Enunciado Prático nº 5

Maria José Borges Pires - A86268

18 de novembro de 2020

1 Exercício 1

Apresenta-se de seguida o carrefamento dos dois primeiros datasets, a sua junção e exploração através de vistas gráficas. Os nodos utilizados para esta exploração encontram-se no metanodo *Exploração Visual*.

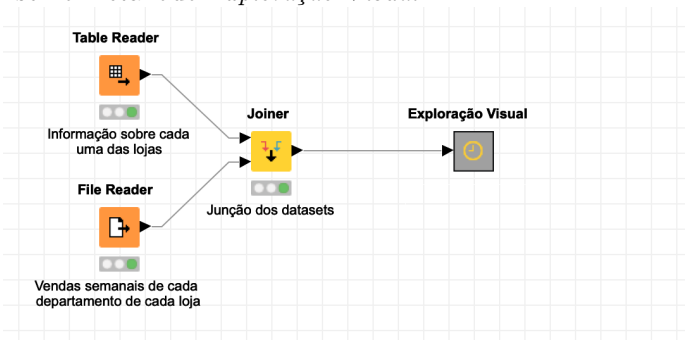

Figure 1: Workflow do exercício 1

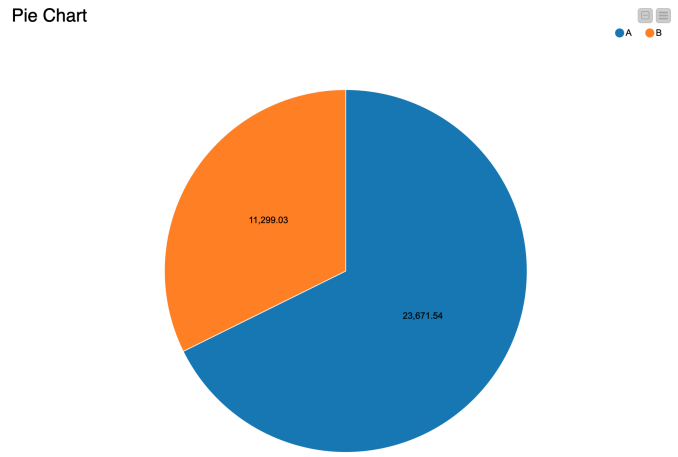

Figure 2: Pie Chart que apresenta as vendas semanais por tipo

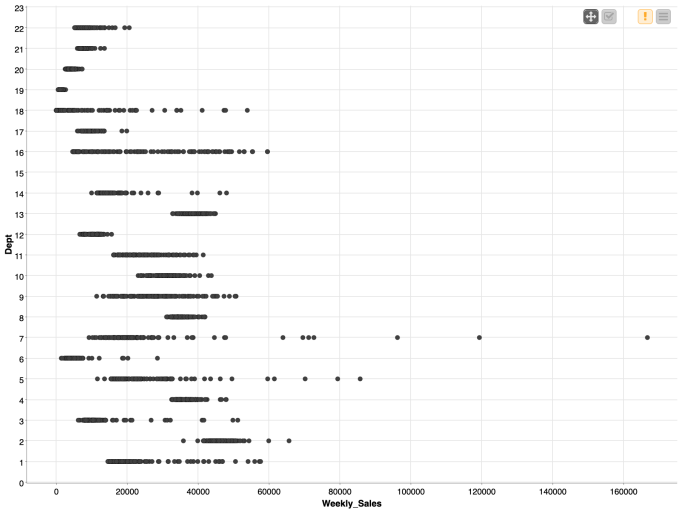

Figure 3: Scatter plot que apresenta as vendas semanais por departamento

2 Exercício 2

2.1 Fazer label encoding à feature isHoliday (1 deve corresponder ao valor True)

Através do nodo *Rule Encoding* traduziram-se os registos da coluna *isHoliday* de *True* *False* para 0 e. 1.

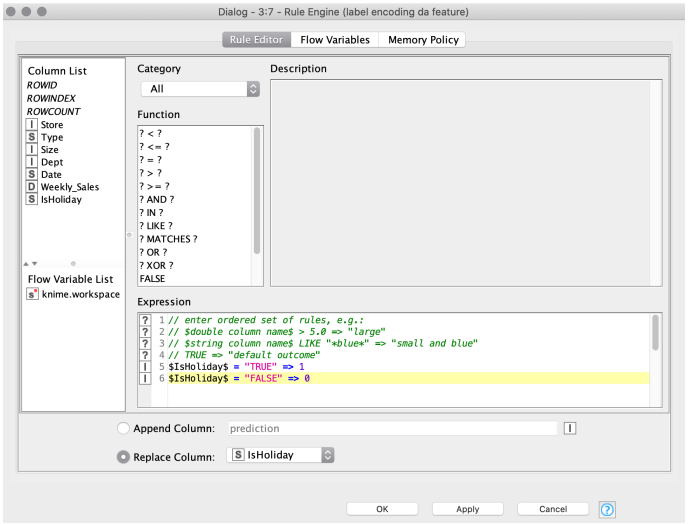

Figure 4: Definições aplicadas ao nodo *Rule Encoding*

Row ID	Store	Type	Size	Dept	Date	Weekl...	IsHoli...
Row0_Row0	1	A	151315	1	05/02/2010	24,924.5	0
Row0_Row1	1	A	151315	1	12/02/2010	46,039.49	1
Row0_Row2	1	A	151315	1	19/02/2010	41,595.55	0
Row0_Row3	1	A	151315	1	26/02/2010	19,403.54	0
Row0_Row4	1	A	151315	1	05/03/2010	21,827.9	0
Row0_Row5	1	A	151315	1	12/03/2010	21,043.39	0
Row0_Row6	1	A	151315	1	19/03/2010	22,136.64	0
Row0_Row7	1	A	151315	1	26/03/2010	26,229.21	0
Row0_Row8	1	A	151315	1	02/04/2010	57,258.43	0
Row0_Row9	1	A	151315	1	09/04/2010	42,960.91	0
Row0_Row10	1	A	151315	1	16/04/2010	17,596.96	0
Row0_Row11	1	A	151315	1	23/04/2010	16,145.35	0
Row0_Row12	1	A	151315	1	30/04/2010	16,555.11	0
Row0_Row13	1	A	151315	1	07/05/2010	17,413.94	0
Row0_Row14	1	A	151315	1	14/05/2010	18,926.74	0
Row0_Row15	1	A	151315	1	21/05/2010	14,773.04	0
Row0_Row16	1	A	151315	1	28/05/2010	15,580.43	0
Row0_Row17	1	A	151315	1	04/06/2010	17,558.09	0
Row0_Row18	1	A	151315	1	11/06/2010	16,637.62	0
Row0_Row19	1	A	151315	1	18/06/2010	16,216.27	0
Row0_Row20	1	A	151315	1	25/06/2010	16,328.72	0
Row0_Row21	1	A	151315	1	02/07/2010	16,333.14	0
Row0_Row22	1	A	151315	1	09/07/2010	17,688.76	0
Row0_Row23	1	A	151315	1	16/07/2010	17,150.84	0
Row0_Row24	1	A	151315	1	23/07/2010	15,360.45	0
Row0_Row25	1	A	151315	1	30/07/2010	15,381.82	0
Row0_Row26	1	A	151315	1	06/08/2010	17,508.41	0
Row0_Row27	1	A	151315	1	13/08/2010	15,536.4	0

Figure 5: Excerto da tabela obtida após o label encoding

2.2 Adicionar, a cada registo, as features ano e mês

Através do nodo *Cell Spliter* é possivel extrair o dia, mês e o ano de uma data, de seguida, através do nodo *Column Filter* é feita a filtragem das colunas, descartando o dia. Finalmente, e para que futuramente não haja conflitos. é feito o rename das colunas.

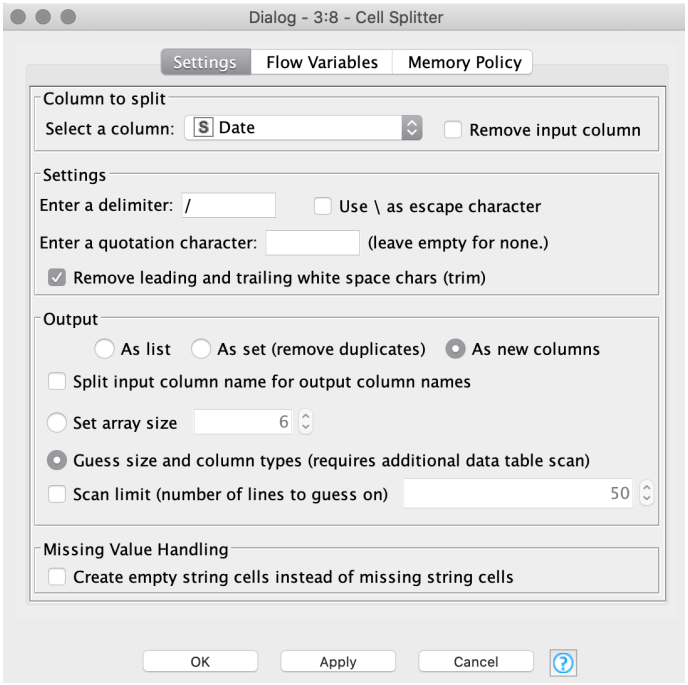

Figure 6: Definições aplicadas ao nodo *Cell Spliter*

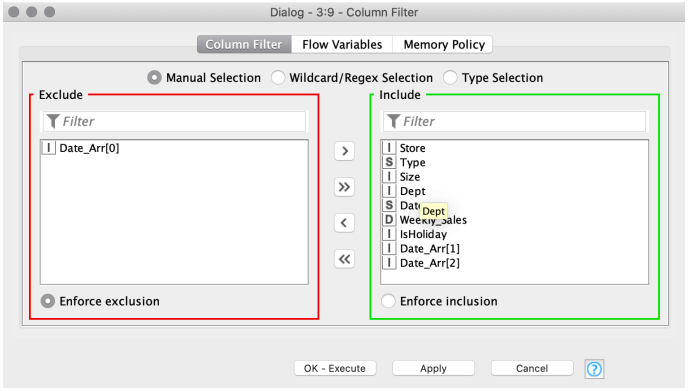

Figure 7

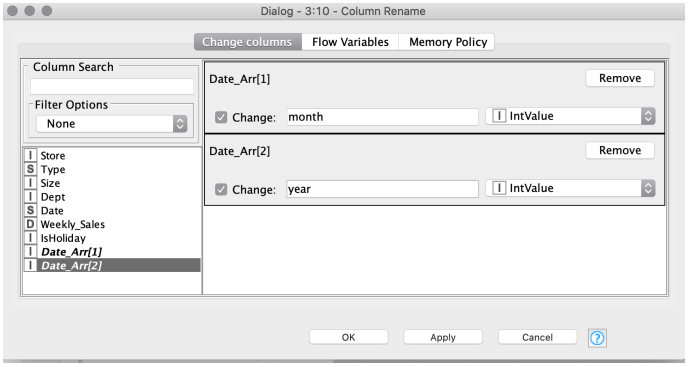

Figure 8: Filtragem das colunas

Row ID	Store	Type	Size	Dept	Date	Weekl...	IsHoli...	month	year
Row0_Row0	1	A	151315	1	05/02/2010	24,924.5	0	2	2010
Row0_Row1	1	A	151315	1	12/02/2010	46,039.49	1	2	2010
Row0_Row2	1	A	151315	1	19/02/2010	41,595.55	0	2	2010
Row0_Row3	1	A	151315	1	26/02/2010	19,403.54	0	2	2010
Row0_Row4	1	A	151315	1	05/03/2010	21,827.9	0	3	2010
Row0_Row5	1	A	151315	1	12/03/2010	21,043.39	0	3	2010
Row0_Row6	1	A	151315	1	19/03/2010	22,136.64	0	3	2010
Row0_Row7	1	A	151315	1	26/03/2010	26,229.21	0	3	2010
Row0_Row8	1	A	151315	1	02/04/2010	57,258.43	0	4	2010
Row0_Row9	1	A	151315	1	09/04/2010	42,960.91	0	4	2010
Row0_Row10	1	A	151315	1	16/04/2010	17,596.96	0	4	2010
Row0_Row11	1	A	151315	1	23/04/2010	16,145.35	0	4	2010
Row0_Row12	1	A	151315	1	30/04/2010	16,555.11	0	4	2010
Row0_Row13	1	A	151315	1	07/05/2010	17,413.94	0	5	2010
Row0_Row14	1	A	151315	1	14/05/2010	18,926.74	0	5	2010
Row0_Row15	1	A	151315	1	21/05/2010	14,773.04	0	5	2010
Row0_Row16	1	A	151315	1	28/05/2010	15,580.43	0	5	2010
Row0_Row17	1	A	151315	1	04/06/2010	17,558.09	0	6	2010
Row0_Row18	1	A	151315	1	11/06/2010	16,637.62	0	6	2010
Row0_Row19	1	A	151315	1	18/06/2010	16,216.27	0	6	2010
Row0_Row20	1	A	151315	1	25/06/2010	16,328.72	0	6	2010
Row0_Row21	1	A	151315	1	02/07/2010	16,333.14	0	7	2010
Row0_Row22	1	A	151315	1	09/07/2010	17,688.76	0	7	2010

Figure 9: Excerto da tabela obtida no final do exercicio

2.3 Agrupar os registos por loja, tipo, tamanho, ano e mês, agregando de forma a obter o somatório das vendas semanais de cada loja e a indicação da existência de feriados nesse mês

Através do nodo *GroupBy* é possivel fazer-se a agregação dos registos.

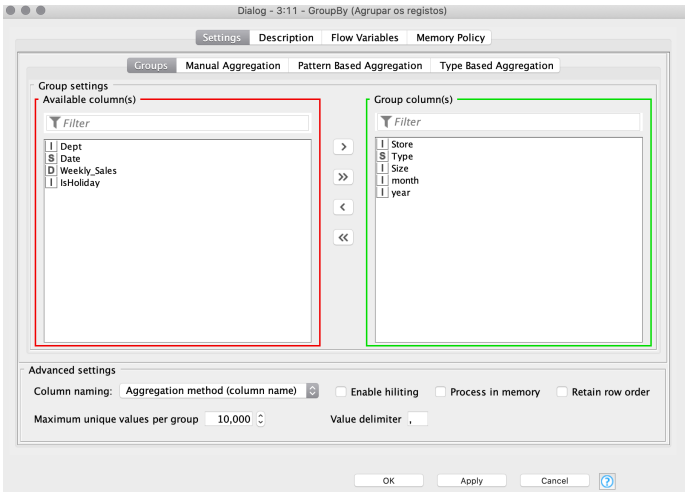

Figure 10: Definições de agrupamento

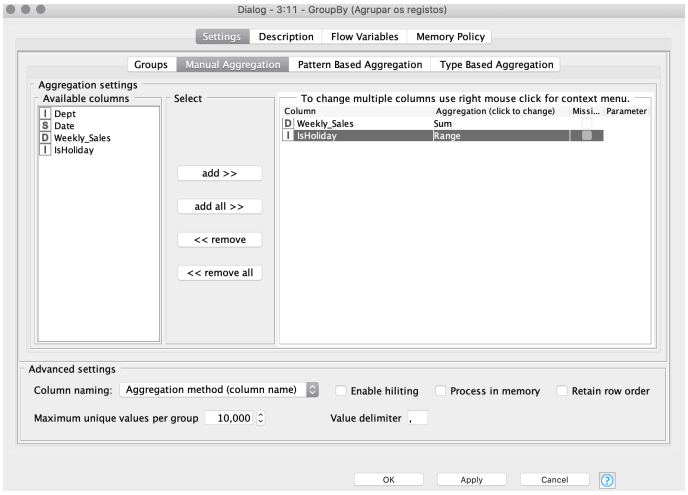

Figure 11: Definições de agregação manual

Row ID	I Store	S Type	I Size	I month	I year	D Sum(We...	D Range(IsHoliday)
Row0	1	A	151315	1	2011	5,480,050....	0
Row1	1	A	151315	1	2012	5,723,690....	0
Row2	1	A	151315	2	2010	6,307,344.1	1
Row3	1	A	151315	2	2011	6,399,887....	1
Row4	1	A	151315	2	2012	6,798,074....	1
Row5	1	A	151315	3	2010	5,871,293....	0
Row6	1	A	151315	3	2011	6,307,375....	0
Row7	1	A	151315	3	2012	8,201,997.4	0
Row8	1	A	151315	4	2010	7,422,801....	0
Row9	1	A	151315	4	2011	7,689,123.6	0
Row10	1	A	151315	4	2012	6,511,214....	0
Row11	1	A	151315	5	2010	5,929,938....	0
Row12	1	A	151315	5	2011	6,128,431.8	0
Row13	1	A	151315	5	2012	6,446,962....	0
Row14	1	A	151315	6	2010	6,084,081....	0
Row15	1	A	151315	6	2011	6,194,971....	0
Row16	1	A	151315	7	2010	7,244,483....	0
Row17	1	A	151315	7	2011	7,227,654....	0
Row18	1	A	151315	8	2010	6,075,952....	0
Row19	1	A	151315	8	2011	6,144,985....	0
Row20	1	A	151315	9	2010	5,829,793....	1
Row21	1	A	151315	9	2011	7,379,542....	1

Figure 12: Excerto da tabela obtido após o agrupamento

2.4 Normalizar o somatório das vendas semanais utilizando a transformação linear Min-Max entre 0 e 1

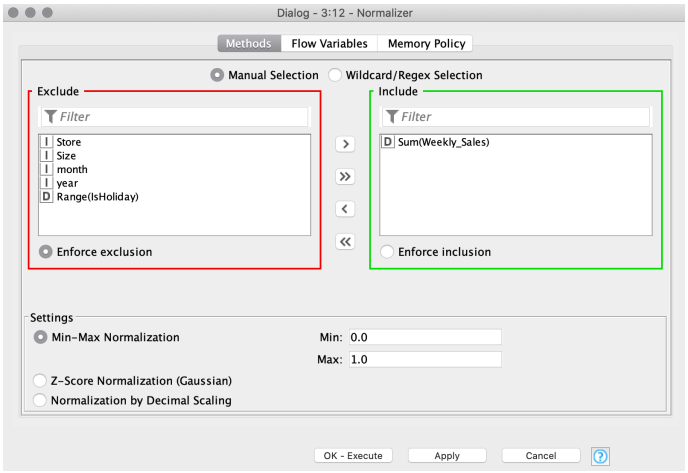

Figure 13: Definições aplicadas para a normalização do somatório das vendas

2.5 Criar 4 bins de igual frequência sobre o valor normalizado no passo anterior (ligando a opção replace target column(s))

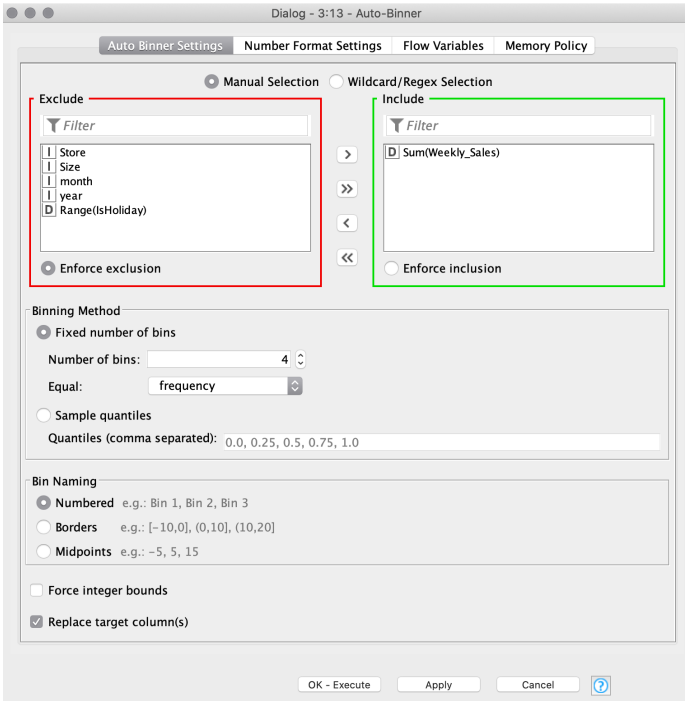

Figure 14: Definições aplicadas ao nodo *Auto-Binner*

Row ID	Store	Type	Size	month	year	Sum(Weekly_Sales)	Range...
Row0	1	A	151315	1	2011	Bin 3	0
Row1	1	A	151315	1	2012	Bin 3	0
Row2	1	A	151315	2	2010	Bin 3	1
Row3	1	A	151315	2	2011	Bin 3	1
Row4	1	A	151315	2	2012	Bin 3	1
Row5	1	A	151315	3	2010	Bin 3	0
Row6	1	A	151315	3	2011	Bin 3	0
Row7	1	A	151315	3	2012	Bin 4	0
Row8	1	A	151315	4	2010	Bin 3	0
Row9	1	A	151315	4	2011	Bin 4	0
Row10	1	A	151315	4	2012	Bin 3	0
Row11	1	A	151315	5	2010	Bin 3	0
Row12	1	A	151315	5	2011	Bin 3	0
Row13	1	A	151315	5	2012	Bin 3	0
Row14	1	A	151315	6	2010	Bin 3	0
Row15	1	A	151315	6	2011	Bin 3	0
Row16	1	A	151315	7	2010	Bin 3	0
Row17	1	A	151315	7	2011	Bin 3	0
Row18	1	A	151315	8	2010	Bin 3	0
Row19	1	A	151315	8	2011	Bin 3	0
Row20	1	A	151315	9	2010	Bin 3	1
Row21	1	A	151315	9	2011	Bin 3	1
Row22	1	A	151315	10	2010	Bin 3	0
Row23	1	A	151315	10	2011	Bin 3	0
Row24	1	A	151315	11	2010	Bin 3	1
Row25	1	A	151315	11	2011	Bin 3	1
Row26	1	A	151315	12	2010	Bin 4	1
Row27	1	A	151315	12	2011	Bin 4	1

Figure 15: Tabela obtida após a aplicação do nodo *Auto-Binner*

2.6 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

Figure 16: Renomeação de bins

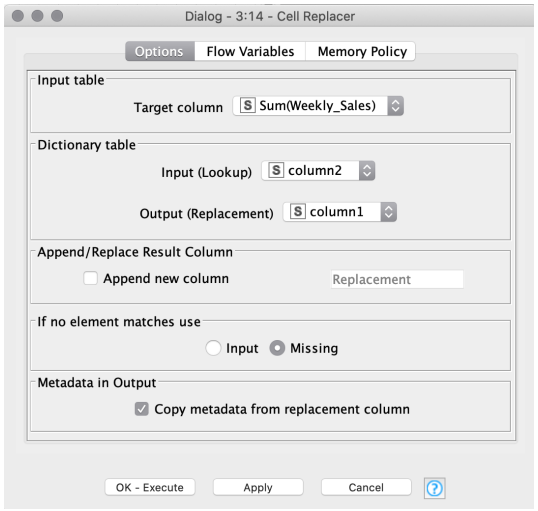

Figure 17: Renomeação de bins

3 Exercício 3

3.1 Treinar uma árvore de decisão

Através do nodo *Partitioning* é possivel dividir o *dataset* e aplicar os nodos *Decision Tree Learner* e *Decision Tree Predictor* para treinar uma àrvore de decisão com os dados tratados até ao momento.

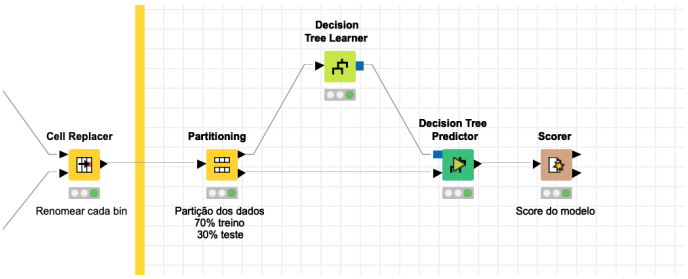

Figure 18: Fluxo de treino da árvore de decisão

3.2 Carregar o dataset de teste e prever o valor de vendas de cada mês para cada uma das 17 lojas

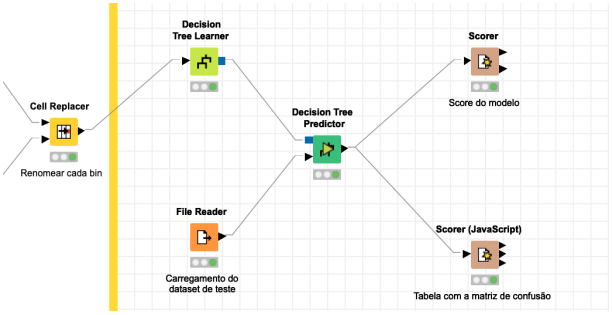

Figure 19: Workflow de carregamento dos dados e previsão do valor de vendas

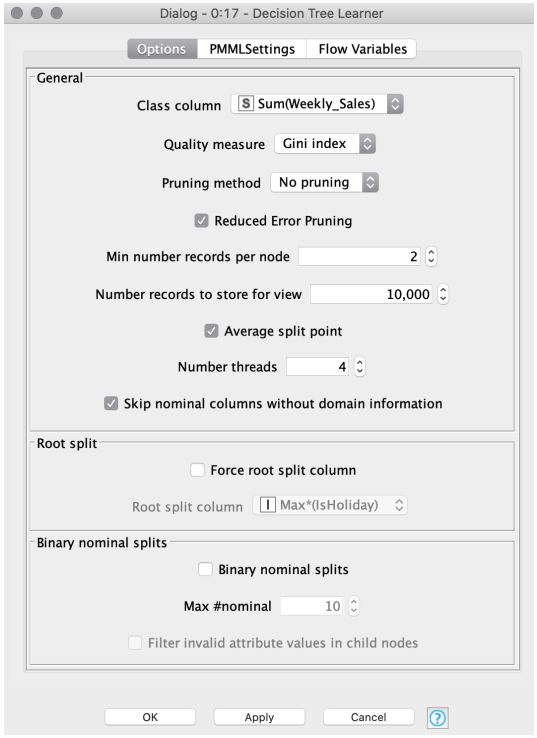

Figure 20: Definições aplicadas ao nodo *Decision Tree Learner*

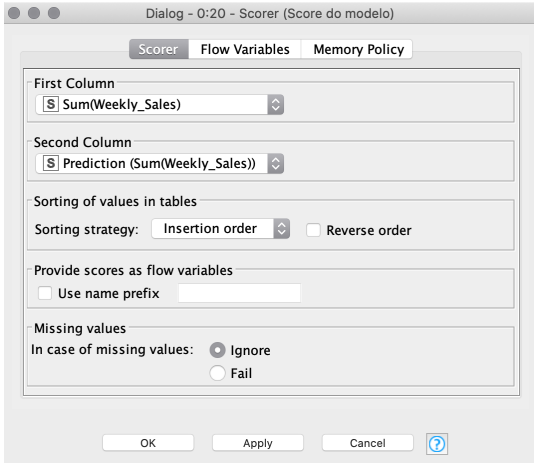

Figure 21: Definições aplicadas ao nodo *Scorer*

A matriz de confusão do modelo apresenta uma previsão de 60%, como de

pode observar na figura seguinte:

Confusion Matrix - 0:20 - Scorer (Score do modelo)

File Hilite

Sum(Weekl...	Very High	High	Low	Medium
Very High	14	6	0	0
High	9	11	0	4
Low	0	0	12	8
Medium	0	1	6	14

Correct classified: 51 Wrong classified: 34

Accuracy: 60 % Error: 40 %

Cohen's kappa (κ) 0.467

Figure 22: Matriz de confusão do modelo

3.3 Mostrar, graficamente, uma tabela com a matriz de confusão do modelo

Para obter a tabela da figura seguinte foi aplicado o nodo *Scorer(JavaScript)*.

Scorer View

Confusion Matrix

	High (Predicted)	Low (Predicted)	Medium (Predicted)	Very High (Predic...	
High (Actual)	11	0	4	9	45.83%
Low (Actual)	0	12	8	0	60.00%
Medium (Actual)	1	6	14	0	66.67%
Very High (Actual)	6	0	0	14	70.00%
	61.11%	66.67%	53.85%	60.87%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
60.00%	40.00%	0.467	51	34

Figure 23: Tabela com a matriz de confusão do modelo

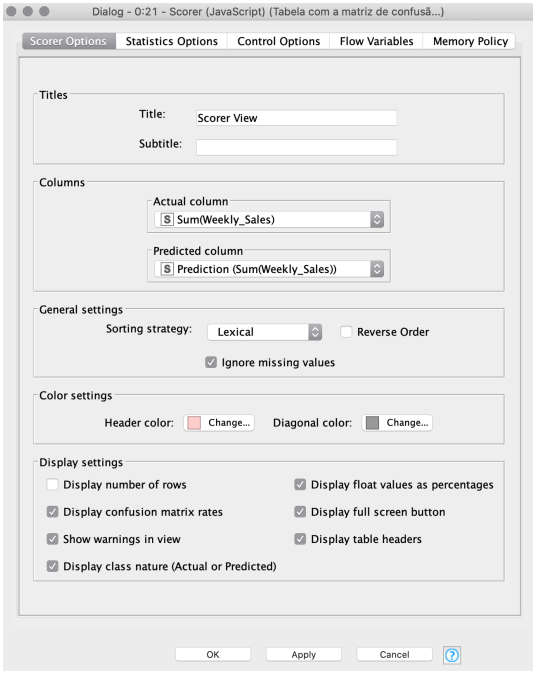

Figure 24: Definições aplicadas ao nodo *Scorer(JavaScript)*

4 Exercício 4

4.1 Fazer o tuning do modelo criado no passo anterior com todos os valores, entre 2 e 10, para o número mínimo de registos por nodo

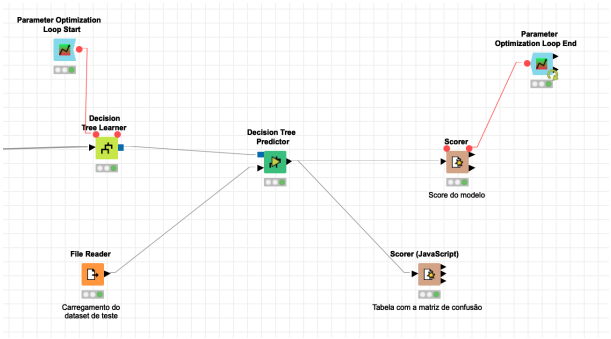

Figure 25: Fluxo de tuning do modelo

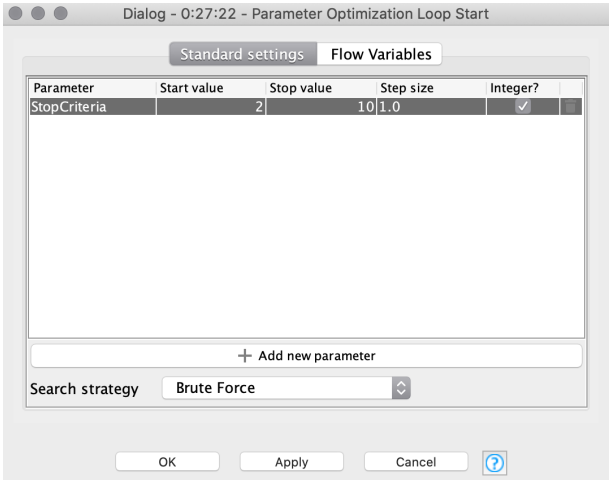

Figure 26: Definições aplicadas ao nodo *Parameter Optimization Loop Start*

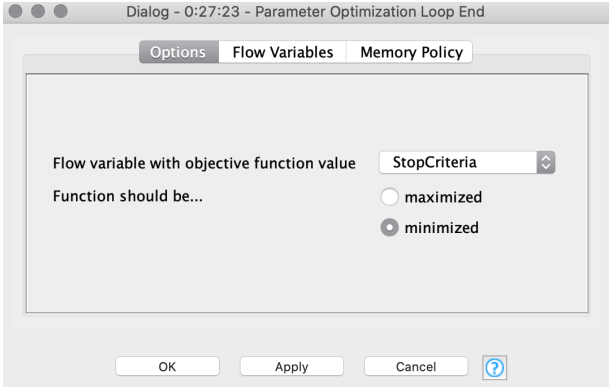

Figure 27: Definições aplicadas ao nodo *Parameter Optimization Loop End*

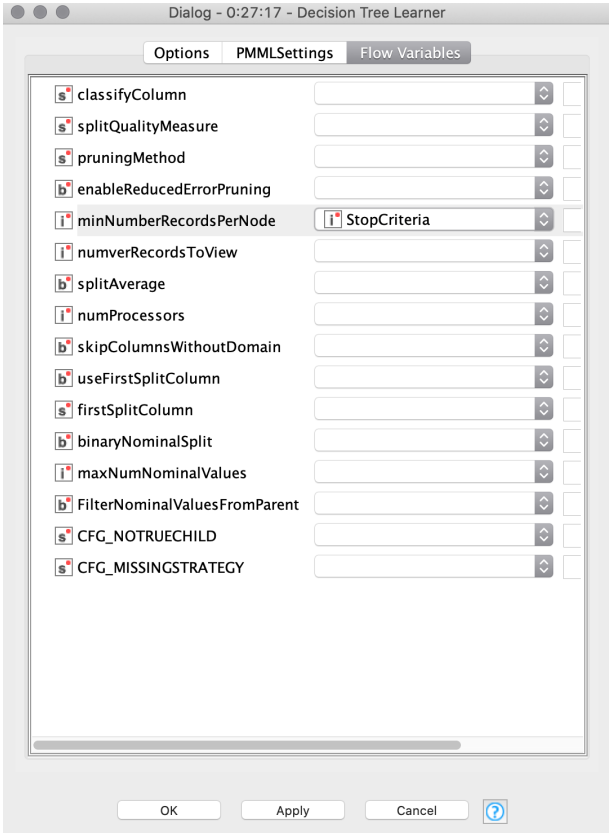

Figure 28: Aplicação da variavel de fluxo ao nodo *Decision Tree Learner*

Row ID	StopCriteria	Objective value
Best parameters	2	2

Figure 29: *Best Parameters*

4.2 Fazer o tuning do modelo com todas as possibilidades para a medida de qualidade

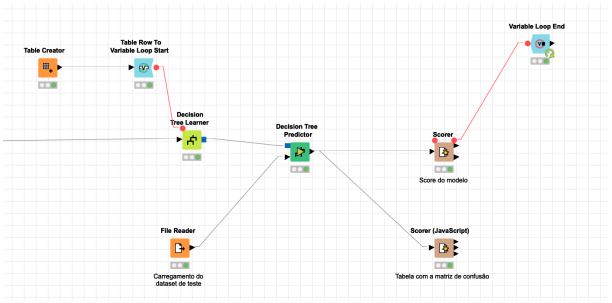

Figure 30: Fluxo de tuning do modelo para a medida de qualidade

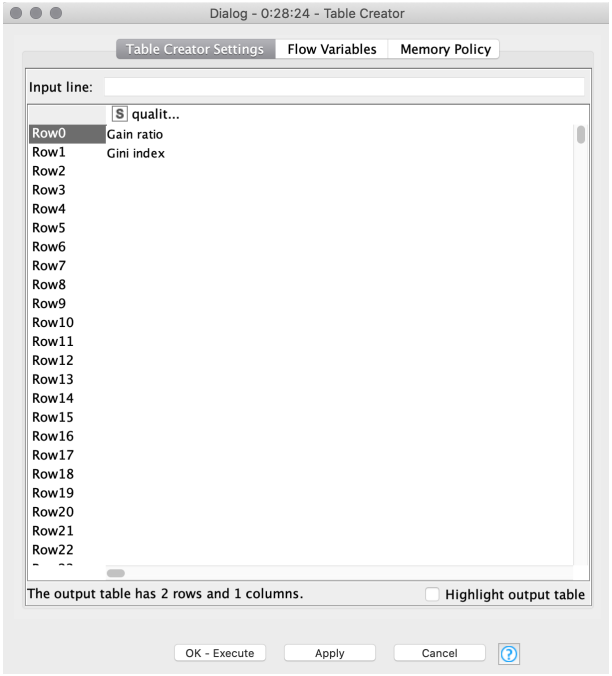

Figure 31: Criação da tabela com as variáveis a iterar

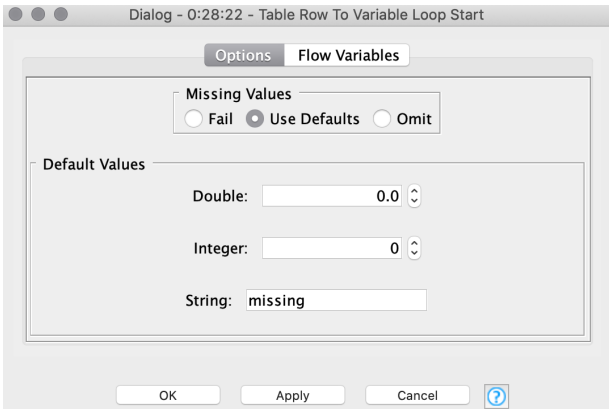

Figure 32: Definições aplicadas ao nodo de iteração da tabela

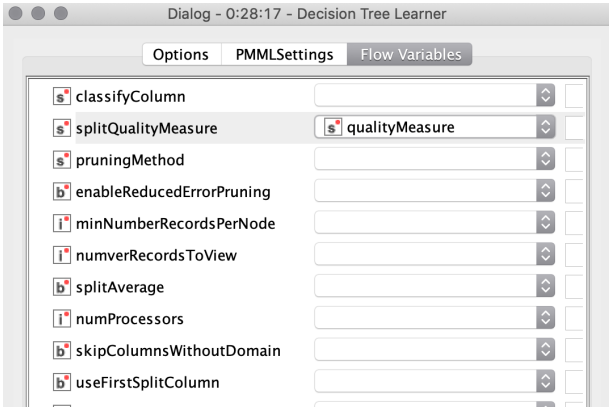

Figure 33: Variavél de fluxo aplicada ao nodo *Decision Tree Learner*

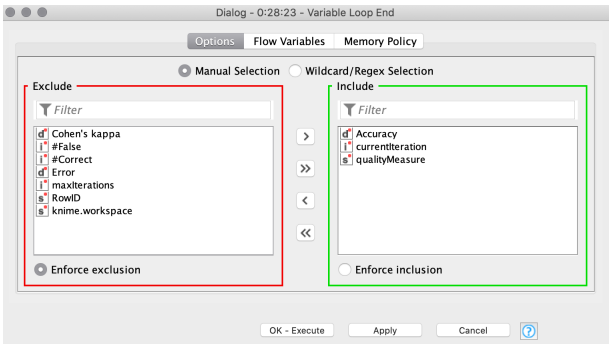

Figure 34: Definições aplicadas ao nodo *Variable Loop End*

A precisão do modelo é mais alta utilizando a feature *Gain Ratio*.

Row ID	Accuracy	currentIteration	qualityMeasure
Row0	0.647	0	Gain ratio
Row1	0.6	1	Gini index

Figure 35

4.3 Fazer o tuning do modelo com todas as possibilidades para o método de pruning

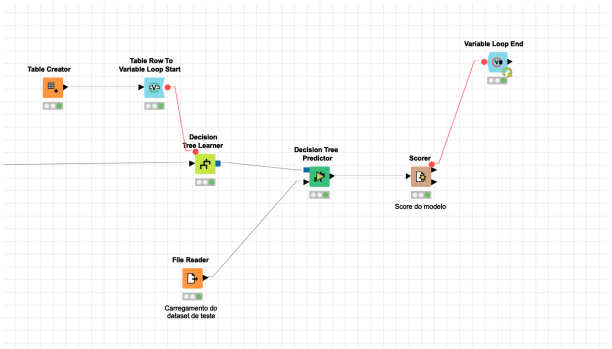

Figure 36: Fluxo de tuning do modelo para o método de pruning

Figure 37: Criação da tabela com as variáveis a iterar

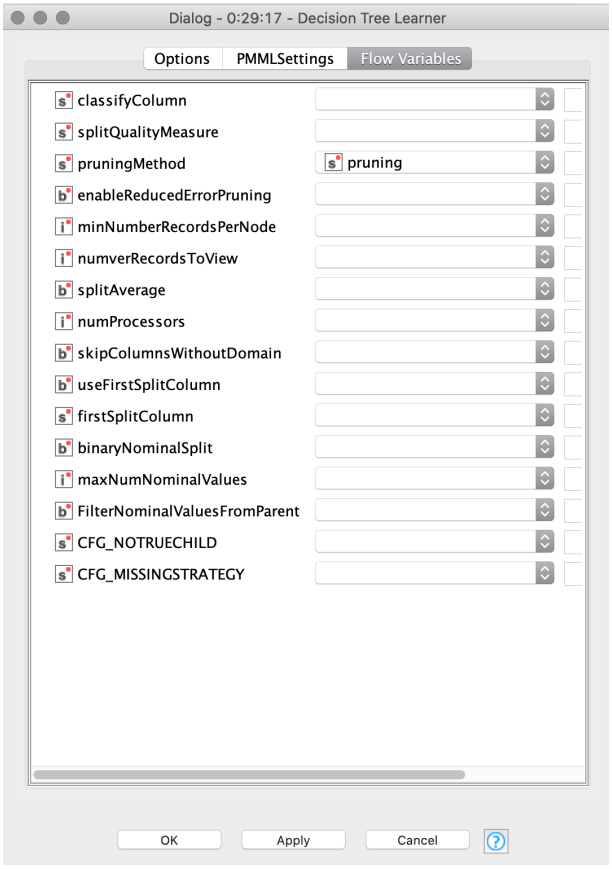

Figure 38: Variavél de fluxo aplicada ao nodo *Decision Tree Learner*

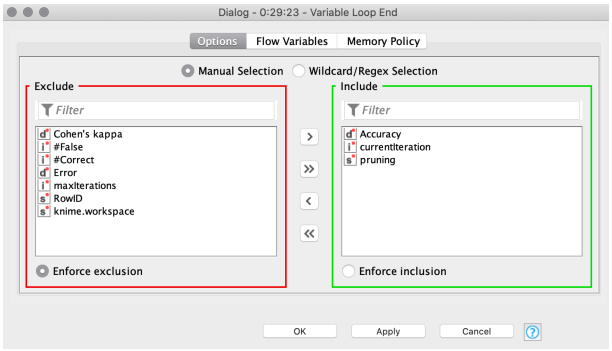

Figure 39: Definições aplicadas ao nodo *Variable Loop End*

A precisão do modelo é maior se a àrvore for treinada com o método de pruning MDL.

Row ID	D Accuracy	I currentIteration	S pruning
Row0	0.6	0	No pruning
Row1	0.694	1	MDL

Figure 40

4.4 Fazer o tuning dos parâmetros anteriores num único workflow. Guardar e analisar todos os resultados obtidos para cada combinação de hiper-parâmetros. Qual a combinação que oferece melhor performance? Existem grandes discrepâncias?

Figure 41: Criação da tabela com as variáveis a iterar

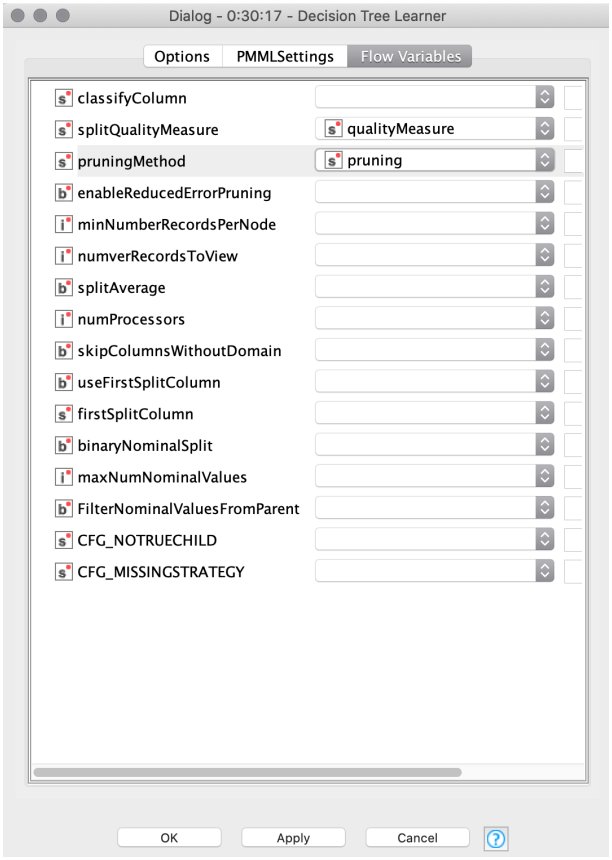

Figure 42: Variáveis de fluxo aplicadas ao nodo Decision Tree Learner

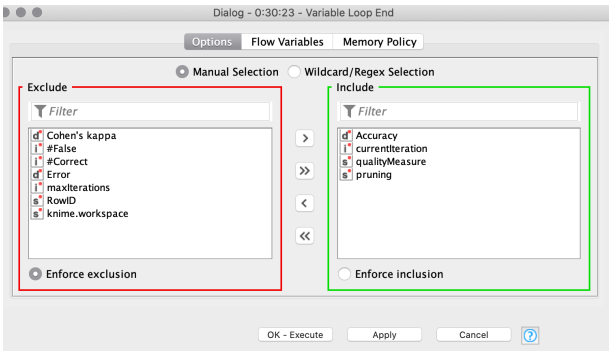

Figure 43: Definições aplicadas ao nodo *Variable Loop End*

A combinação que obtém uma precisão maior no treino da árvore é usar

como medida de qualidade Gain Ration e como método de pruning MDL. Na tabela seguinte apresenta-se a precisão obtida para cada uma das combinações.

Row ID	D Accur...	I curren...	S qualit...	S pruning
Row0	0.647	0	Gain ratio	No pruning
Row1	0.694	1	Gini index	MDL
Row2	0.706	2	Gain ratio	MDL
Row3	0.6	3	Gini index	No pruning

Figure 44: Precisão das combinações testadas

5 Exercício 5

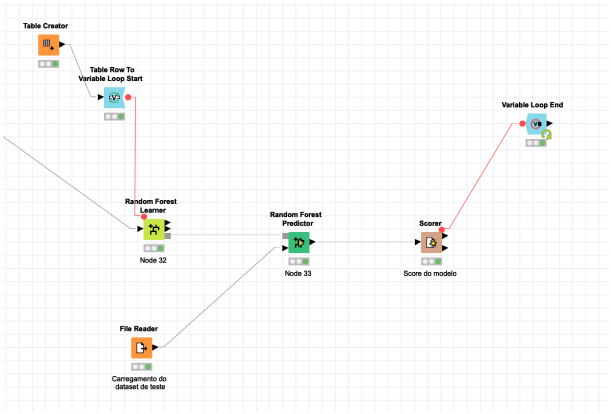

Figure 45: Fluxo para treino e tuning de uma *Random Forest*

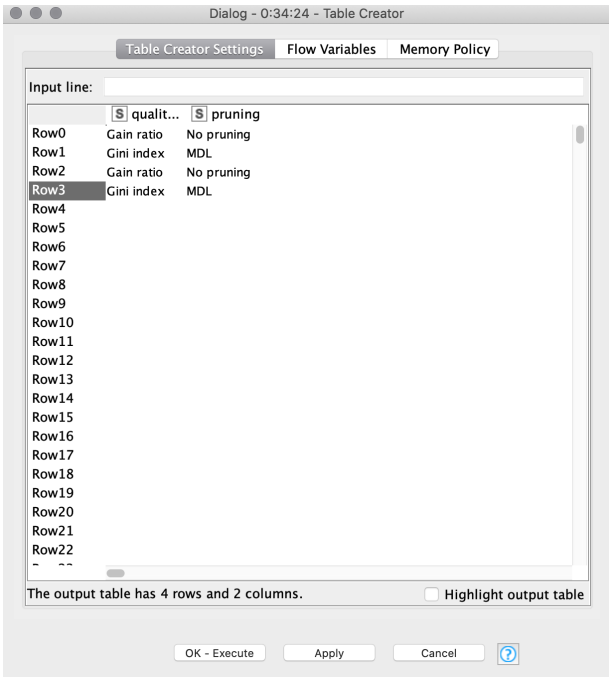

Figure 46: Criação da tabela com as variáveis a iterar

Na tabela seguinte apresenta-se a precisão obtida para cada uma das combinações.

Row ID	D Accur...	I curren...	S qualit...	S pruning
Row0	0.694	0	Gain ratio	No pruning
Row1	0.694	1	Gini index	MDL
Row2	0.694	2	Gain ratio	No pruning
Row3	0.694	3	Gini index	MDL

Figure 47: Precisão das combinações testadas

6 Exercício 6

Analisando as figuras 44 e 47 pode concluir-se que treinando uma *Decision Tree* com o dataset fornecido se obtém uma maior precisão do que treinando uma *Random Forest Tree*. A precisão mais alta obtida foi cerca de 70% aplicando a feature Gain Ratio e o método de pruning MDL ao treino de uma emphDecision Tree. Os restantes modelos não apresentam uma precisão muito distinta sendo o mais baixo 60%.