

Universidade do Minho

Departamento de Informática

Mestrado em Engenharia de Sistemas

Perfil de Machine Learning

Fundamentos e Aplicações Sistemas Baseados em Similaridade

1º Ano, 1º Semestre

Ano letivo 2020/2021

Enunciado Prático nº 1

16 de outubro de 2020

Joel Costa Carvalho

PG42837

Índice

[Tarefas - 4 -](#_Toc53911505)

[T1 – Instalar a plataforma *Knime* - 4 -](#_Toc53911506)

[T2 – Leitura do *dataset* - 4 -](#_Toc53911507)

[T3 – Conjunto de nodos - 5 -](#_Toc53911508)

[a) Filtrar as colunas “Age”, “Ticket” e “Cabin”; - 5 -](#_Toc53911509)

[b) Fazer o cast da coluna “Survived” para *String*; - 5 -](#_Toc53911510)

[c) Particionar os dados, de forma aleatória, utilizando 80% para aprendizagem e 20% para teste; - 6 -](#_Toc53911511)

[d) Aplicar um nodo *Decision Tree Learner* para treinar uma Árvore de Decisão e um *Decision Tree Predictor* para obter previsões utilizando o modelo treinado; - 6 -](#_Toc53911512)

[e) Avaliar a precisão do modelo utilizando o nodo *Scorer* e a respetiva matriz de confusão. - 7 -](#_Toc53911513)

[T4 – Experimentar várias combinações de parâmetros no nodo *Decision Tree Learner* e documentar as performances obtidas. - 8 -](#_Toc53911514)

Índice de figuras

[Figura 1 - Plataforma Knime - 4 -](#_Toc53910539)

[Figura 2 - Leitura dataset - 4 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP1/TP1.docx#_Toc53910540)

[Figura 3 - Operação de filtragem - 5 -](#_Toc53910541)

[Figura 4 - Operação de casting - 5 -](#_Toc53910542)

[Figura 5 - Operação de partição - 6 -](#_Toc53910543)

[Figura 6 - Previsões do modelo treinado - 6 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP1/TP1.docx#_Toc53910544)

[Figura 7 - Matriz confusão - 7 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP1/TP1.docx#_Toc53910545)

[Figura 8 - Avaliação da precisão do modelo - 7 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP1/TP1.docx#_Toc53910546)

[Figura 9 - Avaliação da precisão - 7 -](#_Toc53910547)

Tarefas

**T1 – Instalar a plataforma *Knime***

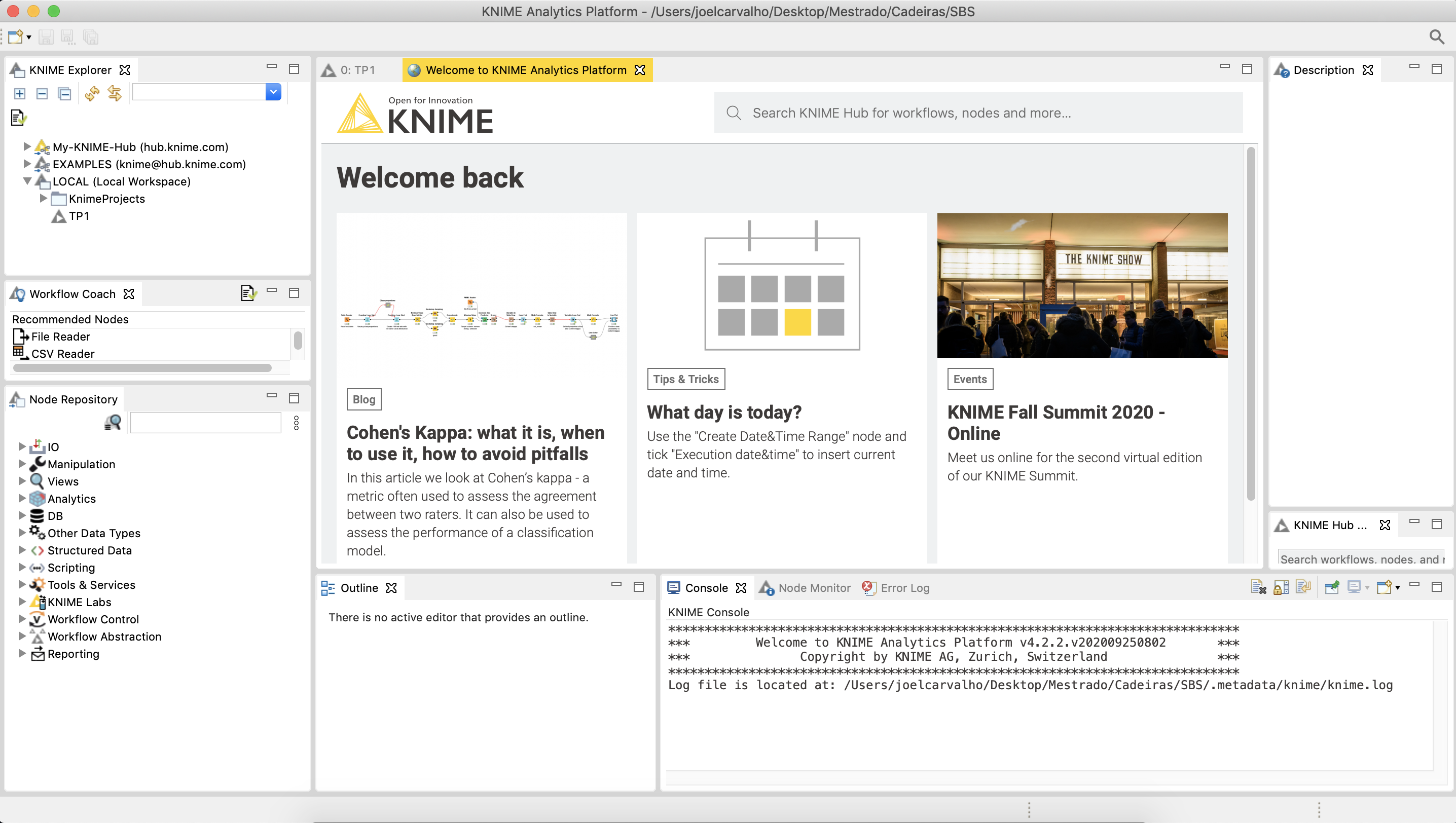


Figura 1 - Plataforma Knime

**T2 – Leitura do *dataset***

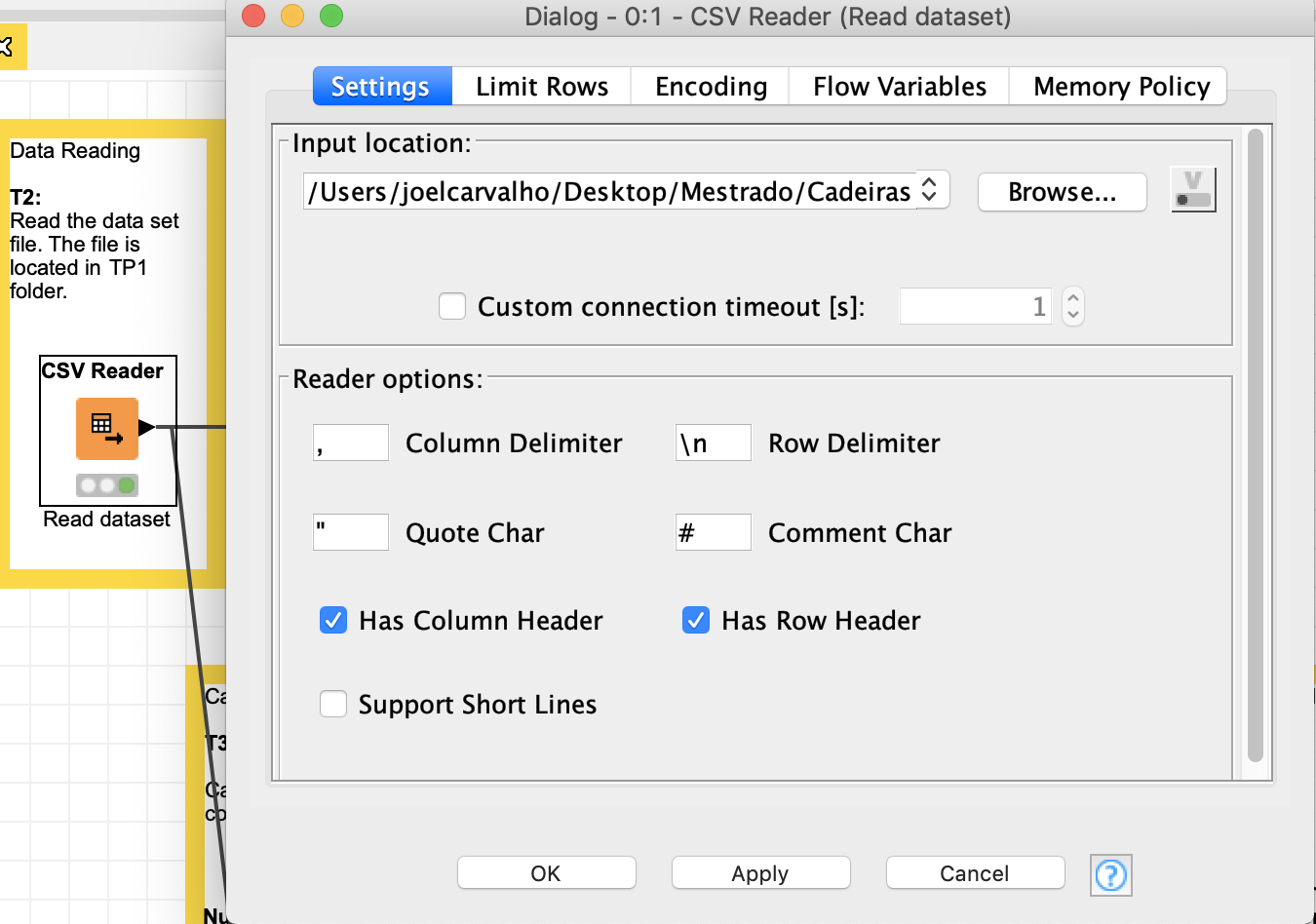


Figura 2 - Leitura dataset

**T3 – Conjunto de nodos**

1. Filtrar as colunas “Age”, “Ticket” e “Cabin”;

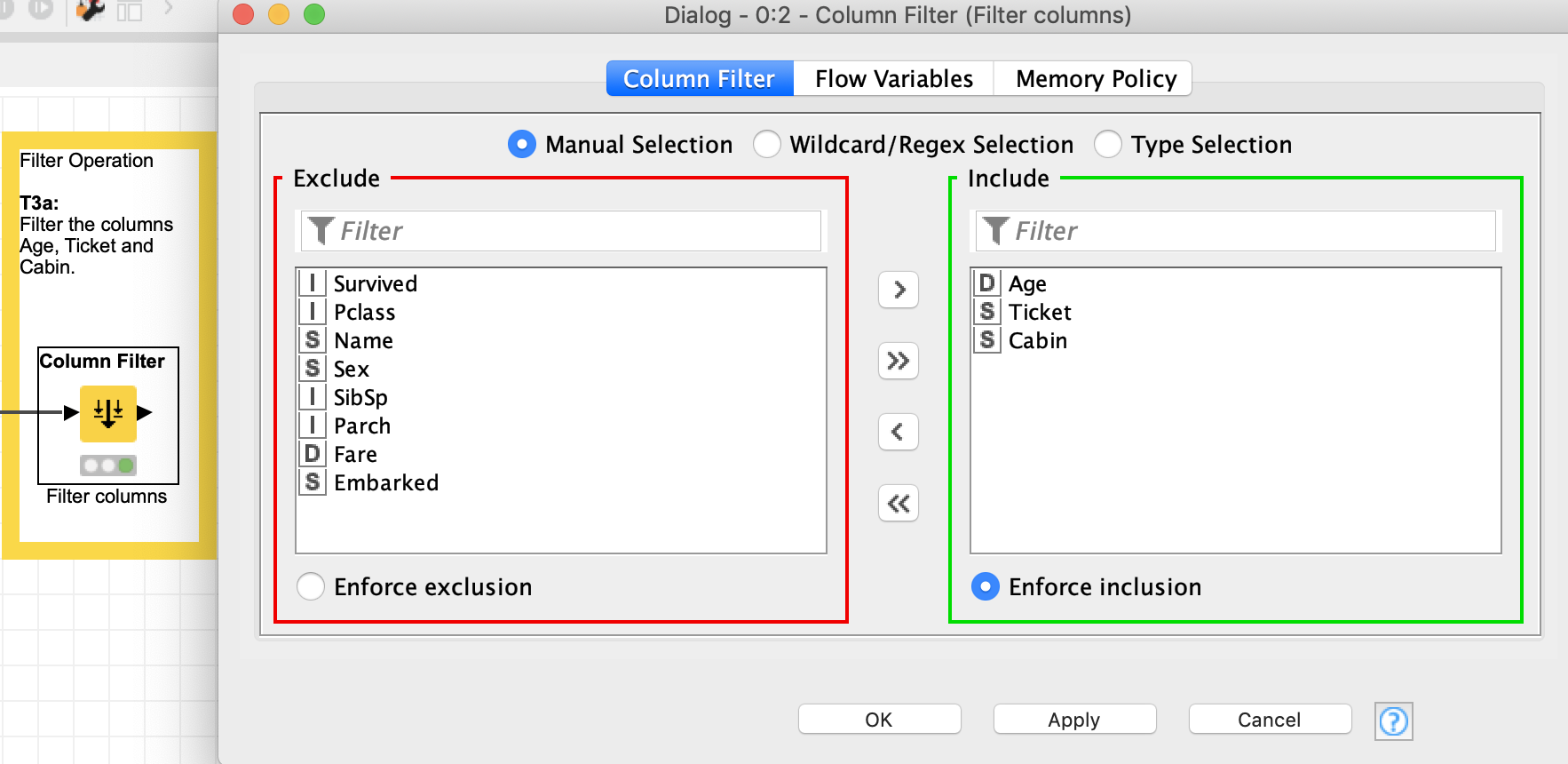


Figura 3 - Operação de filtragem

1. Fazer o cast da coluna “Survived” para *String*;

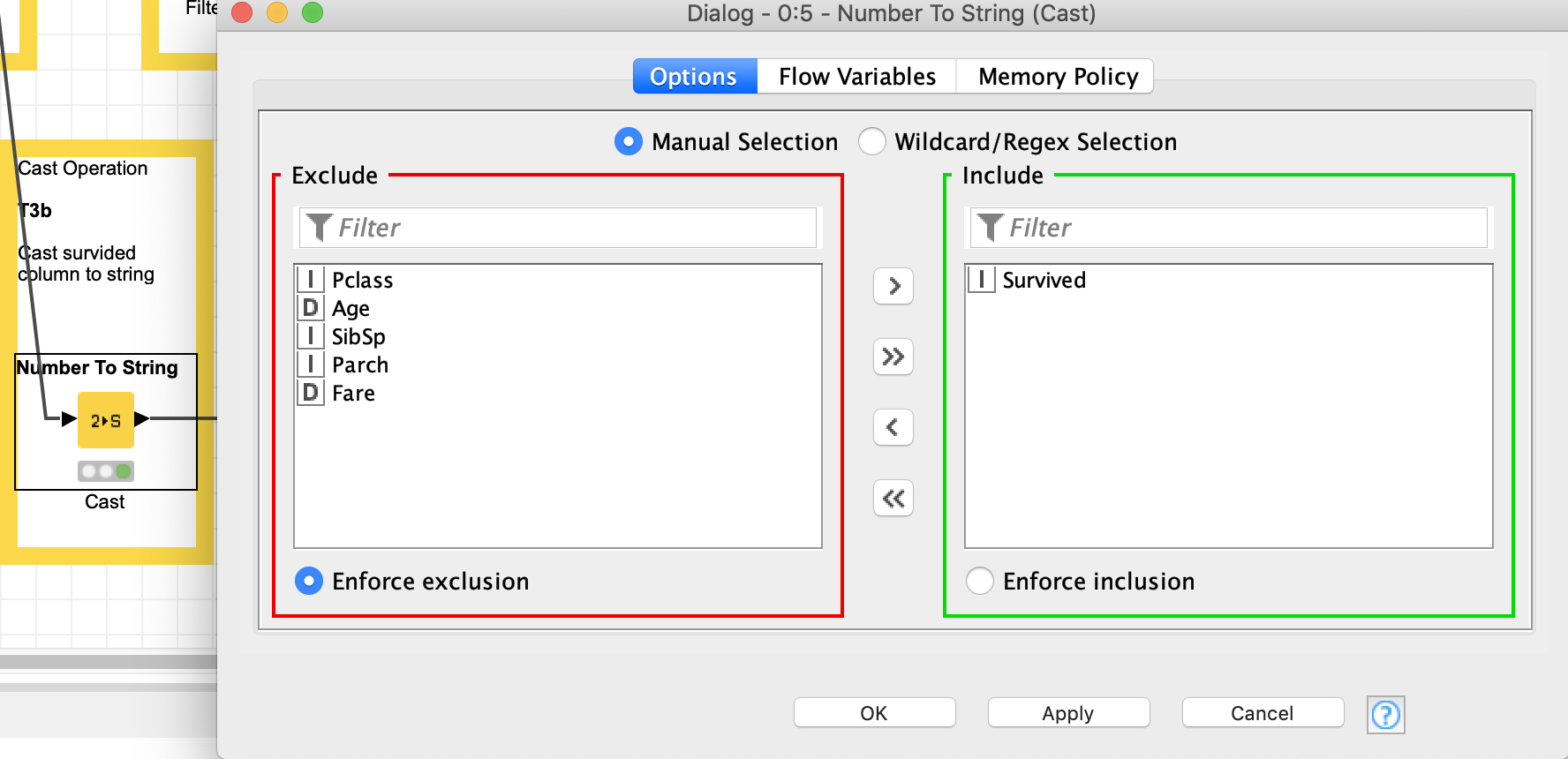


Figura 4 - Operação de casting

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

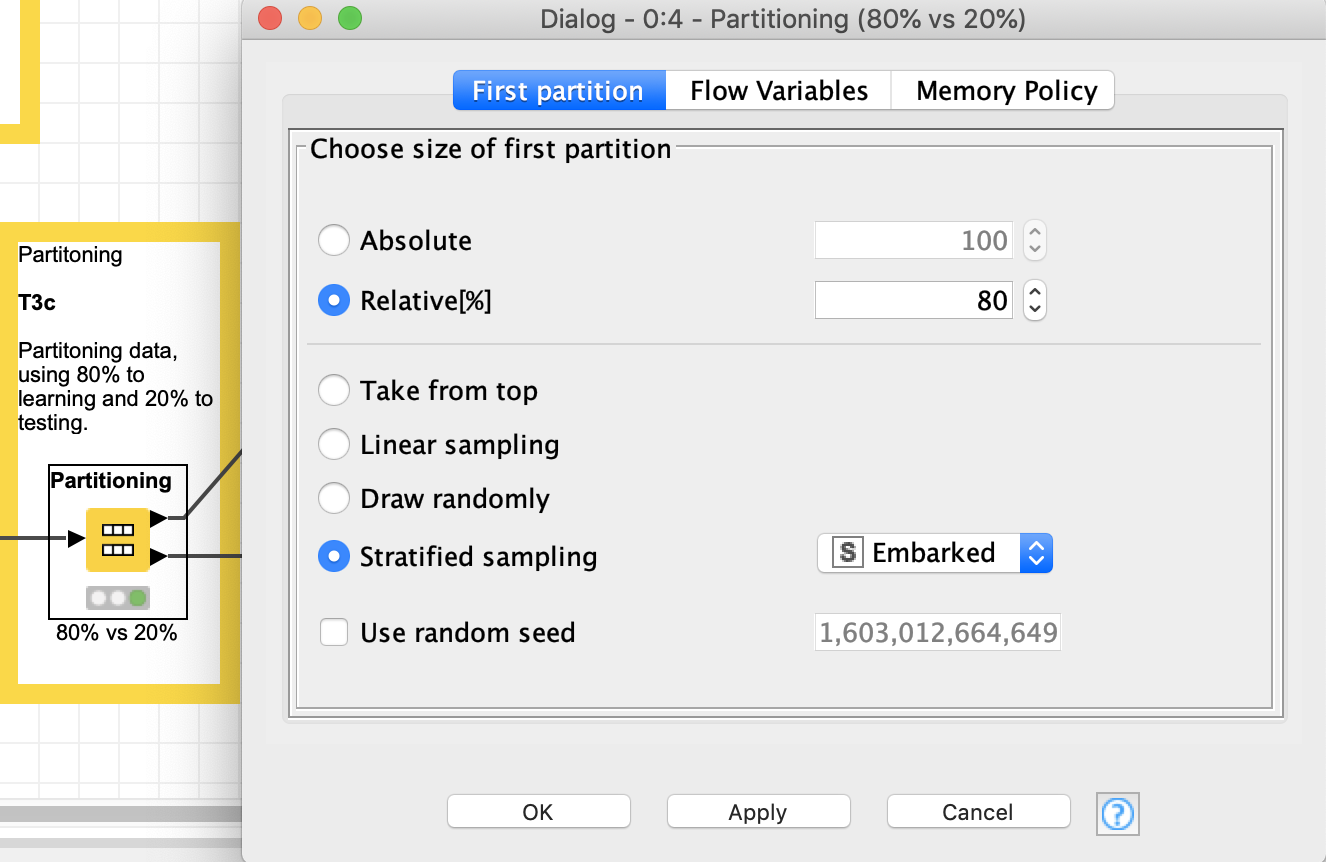
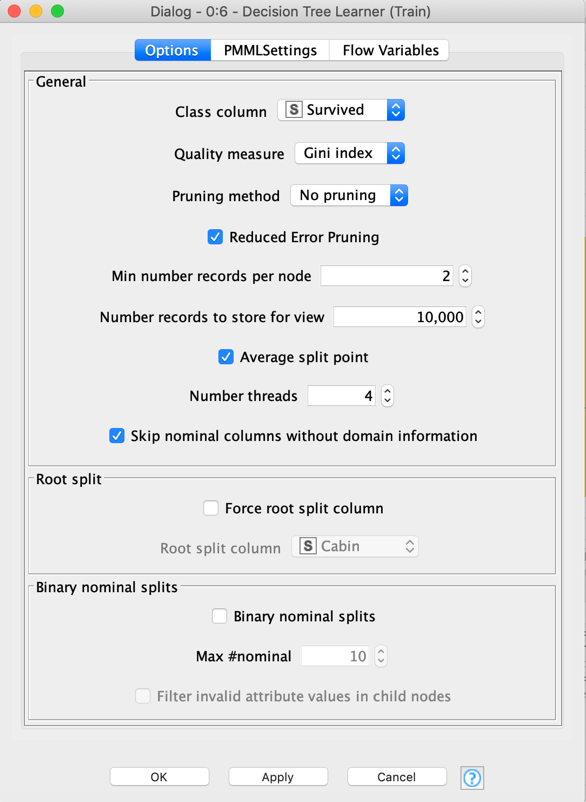
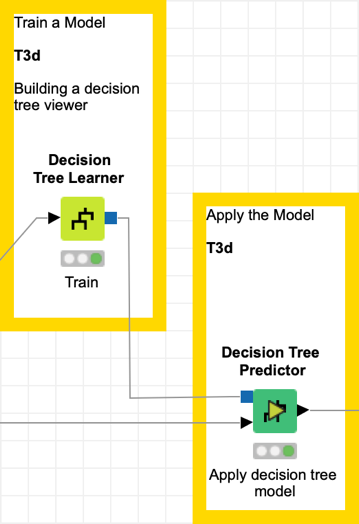


Figura 5 - Operação de partição

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

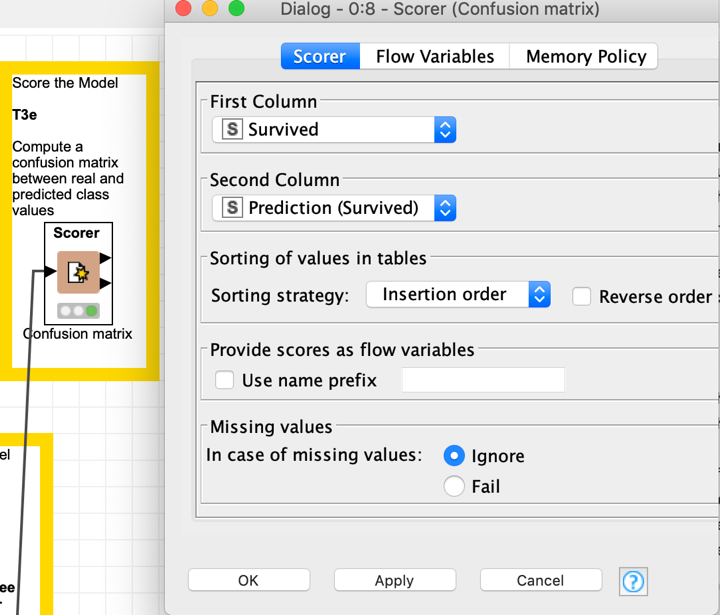


Uma imagem com texto

Descrição gerada automaticamente

Figura 6 - Previsões do modelo treinado

1. Avaliar a precisão do modelo utilizando o nodo *Scorer* e a respetiva matriz de confusão.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 8 - Avaliação da precisão do modelo

Figura 7 - Matriz confusão

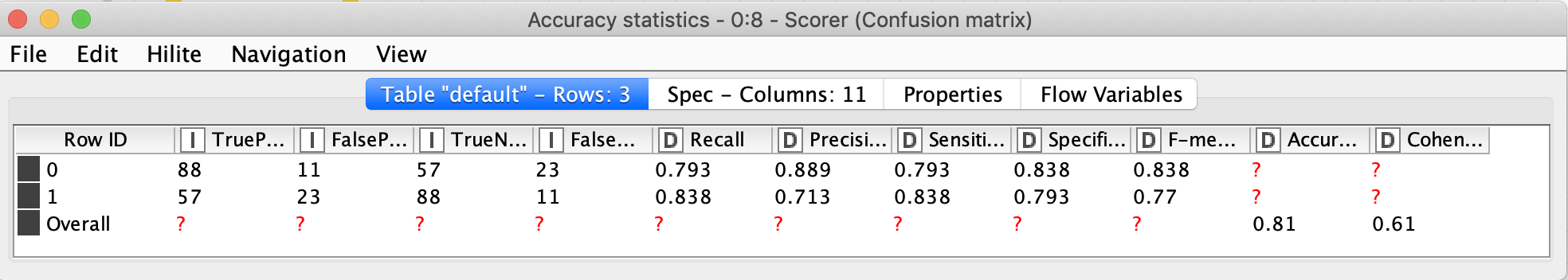


Figura 9 - Avaliação da precisão

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

De modo a elevar o nível de precisão, esta tarefa resume-se a manipular os diversos parâmetros no nodo *Decision Tree Learner,* para que seja visível a alteração da precisão obtida na tarefa T3e).

Nesta tarefa, é possível nos aproximar das melhores combinações possíveis. Observando a Tabela 1, obtemos um maior nível de precisão na manipulação número 2 e número 9, correspondentes às colunas “Survided” e “Sex”, respetivamente.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Manipulação | *Class column* | *Quality measure* | *Pruning method* | *Min num. records/node* | *Min num. records to store/view* | *Num. threads* | *Correct classified* | *Wrong classified* | *Accuracy (~)* | *Error (~)* |
| 1 | Survided | Gini index | No | 2 | 10.000 | 4 | 145 | 34 | 0.81 | 0.19 |
| 2 | **Survided** | **Gini index** | **No** | **4** | **5.000** | **2** | **151** | **28** | **0.84** | **0.16** |
| 3 | Survided | Gini index | MDL | 2 | 10.000 | 4 | 147 | 32 | 0.82 | 0.18 |
| 4 | Survided | Gini index | MDL | 4 | 5.000 | 2 | 147 | 32 | 0.82 | 0.18 |
| 5 | Survided | Gini ratio | No | 2 | 10.000 | 4 | 144 | 35 | 0.80 | 0.20 |
| 6 | Survided | Gini ratio | MDL | 3 | 2.000 | 4 | 144 | 35 | 0.80 | 0.20 |
| 7 | Sex | Gini ratio | MDL | 3 | 2.000 | 4 | 140 | 39 | 0.78 | 0.22 |
| 8 | Sex | Gini index | No | 2 | 10.000 | 4 | 137 | 42 | 0.76 | 0.24 |
| 9 | **Sex** | **Gini index** | **No** | **5** | **7.000** | **4** | **142** | **37** | **0.80** | **0.20** |

Tabela 1 - Combinações no nodo Decision Tree Learner