.674
MONTH_ID	6.992	7.129	7.123	7.158	6.884
YEAR_ID	2003.841	2003.817	2003.782	2003.806	2003.979
MSRP	118.037	69.829	96.294	138.479	159.611

Centroid Table com agrupamento em cinco clusters.



Folder view com agrupamento em doze clusters.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7	cluster_8	cluster_9	cluster_...	cluster...
ORDERNU...	10255.990	10323.333	10262.201	10292.355	10251.073	10265.936	10258.316	10259.250	10254.878	10407	10259.954	10263.496
QUANTITY...	41.914	54.250	26.771	49.677	34.927	44.150	30.730	39.765	32.903	76	45.585	36.747
PRICEEACH	99.951	100	50.469	99.783	89.897	99.904	68.501	99.485	82.224	100	99.827	95.566
ORDERLIN...	6.220	5.333	6.439	5.806	6.537	5.786	6.544	6.642	6.928	2	5.231	6.300
SALES	5746.976	11426.483	1281.426	9230.015	3342.071	6808.780	1990.664	4871.936	2669.310	14082.800	7998.238	4057.202
QTR_ID	2.751	2.750	2.721	2.516	2.732	2.671	2.720	2.665	2.737	2	2.831	2.703
MONTH_ID	7.196	7.083	7.157	6.323	7.113	7.014	7.091	6.865	7.126	4	7.385	7.123
YEAR_ID	2003.780	2004.250	2003.840	2004.097	2003.758	2003.871	2003.811	2003.842	2003.796	2005	2003.785	2003.837
MSRP	134.096	167.750	61.066	149.968	100.869	143.071	75.144	120.858	90.794	170	162.323	115.294

Centroid Table com agrupamento em cinco clusters.

e. Reportar todos os passos anteriores e as evidências encontradas. Discutir as iterações no modelo, e de que forma o que foi encontrado permite responder ao problema inicial.

R: Pode-se verificar que existe uma separação não homogénea quando k=12, verificando-se clusters com, por exemplo, preços unitários médios inferiores ao normal de 100. É possível também ver que há estas organizações que resultam em médias de vendas muito distintas, sendo que onde se verifica mais vendas é no mês 7.

5. Experimentar o mesmo dataset com diferentes operadores de k-Means como o Kernel ou Fast. Em que medida diferem do modelo original. Estes operadores mudam os clusters originais? Se sim, em que medida?

R:

K-Means Normal

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
ORDERNUMBER	10261.831	10259.067	10253.795	10259.177	10281.147
QUANTITYORDERED	38.032	29.314	33.920	43.022	48.105
PRICEEACH	97.263	61.802	86.412	99.903	99.929
ORDERLINENUMBER	6.408	6.504	6.765	6.017	5.347
SALES	4411.354	1731.764	3024.864	6229.508	8969.228
QTR_ID	2.679	2.727	2.734	2.734	2.674
MONTH_ID	6.992	7.129	7.123	7.158	6.884
YEAR_ID	2003.841	2003.817	2003.782	2003.806	2003.979
MSRP	118.037	69.829	96.294	138.479	159.611

K-Means Kernel - Não são apresentados centróides, uma vez que os centros não são pontos mas sim “zonas”.

K-Means Fast

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
ORDERNUMBER	10259.067	10281.147	10261.831	10259.177	10253.795
QUANTITYORDERED	29.314	48.105	38.032	43.022	33.920
PRICEEACH	61.802	99.929	97.263	99.903	86.412
ORDERLINENUMBER	6.504	5.347	6.408	6.017	6.765
SALES	1731.764	8969.228	4411.354	6229.508	3024.864
QTR_ID	2.727	2.674	2.679	2.734	2.734
MONTH_ID	7.129	6.884	6.992	7.158	7.123
YEAR_ID	2003.817	2003.979	2003.841	2003.806	2003.782
MSRP	69.829	159.611	118.037	138.479	96.294

Entre o normal e o fast a diferença é diminuta. No entanto, o kernel apresenta resultados muito diferentes quanto ao clustering, fazendo-o de forma muito mais homogénea do que os anteriores, com cerca de 500 instâncias por cluster.