

Universidade do Minho

Departamento de Informática

Sistemas Baseados em Similaridade

Trabalho Prático Individual 1

Gonalo Almeida (A84610)

19 de Outubro de 2020

T1. Instalar a plataforma Knime.

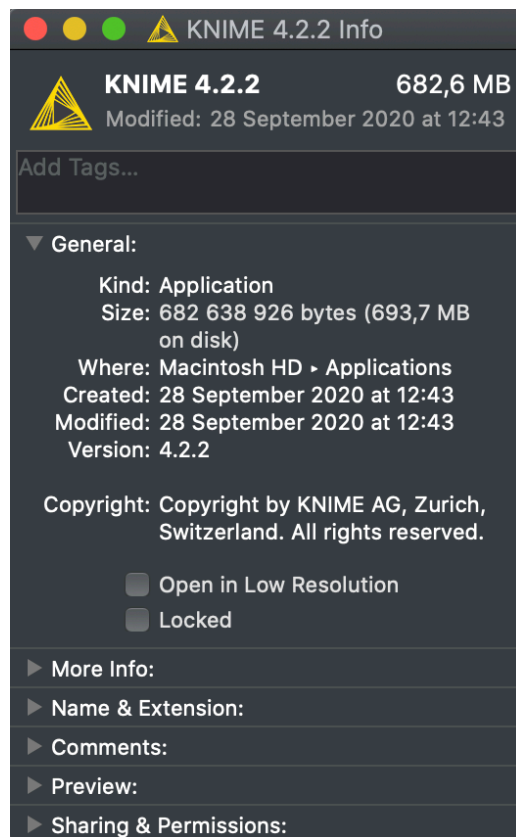


Figura 1 - Informações da Aplicação Knime

T2. Desenvolver um workflow que, utilizando um nodo reader, faz a correta leitura do dataset disponível em <https://bit.ly/3hXCwIG>.

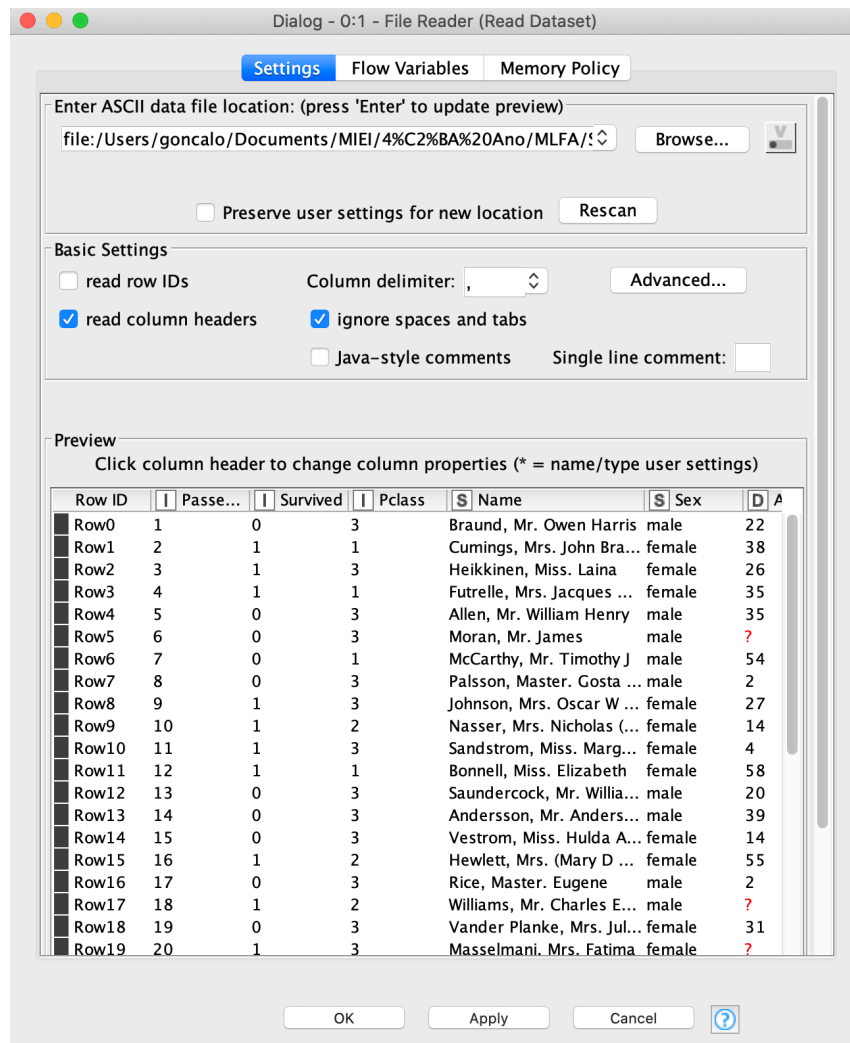
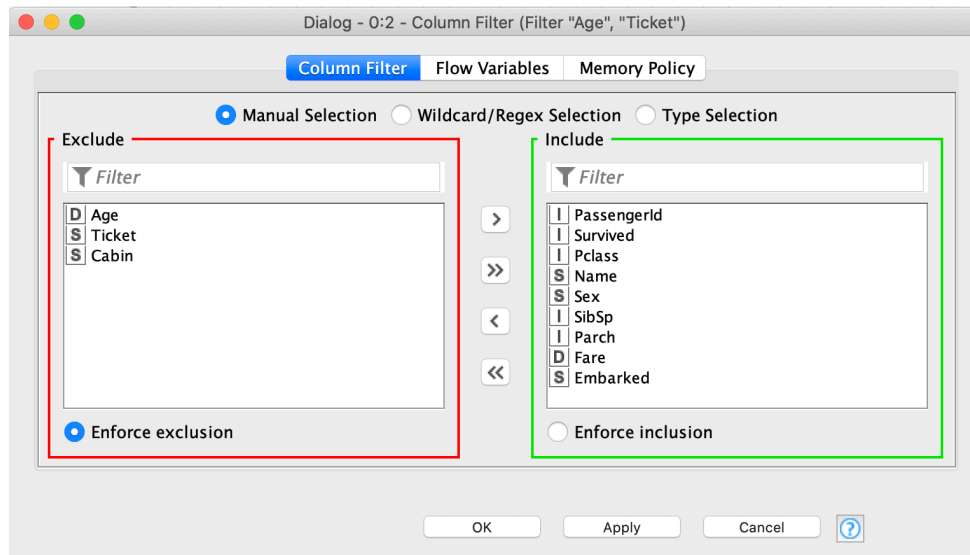


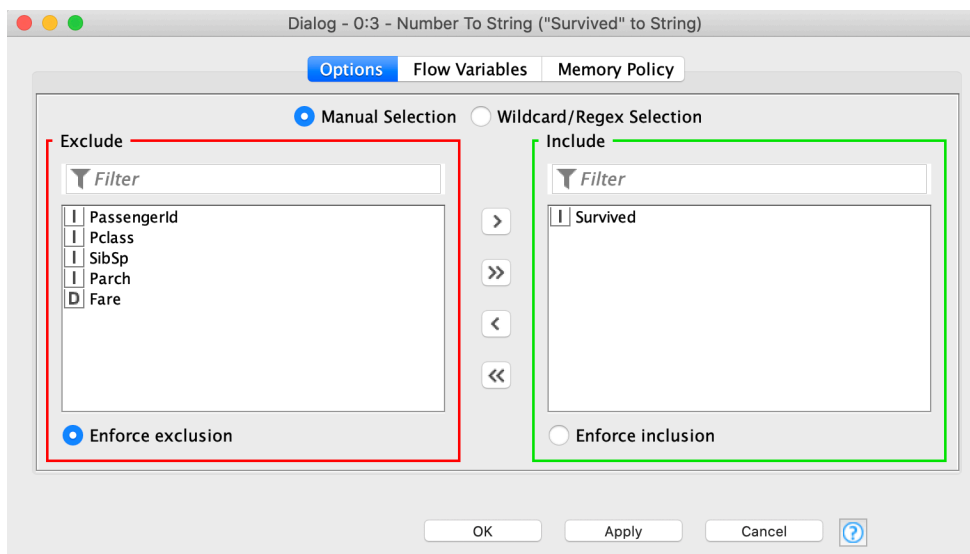
Figura 2 - Nodo Reader

T3. Utilizar um conjunto de nodos para:

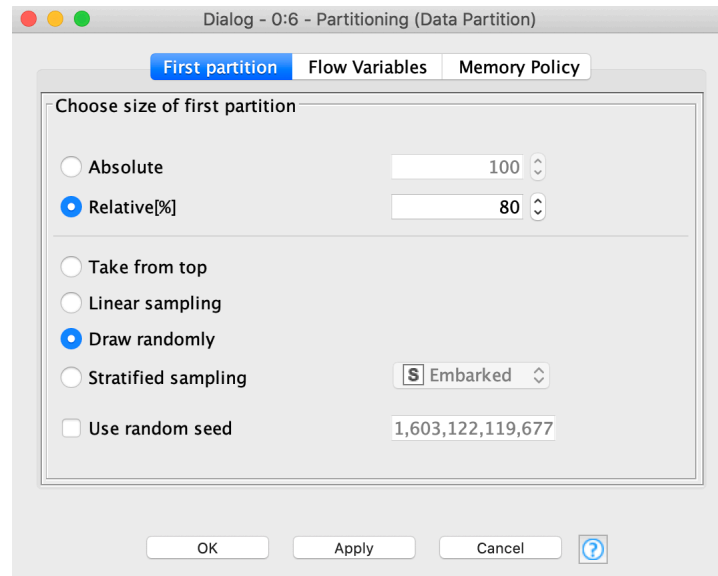
- Filtrar as colunas “Age”, “Ticket” e “Cabin”;



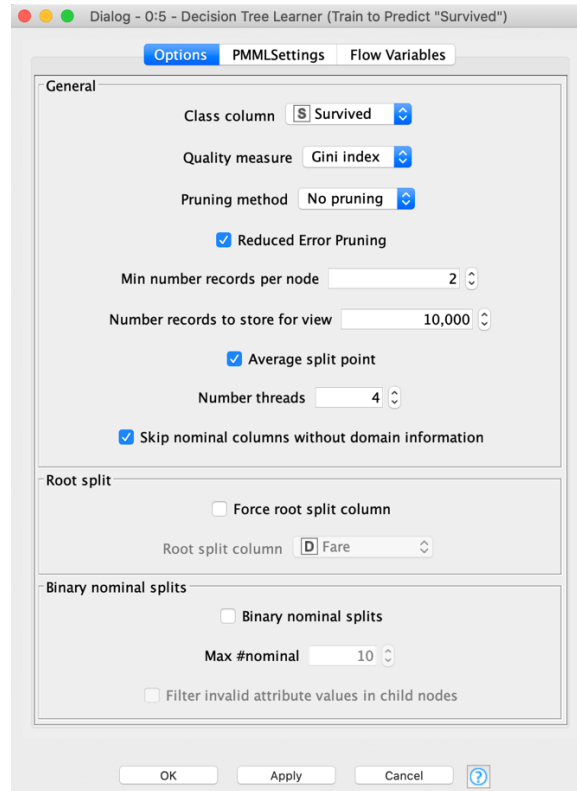
- Fazer o cast da coluna “Survived” para String;

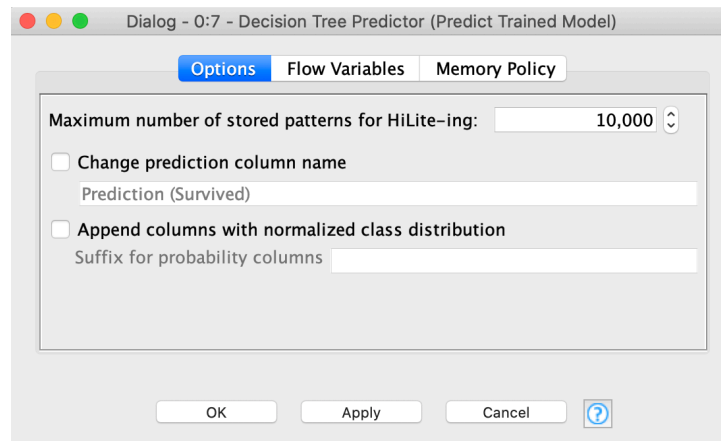


- Particionar os dados, de forma aleatória, utilizando 80% para aprendizagem e 20% para teste;

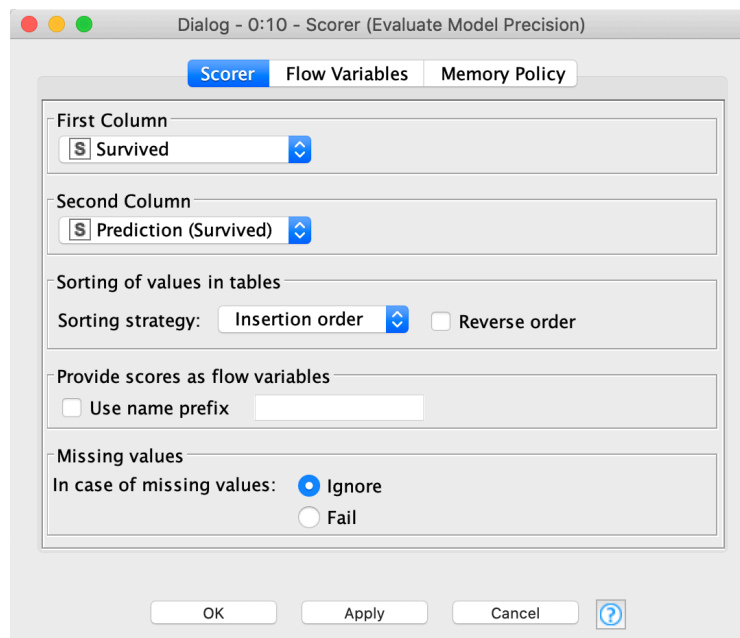


- Aplicar um nodo Decision Tree Learner para treinar uma Árvore de Decisão e um Decision Tree Predictor para obter previsões utilizando o modelo treinado;





- Avaliar a precisão (accuracy) do modelo utilizando o nodo Scorer e a respetiva matriz de confusão.



T4. Experimentar várias combinações de parâmetros no nodo *Decision Tree Learner* e documentar as performances obtidas.

The screenshot shows the 'PMMLSettings' dialog box for the 'Decision Tree Learner' node. The 'General' tab is selected, displaying various configuration options. The 'Class column' is set to 'Survived', the 'Quality measure' is 'Gini index', and the 'Pruning method' is 'No pruning'. The 'Reduced Error Pruning' checkbox is checked. The 'Min number records per node' is set to 2, and the 'Number records to store for view' is set to 10,000. The 'Average split point' checkbox is checked, and the 'Number threads' is set to 4. The 'Skip nominal columns without domain information' checkbox is also checked. The 'Root split' section has the 'Force root split column' checkbox unchecked, and the 'Root split column' is set to 'Fare'. The 'Binary nominal splits' section has the 'Binary nominal splits' checkbox unchecked, and the 'Max #nominal' is set to 10. The 'Filter invalid attribute values in child nodes' checkbox is unchecked. At the bottom, there are buttons for 'OK', 'Apply', 'Cancel', and a help icon.

Figura 3 - Configuração 1

The screenshot shows the 'Confusion Matrix - 0:10 - Scorer (Evaluate Model Precision)' window. It displays a confusion matrix for the 'Survived' variable with two classes (0 and 1). The matrix shows 85 correct classifications for class 0, 14 incorrect classifications for class 0, 29 incorrect classifications for class 1, and 51 correct classifications for class 1. Summary statistics at the bottom indicate 136 correct classifications, 43 wrong classifications, an accuracy of 75.978%, an error rate of 24.022%, and a Cohen's kappa (κ) of 0.505.

Survived \ ...	0	1
0	85	14
1	29	51

Correct classified: 136 Wrong classified: 43
Accuracy: 75.978 % Error: 24.022 %
Cohen's kappa (κ) 0.505

Figura 4 - Matriz de Confusão 1

Options PMMLSettings Flow Variables

General

Class column

Quality measure

Pruning method

☒ Reduced Error Pruning

Min number records per node

Number records to store for view

☒ Average split point

Number threads

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column

Binary nominal splits

☐ Binary nominal splits

Max #nominal

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Figura 5 - Configuração 2

Confusion Matrix - 0:10 - Scorer (Evaluate Model Precision)

File	Hilite
Survived \ ...	0 1
0	93 6
1	33 47

Correct classified: 140 Wrong classified: 39

Accuracy: 78.212 % Error: 21.788 %

Cohen's kappa (κ) 0.545

Figura 6 - Matriz de Confusão 1

Options PMMLSettings Flow Variables

General

Class column

Quality measure

Pruning method

☒ Reduced Error Pruning

Min number records per node

Number records to store for view

☒ Average split point

Number threads

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column

Binary nominal splits

☐ Binary nominal splits

Max #nominal

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Figura 7 - Configuração 3

Confusion Matrix - 0:10 - Scorer (Evaluate Model Precision)

File	Hilite		
Survived \ ...	0	1	
0	78	21	
1	24	56	

Correct classified: 134 Wrong classified: 45

Accuracy: 74.86 % Error: 25.14 %

Cohen's kappa (κ) 0.49

Figura 8 - Matriz de Confusão 3

Options PMMLSettings Flow Variables

General

Class column

Quality measure

Pruning method

☒ Reduced Error Pruning

Min number records per node

Number records to store for view

☒ Average split point

Number threads

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column

Binary nominal splits

☐ Binary nominal splits

Max #nominal

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Figura 9 - Configuração 4

Confusion Matrix - 0:10 - Scorer (Evaluate Model Precision)

File	Hilite
Survived \ ...	0 1
0	93 6
1	33 47

Correct classified: 140 Wrong classified: 39

Accuracy: 78.212 % Error: 21.788 %

Cohen's kappa (k) 0.545

Figura 10 - Matriz de Confusão 4