

Sistemas Baseados em Similaridade - SBS 2020

Mestrado Integrado em Engenharia Informática

Enunciado Prático 5



José Pinto

A84590

T1

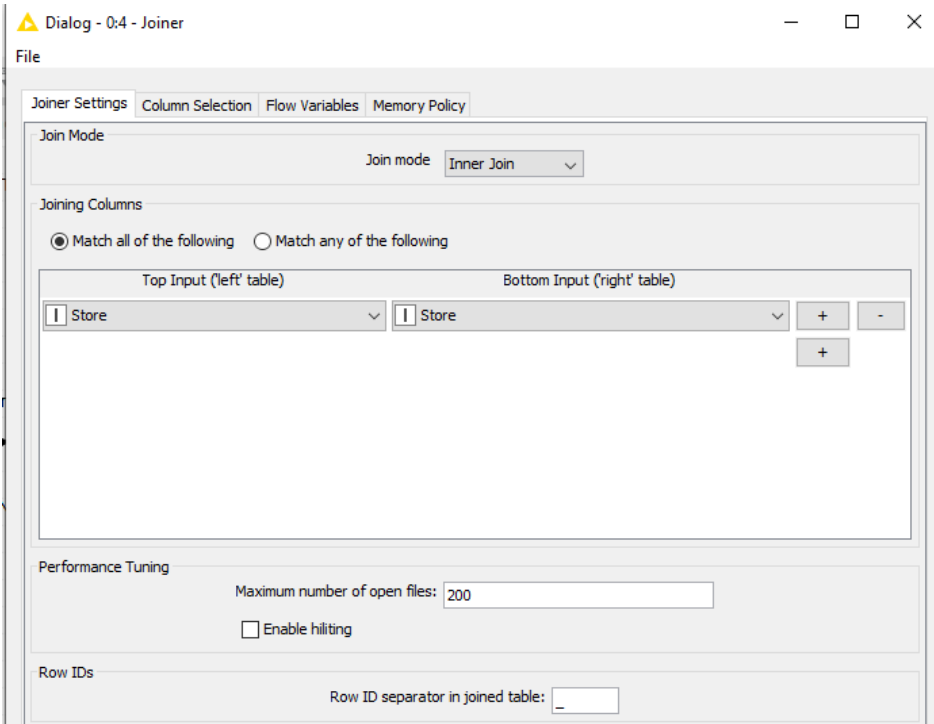


Figura 1 - Configuração Joiner

● A ● B

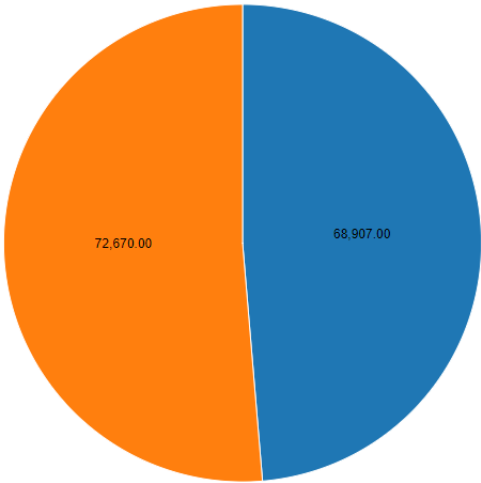


Figura 2 - Vendas por tipo

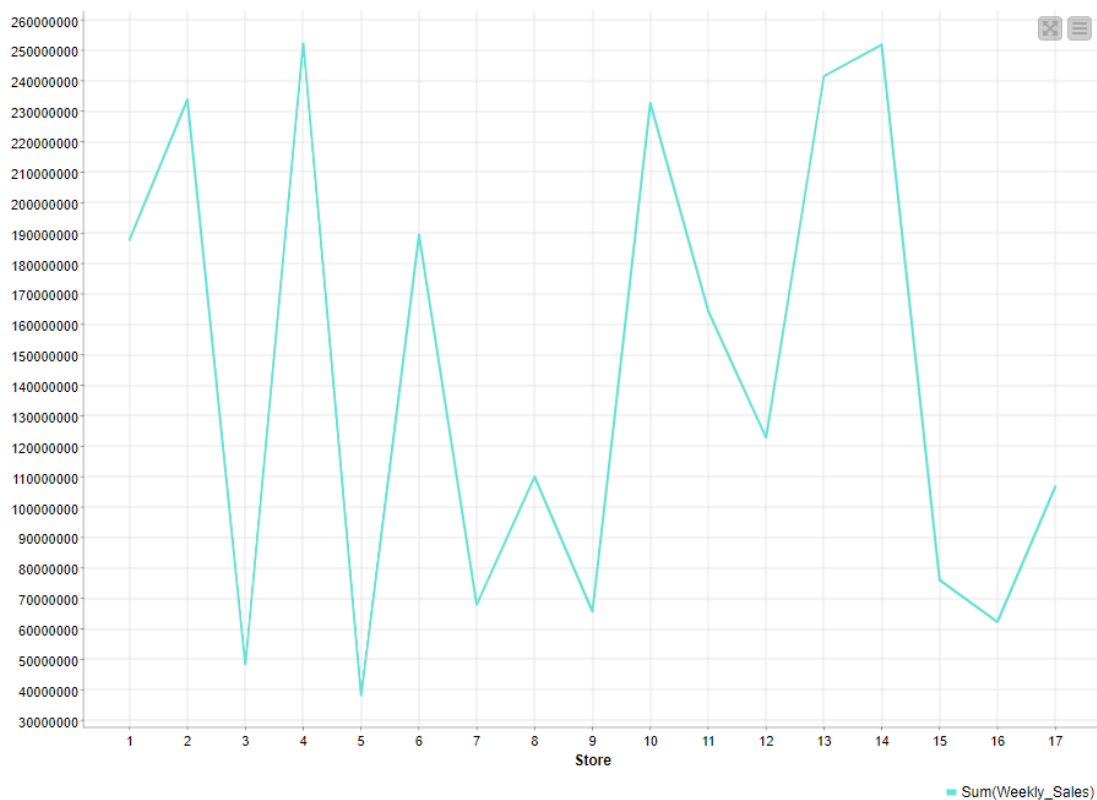


Figura 3 - Soma das vendas semanais por loja

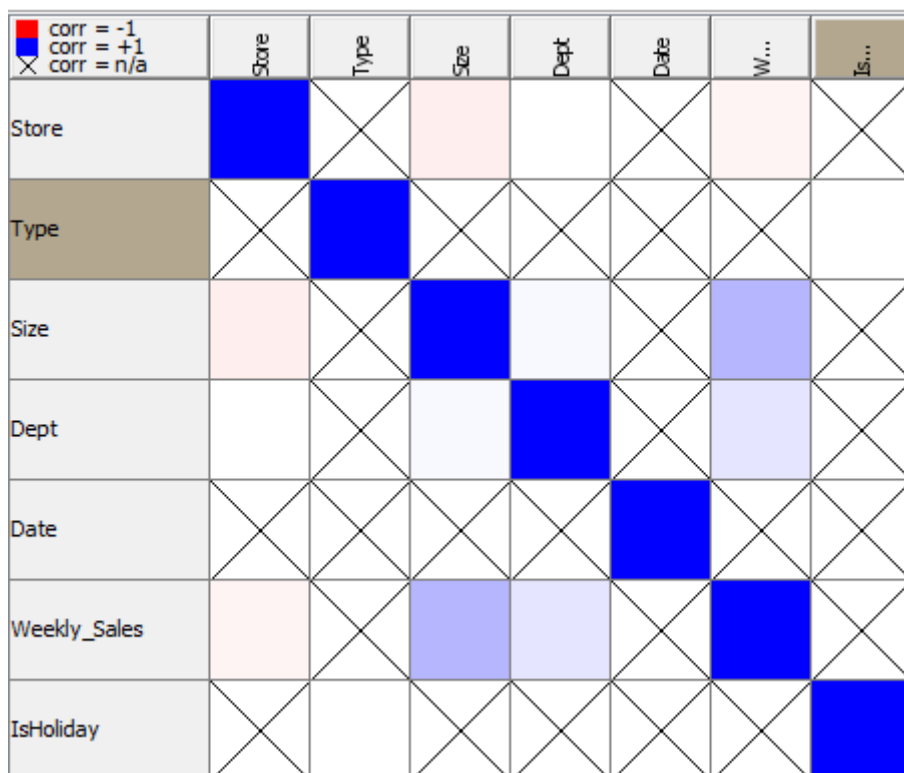


Figura 4 - Correlação entre features

T2

a)

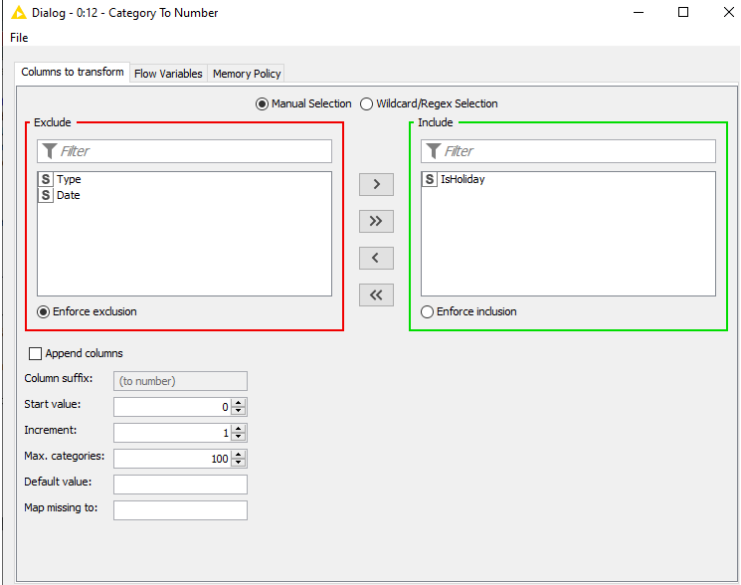


Figura 5 - Configuração Category To Number

b)

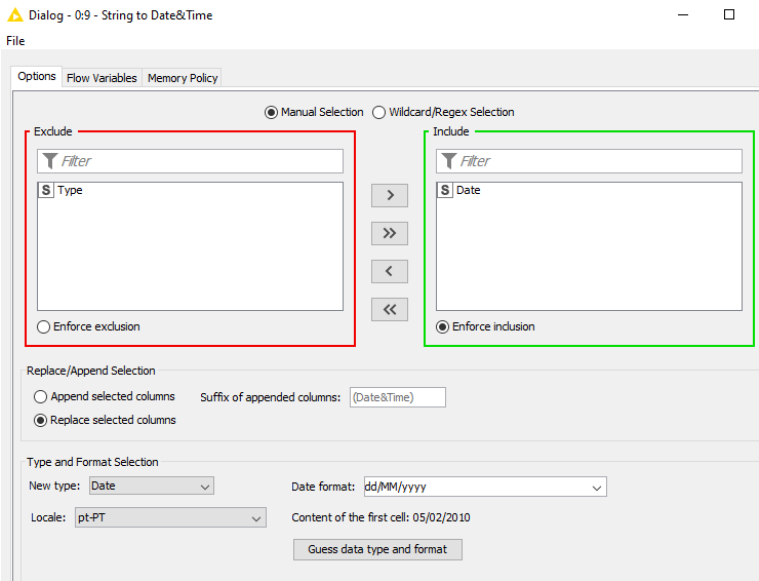


Figura 6 - Configuração String to Date&Time

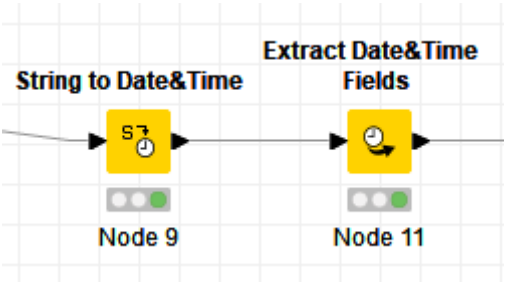


Figura 7 - Extração do ano e mês

c)

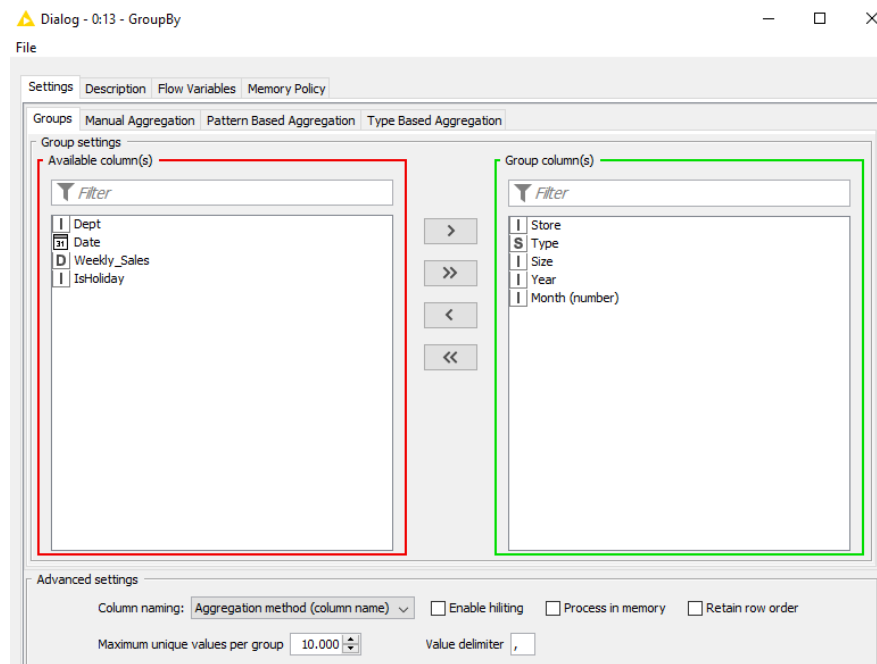


Figura 8 - Configuração GroupBy

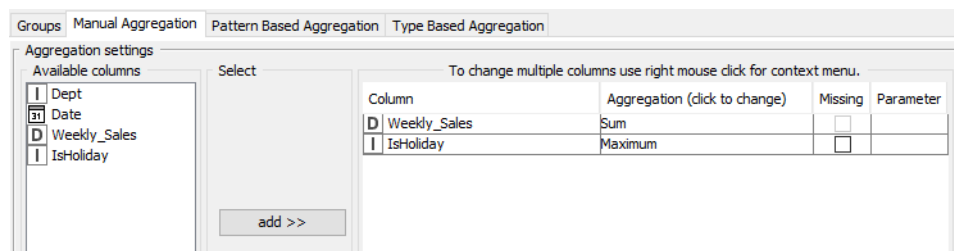


Figura 9 - GroupBy aggregation

d)

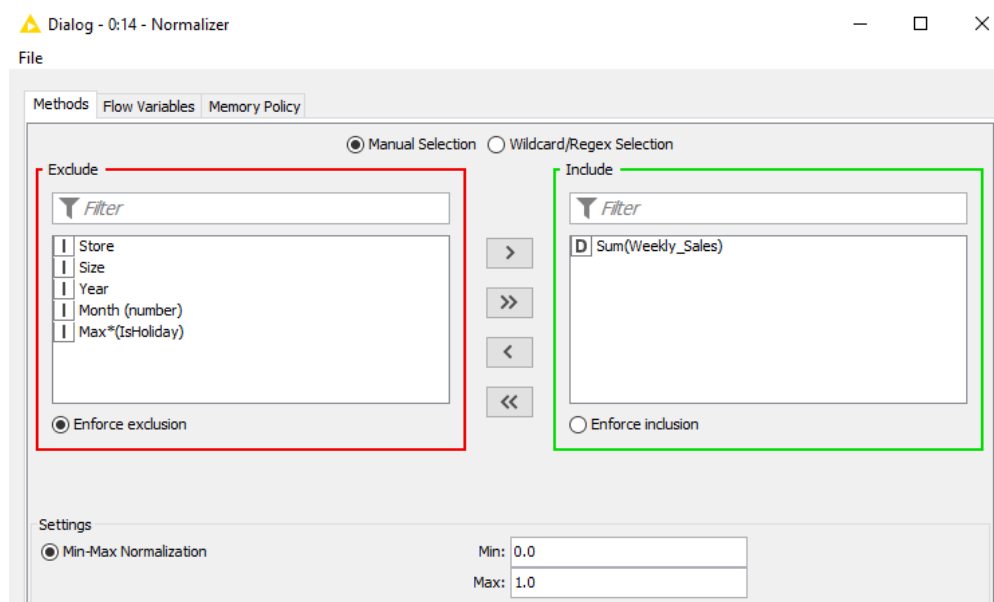


Figura 10 - Configuração Normalizer

e)

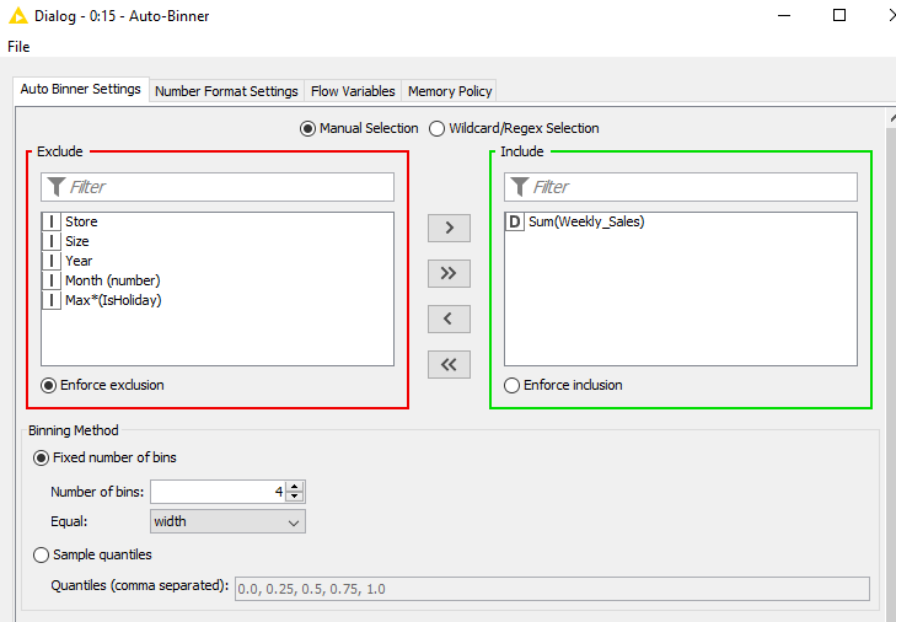


Figura 11 - Configuração Auto-Binner

f)

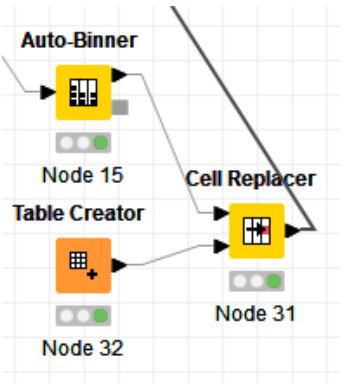


Figura 12 - Nodos para renomear os bin

Sum(W...
Medium
Medium
High
Medium
Medium
High
Medium
Medium
High
Medium
High
Medium
Medium
High
Medium
Medium
High
Medium
High
Medium

Figura 13 - Bin finais

T3

a)

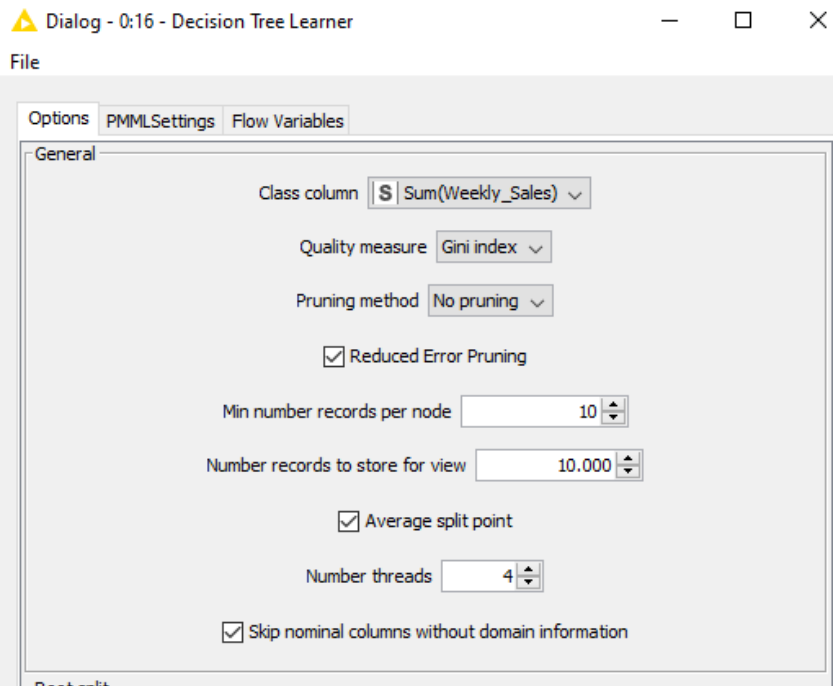


Figura 14 - Decision Tree Learner

Accuracy statistics - 0:18 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 5 Spec - Columns: 11 Properties Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifiaty	D F-meas...	D Accuracy
Low	59	5	75	4	0.937	0.922	0.937	0.938	0.929	?
Medium	21	7	107	8	0.724	0.75	0.724	0.939	0.737	?
High	43	8	89	3	0.935	0.843	0.935	0.918	0.887	?
Very High	0	0	138	5	0	?	0	1	?	?
Overall	?	?	?	?	?	?	?	?	?	0.86

Figura 15 - Precisão

b)

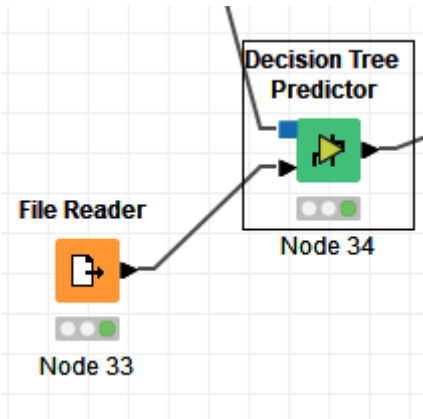


Figura 16 - Nodos para a previsão de vendas dataset teste

S	Predicti...
	Medium
	Medium
	Medium
	Medium
	Medium
	High
	High
	High
	High
	High

Figura 17 - Excerto coluna de previsões

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy
High	9	16	45	15	0.375	0.36	0.375	0.738	0.367	?
Low	20	23	42	0	1	0.465	1	0.646	0.635	?
Medium	2	15	49	19	0.095	0.118	0.095	0.766	0.105	?
Very High	0	0	65	20	0	?	0	1	?	?
Overall	?	?	?	?	?	?	?	?	?	0.365

Figura 18 - Precisão do modelo

c)

Row ID	I High	I Low	I Medium	I Very High
High	9	4	11	0
Low	0	20	0	0
Medium	0	19	2	0
Very High	16	0	4	0

Figura 19 - Matriz de confusão

Row ID	I TruePo...	I Count (...)	D Relativ...	I FalsePo...	I Count (...)	D Relativ...	I TrueNe...	I Count (...)	D Relativ...	I FalseN...	I Count (...)	D Relativ...
Row0	9	1	0.2	16	1	0.2	45	1	0.2	15	1	0.2
Row1	20	1	0.2	23	1	0.2	42	1	0.2	0	1	0.2
Row2	2	1	0.2	15	1	0.2	49	1	0.2	19	1	0.2
Row3	0	1	0.2	0	1	0.2	65	1	0.2	20	1	0.2
Row4	?	1	0.2	?	1	0.2	?	1	0.2	?	1	0.2

Figura 20 - Statistics Occurences Table

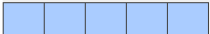

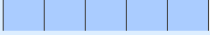
Column	No. missings	Histogram
TruePositives	1	
FalsePositives	1	
TrueNegatives	1	
FalseNegatives	1	

Figura 21 - Histograma com o resultado das previsões

T4

a)

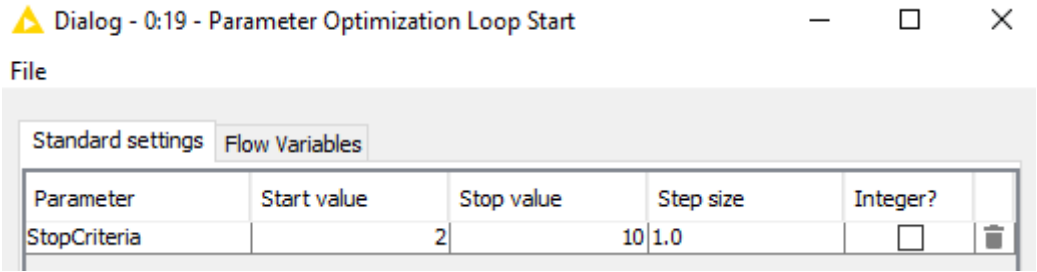


Figura 22 - Configuração Parameter Optimization Loop Start

b & c)

S	Pruning	S	Quality ...
	No pruning		Gain ratio
	MDL		Gini index
	No pruning		Gini index
	MDL		Gain ratio

Figura 23 - Possibilidades qualidade e pruning

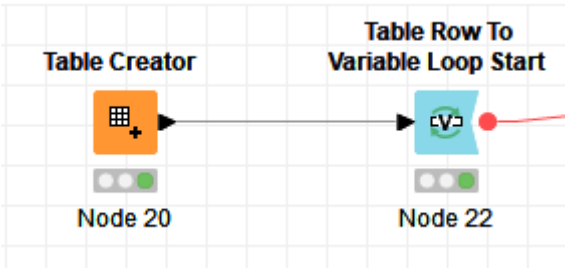


Figura 24 - Loop para testar possibilidades

d)

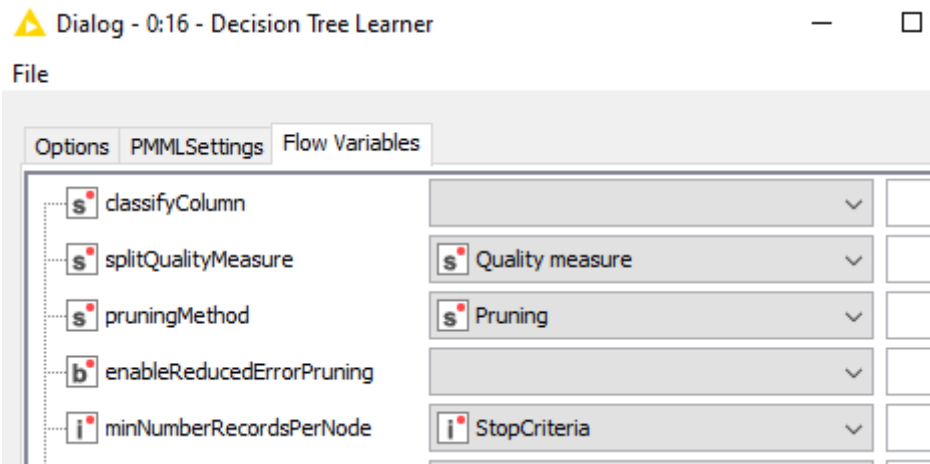


Figura 25 - Decision Tree Learner Flow Variables

Row ID	I stopCri...	D Objecti...	S Pruning	S Quality ...
Row0	3	0.86	No pruning	Gain ratio
Row1	2	0.832	MDL	Gini index
Row2	3	0.839	No pruning	Gini index
Row3	2	0.86	MDL	Gain ratio

Figura 26 - Combinação de hiper-parâmetros e precisão dataset treino

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy
High	9	16	45	15	0.375	0.36	0.375	0.738	0.367	?
Low	20	20	45	0	1	0.5	1	0.692	0.667	?
Medium	3	17	47	18	0.143	0.15	0.143	0.734	0.146	?
Very High	0	0	65	20	0	?	0	1	?	?
Overall	?	?	?	?	?	?	?	?	?	0.376

Figura 27 - Precisão melhor combinação de hiper-parâmetros para dataset teste

Como podemos ver pela análise de resultados as melhores combinações de hiper-parâmetros são:

- Nº de registos = 3, qualidade = Gain ratio, pruning = No pruning;
- Nº de registos = 2, qualidade = Gain ratio, pruning = MDL;

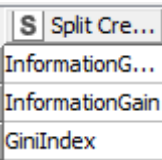


Figura 28 - Split Criterion values

Parameter	Start value	Stop value	Step size	Integer?	
limit number of levels	10	100	10.0	<input checked="" type="checkbox"/>	
Minimum node size	2	10	1.0	<input checked="" type="checkbox"/>	

Figura 29 - maxLevels & minNodeSize values

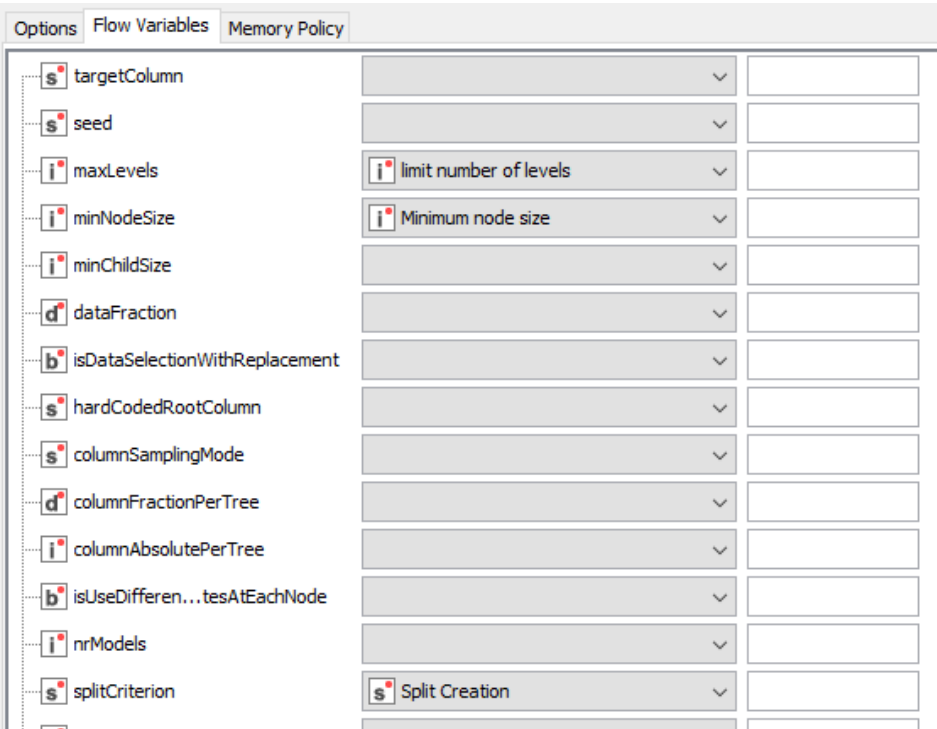


Figura 30 - Random Forest flow variables

Row ID	I limit nu...	I Minimu...	D Objecti...	S Split Creation
Row0	10	2	0.376	InformationGainRatio
Row 1	10	2	0.376	InformationGain
Row2	10	2	0.365	Gini

Figura 31 - Precisão Random Forest dataset teste

T6

À priori seria de esperar uma considerável subida de precisão ao passar de um modelo de decision tree para um com random forest. Isto porque uma random forest é na verdade um conjunto de decision trees, onde cada árvore individual faz uma previsão e a classe com mais votos torna-se a previsão do modelo. O que significa que à partida modelos que utilizem random forest terão melhores resultados que um modelo com apenas uma decision tree.

No entanto, a análise das performances dos dois modelos não vem defender o que inicialmente era previsto, visto que as melhores combinações dos dois modelos têm exatamente a mesma precisão. Para além disso, a precisão final em ambos os modelos para o dataset teste é extremamente baixa relativamente aos valores alcançados na precisão do dataset treino.

Estes resultados podem ser devidos à distribuição dos dados dos datasets, ou algum erro no desenvolvimento de ambos os modelos. Infelizmente, não foi possível apurar a verdadeira causa.