

CISC 333

Assignment 2: Titanic Name Analysis

Name: Sam Ruscica

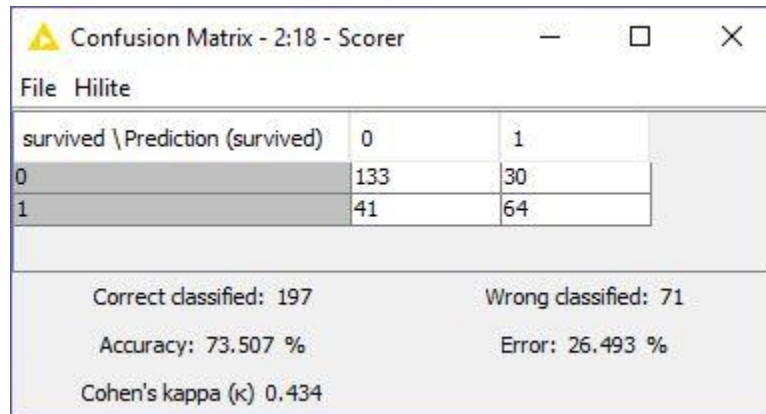
Student Number: 10148585

Course: CISC 333

Date: September 29, 2017

Procedure & Results

This analysis was conducted to determine if a person's title correlates to other data. The model was made to see if their title could be used to predict if they survived the tragedy. This was done by isolating their title from their name and feeding them into a decision tree. To see the effect of adding in the titles category, a model was made without it. Below are the results.



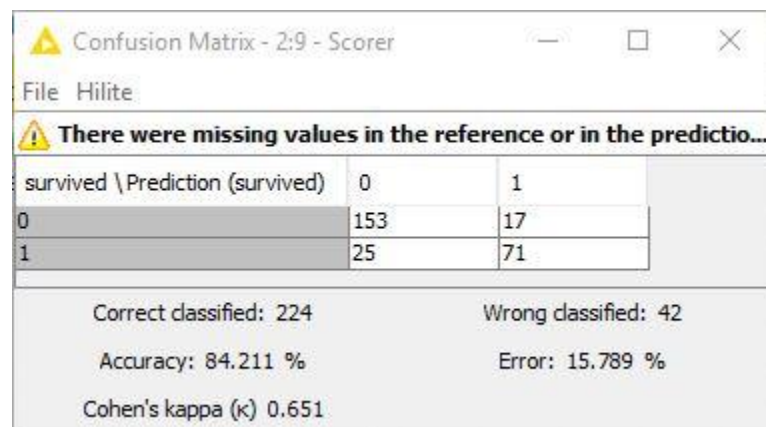
Confusion Matrix - 2:18 - Scorer

File Hilite

survived \ Prediction (survived)	0	1
0	133	30
1	41	64

Correct classified: 197 Wrong classified: 71
Accuracy: 73.507 % Error: 26.493 %
Cohen's kappa (κ) 0.434

Figure 1: Scorer no Titles



Confusion Matrix - 2:9 - Scorer

File Hilite

There were missing values in the reference or in the prediction...

survived \ Prediction (survived)	0	1
0	153	17
1	25	71

Correct classified: 224 Wrong classified: 42
Accuracy: 84.211 % Error: 15.789 %
Cohen's kappa (κ) 0.651

Figure 2: Scorer with Titles

Table 1: Model Accuracy

Group	Accuracy %					Average
	Test 1	Test 2	Test 3	Test 4	Test 5	
No Titles	73.5	83.2	79.1	80.6	78.7	79.02
Titles	84.2	82.6	80.5	78.7	82.4	81.68

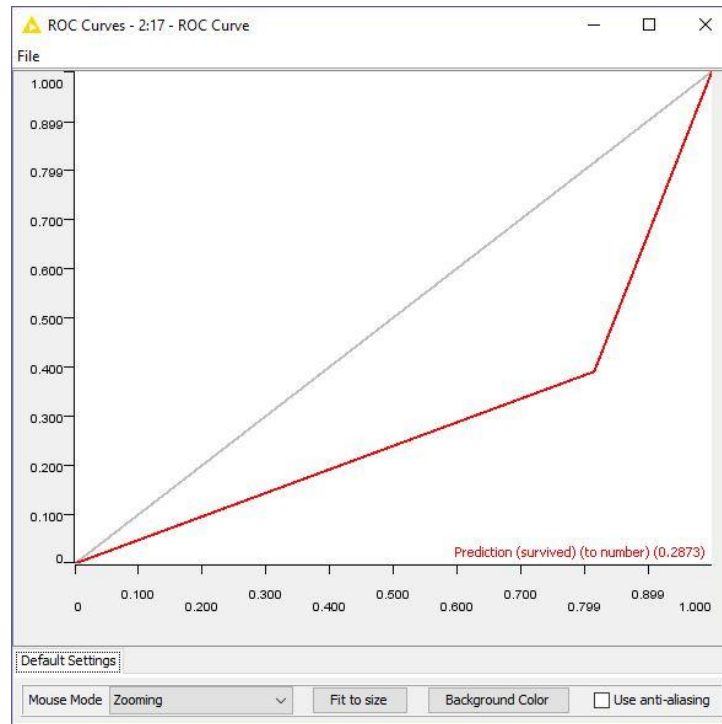


Figure 3: ROC Curve no Titles

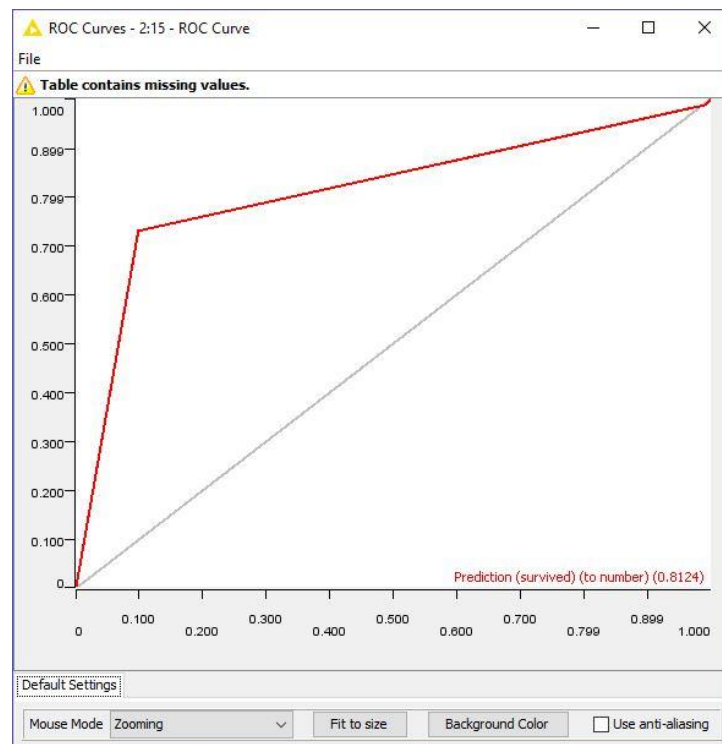



Figure 4: ROC Curve with Titles

Analysis

The results from the no title model showed that it is good at predicting survivors. This can be seen in the prediction accuracy, confusion matrix, and the ROS curve. The prediction accuracy averaged out to 79%. The errors for the confusion matrix were not balanced though. The ROS curve showed that it did not sit on the grey line. The grey line represents the worst possible model. The addition of the titles improved the average accuracy by nearly 3%. The model still does not have balanced errors for the confusion matrix. After seeing the constant unbalanced errors, a new model was made using only the titles category to predict survivors. Below are the results.

Table 2: Model Accuracy Only Titles


Group	Accuracy %					Average
	Test 1	Test 2	Test 3	Test 4	Test 5	
Only Titles	80.5	78.9	78.1	77.8	80.3	79.1



Confusion Matrix - 2:9 - Scorer

File

Hilite



There were missing values in the reference or in the prediction cla...

survived \ Prediction (survived)	0	1	
0	132	25	
1	23	86	

Correct classified: 218

Wrong classified: 48

Accuracy: 81.955 %

Error: 18.045 %

Cohen's kappa (κ) 0.628

Figure 5: Scorer Results Only Titles

The only titles model still performed better than the no title model. Additionally, the errors in the confusion matrix have drastically improved by balancing out. This is most likely a result of filtering out the extra data. The titles category is simply a cleaned-up version of the name category. Since, the name category was used in the other models, it still conveyed the similar trend.

KNIME Nodes

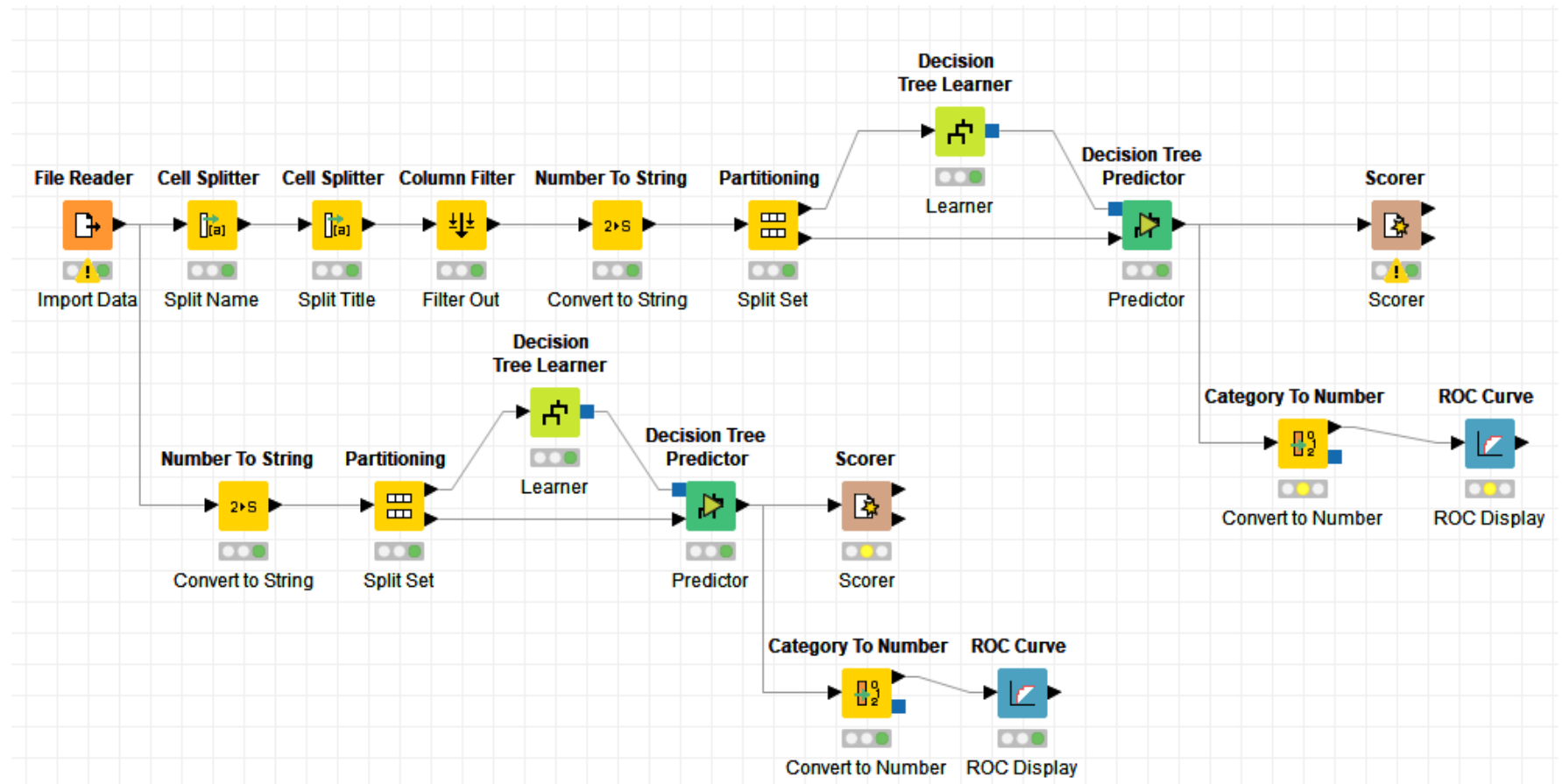


Figure 6: Node Layout in KNIME