



UNIVERSIDADE DO MINHO
MESTRADO EM MATEMÁTICA E COMPUTAÇÃO

Sistemas Baseados em Similaridade

Ficha Prática Individual 1

Hugo Filipe de Sá Rocha (PG52250)

27 de setembro de 2023

Conteúdo

1	Concepção das soluções	3
1.1	Tarefa 1	3
1.2	Tarefa 2	3
1.3	Tarefa 3	4
1.3.1	Alínea a)	4
1.3.2	Alínea b)	4
1.3.3	Alínea c)	5
1.3.4	Alínea d)	5
1.3.5	Alínea e)	6
1.3.6	Workflow completo	7
1.4	Tarefa 4	7

Capítulo 1

Concepção das soluções

1.1 Tarefa 1

(Instalar a plataforma *Knime*)

1.2 Tarefa 2

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

Para efetuar a leitura correta dos dados fornecidos, em formato **CSV**, usei o nodo **CSV Reader** configurando-o com o caminho para o respetivo ficheiro, acabando por obter a seguinte tabela que representa o dataset:

► 1: File Table 📄 Flow Variables

Rows: 891 | Columns: 12

Table Statistics

#	Row...	PassengerId Number (integer)	Survived String	Pclass Number (integer)	Name String	Sex String	Age Number (double)	SibSp Number (integer)	Parch Number (integer)	Ticket String	Fare Number (double)	Cabin String	Embarked String
121	Row...	121	N	2	Hickman, Mr. St...	male	21	2	0	S.O.C. 14879	73.5	🚫	S
122	Row...	122	N	3	Moore, Mr. Leo...	male	🚫	0	0	A4. 54510	8.05	🚫	S
123	Row...	123	N	2	Nasser, Mr. Nic...	male	32.5	1	0	237736	30.071	🚫	C
124	Row...	124	Y	2	Webber, Miss. S...	female	32.5	0	0	27267	13	E101	S
125	Row...	125	N	1	White, Mr. Perci...	male	54	0	1	35281	77.287	D26	S
126	Row...	126	Y	3	Nicola-Yarred, ...	male	12	1	0	2651	11.242	🚫	C
127	Row...	127	N	3	McMahon, Mr. ...	male	🚫	0	0	370372	7.75	🚫	Q
128	Row...	128	Y	3	Madsen, Mr. Fri...	male	24	0	0	C 17369	7.142	🚫	S
129	Row...	129	Y	3	Peter, Miss. Anna	female	🚫	1	1	2668	22.358	F E69	C
130	Row...	130	N	3	Ekstrom, Mr. Jo...	male	45	0	0	347061	6.975	🚫	S
131	Row...	131	N	3	Drazenoic, Mr. J...	male	33	0	0	349241	7.896	🚫	C
132	Row...	132	N	3	Coelho, Mr. Do...	male	20	0	0	SOTON/O.Q. 31...	7.05	🚫	S
133	Row...	133	N	3	Robins, Mrs. AL...	female	47	1	0	A/5. 3337	14.5	🚫	S
134	Row...	134	Y	2	Weisz, Mrs. Leo...	female	29	1	0	228414	26	🚫	S
135	Row...	135	N	2	Sobey, Mr. Sam...	male	25	0	0	C.A. 29178	13	🚫	S
136	Row...	136	N	2	Richard, Mr. Emi...	male	23	0	0	SC/PARIS 2133	15.046	🚫	C
137	Row...	137	Y	1	Newsom, Miss. ...	female	19	0	2	11752	26.283	D47	S

1.3 Tarefa 3

1.3.1 Alínea a)

- Filtrar (i.e., remover) as colunas “*Age*”, “*Ticket*” e “*Cabin*”.

Para remover as respectivas colunas do dataset, utilizei o nodo **Column Filter** onde, nas configurações do mesmo, selecionei as colunas que queria excluir, neste caso: “*Age*”, “*Ticket*” e “*Cabin*”, obtendo o seguinte dataset:

► 1: Filtered table

📄 Flow Variables

Rows: 891 | Columns: 9

Table

Statistics

#	Row...	Passengerid <small>Number (integer)</small>	Survived <small>String</small>	Pclass <small>Number (integer)</small>	Name <small>String</small>	Sex <small>String</small>	SibSp <small>Number (integer)</small>	Parch <small>Number (integer)</small>	Fare <small>Number (double)</small>	Embarked <small>String</small>
121	Row...	121	N	2	Hickman, Mr. Stanley ...	male	2	0	73.5	S
122	Row...	122	N	3	Moore, Mr. Leonard C...	male	0	0	8.05	S
123	Row...	123	N	2	Nasser, Mr. Nicholas	male	1	0	30.071	C
124	Row...	124	Y	2	Webber, Miss. Susan	female	0	0	13	S
125	Row...	125	N	1	White, Mr. Percival Wa...	male	0	1	77.287	S
126	Row...	126	Y	3	Nicola-Yarred, Master. ...	male	1	0	11.242	C
127	Row...	127	N	3	McMahon, Mr. Martin	male	0	0	7.75	Q
128	Row...	128	Y	3	Madsen, Mr. Fridtjof A...	male	0	0	7.142	S
129	Row...	129	Y	3	Peter, Miss. Anna	female	1	1	22.358	C
130	Row...	130	N	3	Ekstrom, Mr. Johan	male	0	0	6.975	S
131	Row...	131	N	3	Drazenoic, Mr. Jozef	male	0	0	7.896	C
132	Row...	132	N	3	Coelho, Mr. Domingos ...	male	0	0	7.05	S
133	Row...	133	N	3	Robins, Mrs. Alexande...	female	1	0	14.5	S
134	Row...	134	Y	2	Weisz, Mrs. Leopold (...)	female	1	0	26	S
135	Row...	135	N	2	Sobey, Mr. Samuel Ja...	male	0	0	13	S
136	Row...	136	N	2	Richard, Mr. Emile	male	0	0	15.046	C
137	Row...	137	Y	1	Newsom, Miss. Helen ...	female	0	2	26.283	S

1.3.2 Alínea b)

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

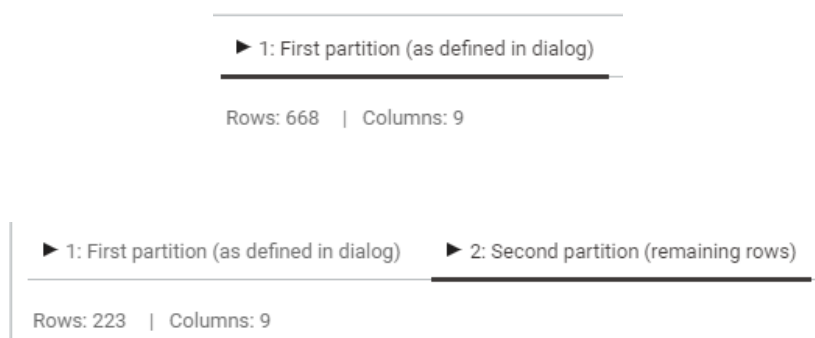
A coluna “*Survived*” já se encontrava no tipo *String* pelo que não foi necessário realizar o *cast*.

Survived <small>String</small>
N
N
N
Y

1.3.3 Alínea c)

- Particionar os dados, de forma aleatória, utilizando 75% para aprendizagem e 25% para teste.

Para dividir os dados para aprendizagem e teste, utilizei o nodo **Partitioning** onde especifiquei que a primeira partição correspondia a 75% dos dados que seriam utilizados para aprendizagem do nosso modelo e os restantes 25% para teste. As duas saídas deste nodo correspondem a estas duas componentes resultantes da partição dos dados que, também nas configurações do nodo, especifiquei para que fosse realizada de forma aleatória.

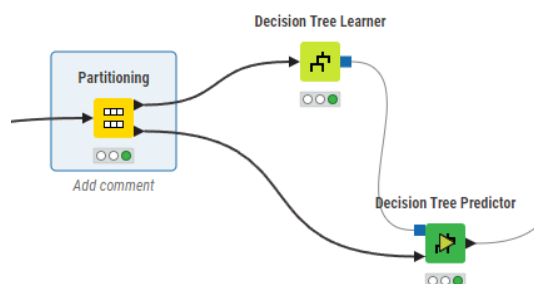


1.3.4 Alínea d)

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

Nas configurações do nodo **Decision Tree Learner** comecei por especificar que a coluna correspondente à classe do nosso modelo (aquilo que queremos prever) era a coluna **"Survived"**. Depois, estabeleci uma conexão entre o nodo **Partitioning**, mais concretamente, a saída correspondente aos dados de aprendizagem e o nodo **Decision Tree Learner** de forma a treinar a Árvore de Decisão.

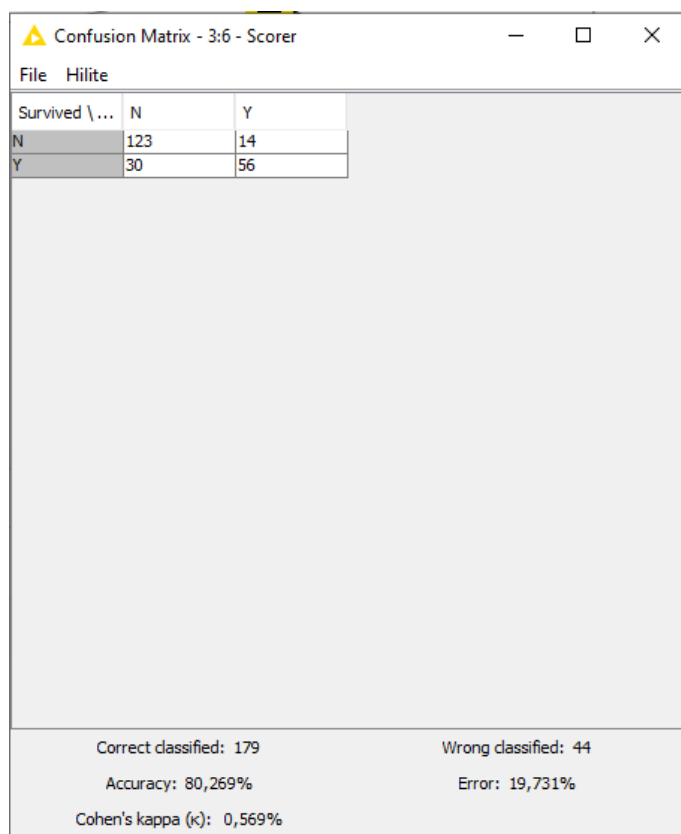
Quanto ao nodo **Decision Tree Predictor**, liguei-o à saída correspondente aos dados de teste do nodo **Partitioning** de forma a obter previsões sobre os mesmos, utilizando o modelo de Árvore de Decisão já treinado e, para isso, criei também uma ligação entre o nodo **Decision Tree Learner** e o nodo **Decision Tree Predictor**.



1.3.5 Alínea e)

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

Para avaliar a precisão do modelo, utilizei o nodo **Scorer** e liguei-o ao nodo **Decision Tree Predictor**. Ao fazer "Open View" sobre o nodo **Scorer**, conseguimos obter a matriz de confusão bem como a precisão do modelo em percentagem e em valores absolutos.



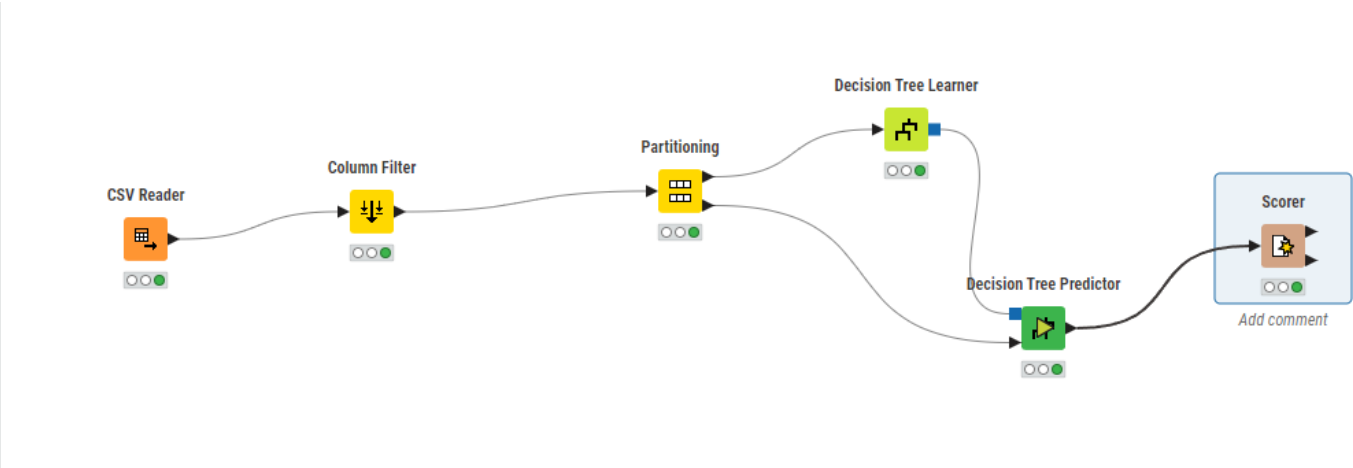
The screenshot shows a window titled "Confusion Matrix - 3:6 - Scorer". It contains a confusion matrix for the "Survived" variable, with actual values "N" and "Y" and predicted values "N" and "Y". The matrix shows 123 correct classifications for "N", 14 misclassifications for "N", 30 misclassifications for "Y", and 56 correct classifications for "Y". Below the matrix, summary statistics are provided: 179 correct classifications, 44 wrong classifications, an accuracy of 80.269%, an error rate of 19.731%, and a Cohen's kappa of 0.569%.

Survived \ ...	N	Y
N	123	14
Y	30	56

Correct classified: 179	Wrong classified: 44
Accuracy: 80,269%	Error: 19,731%
Cohen's kappa (κ): 0,569%	

1.3.6 Workflow completo

O *workflow* completo está representado abaixo:



1.4 Tarefa 4

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

Comecei por alterar os parâmetros *"Quality Measure"* e *"Pruning Method"* (um de cada vez), de *"Gini Index"* para *"Gain Ratio"* e de *"No Pruning"* para *"MDL"* e acabou por não afetar a precisão do algoritmo.

Relativamente aos parâmetros numéricos acabei por testar diferentes combinações de valores numéricos.

Algumas delas foram:

- *Min number records per node:* 15
- *Number records to store for view:* 10 000
- *Number threads:* 4

Percentagem de acerto do algoritmo: em torno de **80%**.

#	Row...	N Number (integer)	Y Number (integer)
1	N	129	11
2	Y	35	48

- *Min number records per node:* 300
- *Number records to store for view:* 10 000
- *Number threads:* 4

Percentagem de acerto do algoritmo acabou por descer significativamente para cerca de **65%**. De notar que, ao contrário da combinação acima, aqui o número de falsos negativos é maior que o número de falsos positivos.

#	Row...	N Number (integer)	Y Number (integer)
1	N	87	45
2	Y	30	61

- *Min number records per node:* 500
- *Number records to store for view:* 10 000
- *Number threads:* 4

Percentagem de acerto do algoritmo: cerca de **61%**. Destaque para o facto que, todos os negativos previstos pelo algoritmo foram acertados. No entanto, todos os positivos previstos foram errados. Ou seja, não temos falsos negativos, no entanto, todos os positivos são falsos o que acaba por fazer com que a precisão do algoritmo seja má.

#	Row...	N Number (integer)	Y Number (integer)
1	N	138	0
2	Y	85	0

Procurei também aumentar e diminuir os parâmetros *Number records to store for view* e *Number threads* mas não detetei alterações significativas naquilo que foi a eficácia e respetiva matriz de confusão do algoritmo comparativamente à primeira combinação aqui apresentada.