CISC 333

Assignment 5: Determining Best Titanic Model

Name: Sam Ruscica

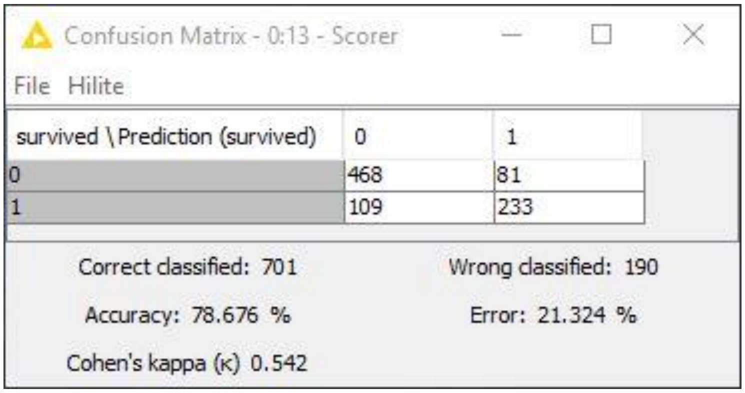
Student Number: 10148585

Course: CISC 333

Date: October 20, 2017

Procedure & Results

The analysis was conducted to examine the effectiveness of two more model types; SVM and Random Forest. Firstly, the data was manipulated to filter out unwanted data, missing values, normalized, and placed into categories. This modified dataset can now be partitioned into training and test data for the different models. The models both use learner nodes and predictor nodes to make their corresponding models. Lastly, a scorer is used to see how accurate the model is. Below are the results.



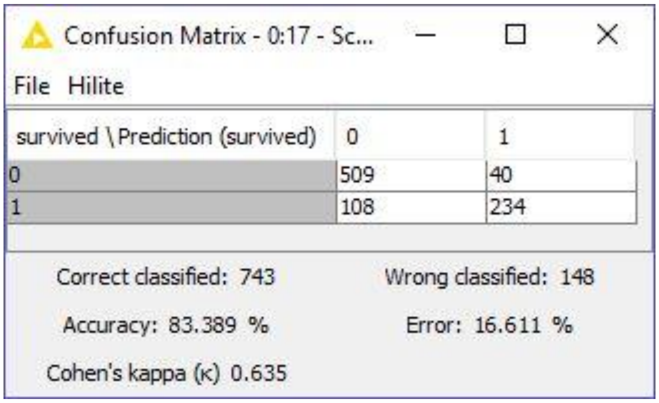
survived \ Prediction (survived)	0	1
0	468	81
1	109	233

Correct classified: 701 Wrong classified: 190

Accuracy: 78.676 % Error: 21.324 %

Cohen's kappa (κ) 0.542

Figure 1: SVM Confusion Matrix



survived \ Prediction (survived)	0	1
0	509	40
1	108	234

Correct classified: 743 Wrong classified: 148

Accuracy: 83.389 % Error: 16.611 %

Cohen's kappa (κ) 0.635

Figure 2: Random Forest Confusion Matrix

Analysis

Throughout the past assignments, numerous models were used to try and predict survival. Below is a table showing all the different models used and their accuracies.

Table 1: Model Accuracy Comparison

Model Type	Accuracy
Decision Tree No Titles	79.02%
Decision Tree with Titles	81.68%
Decision Tree Only Titles	79.10%
Normalization	80.60%
Supervised Neural Network	76.32%
SVM	78.70%
Random Forest	83.39%

Based on the table above, the best choice for a model is the random forest. This model had the highest consistent accuracy overall. If this was a larger dataset, the random forest would take far longer to run than the other models. This could prove to be an issue as the results of the analysis might be readily needed. If this were the case, I would recommend using the decision tree with titles. This was the second most accurate model and it ran faster. This was the best decision tree as it had cleaned up the title category so it could be used cleaner.

KNIME Nodes

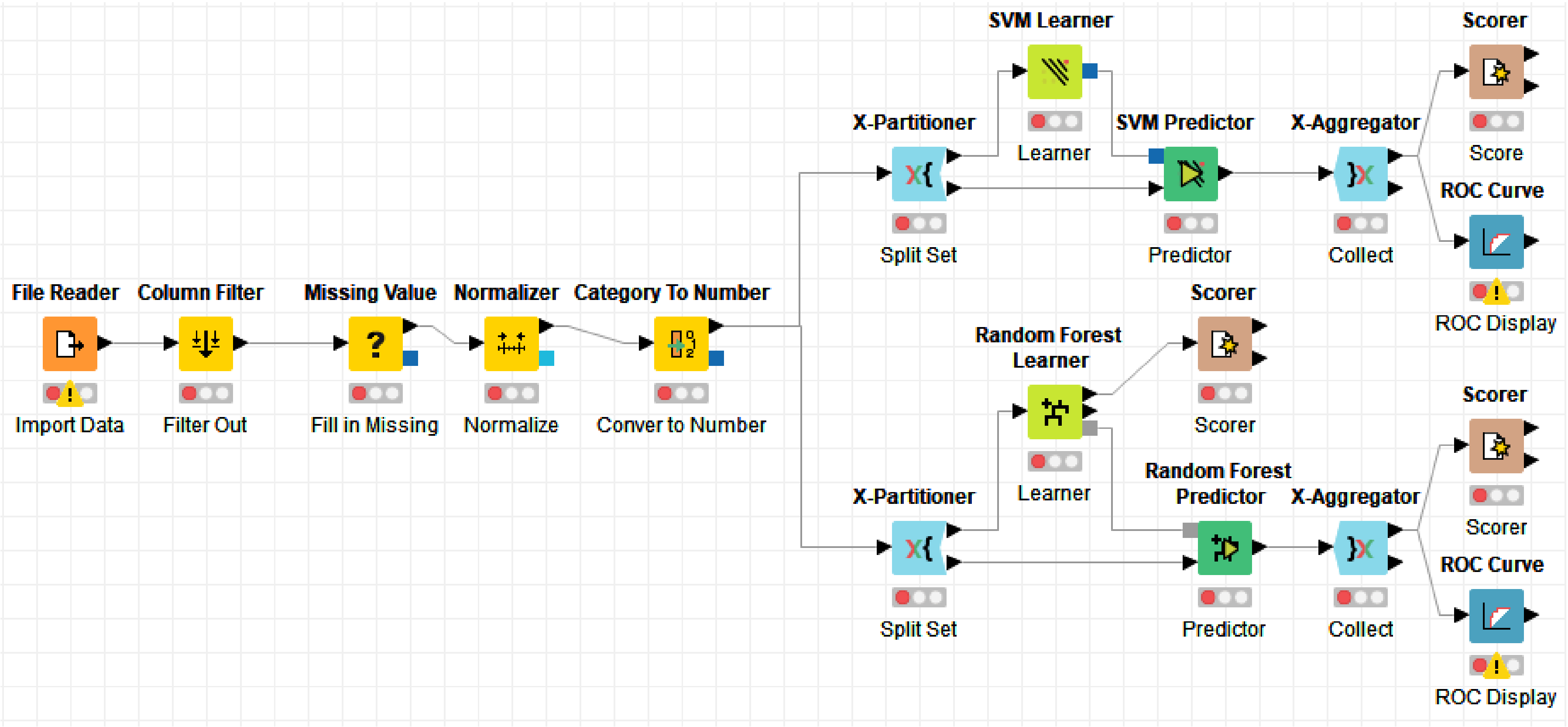


Figure 3: KNIME Nodes Used for Analysis