Titanic Survival Prediction Project Report

Summary

This project employs the RapidMiner Studio platform to predict survival rates of passengers aboard the Titanic using machine learning techniques. A Random Forest algorithm was utilized for its high-performance metrics and ability to handle complex datasets.

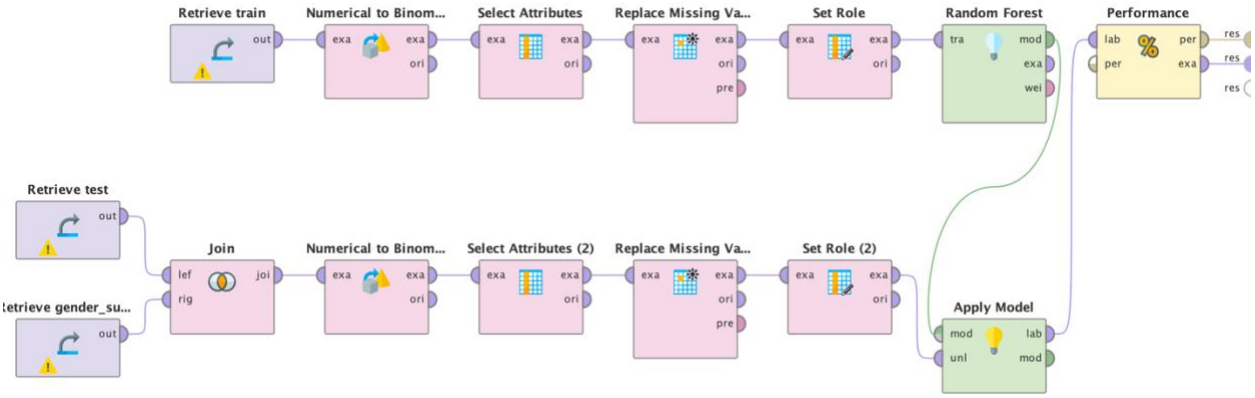
Project Objectives

- To predict the "Survived" status of passengers using a trained Random Forest model.
- To handle data preprocessing including cleaning, transformation, and missing value imputation.
- To merge model predictions with the passenger identification data for submission.

Methodology

Data Preprocessing

1. **Data Importation:** The Train.csv dataset was imported into RapidMiner Studio.
2. **Cleaning and Transformation:**
   - Unnecessary columns were removed.
   - Categorical variables were converted to numerical format, specifically to binomial values.
   - Missing values were replaced.
   - Relevant attributes were selected for model training.
3. **Label Assignment:** The 'Survived' column was established as the label for prediction.



Model Training

1. **Random Forest Algorithm:** The cleaned data was used to train a Random Forest model.
2. **Model Evaluation:** The "Performance (Classification)" operator assessed the model's accuracy.

Prediction and Validation

1. **Prediction Generation:** The trained model predicted the 'Survived' status on both the training set and the unseen test data.
2. **Data Merging:** Predictions were joined with the 'PassengerId' from the test set and the gender submission file.

Results

- The model achieved an accuracy of 92.58%.
- Precision and recall metrics were calculated, with a class precision of 93.36% for 'Survived = True' predictions and a class recall of 88.16%.
- A confusion matrix and a sample of prediction results were provided to illustrate the model's performance.

accuracy: 92.58%

	true false	true true	class precision
pred. false	253	18	93.36%
pred. true	13	134	91.16%
class recall	95.11%	88.16%	

Conclusion

The project successfully applied a Random Forest classifier within RapidMiner Studio to predict survival on the Titanic with high accuracy. This showcases the potential of machine learning in historical data analysis and prediction.

Row No.	Survived	prediction(...	confidence(...	confidence(...	Pclass	Sex	Age	SibSp	Parch
1	false	false	0.897	0.103	3	male	34.500	0	0
2	true	false	0.857	0.143	3	female	47	1	0
3	false	false	0.897	0.103	2	male	62	0	0
4	false	false	0.897	0.103	3	male	27	0	0
5	true	true	0.409	0.591	3	female	22	1	1
6	false	false	0.897	0.103	3	male	14	0	0
7	true	true	0.409	0.591	3	female	30	0	0
8	false	false	0.897	0.103	2	male	26	1	1
9	true	true	0.409	0.591	3	female	18	0	0
10	false	false	0.897	0.103	3	male	21	2	0
11	false	false	0.897	0.103	3	male	30.273	0	0
12	false	false	1	0	1	male	46	0	0
13	true	true	0	1	1	female	23	1	0
14	false	false	0.897	0.103	2	male	63	1	0
15	true	true	0	1	1	female	47	1	0
16	true	true	0.090	0.910	2	female	24	1	0
17	false	false	0.897	0.103	2	male	35	0	0
18	false	false	0.897	0.103	3	male	21	0	0
19	true	true	0.409	0.591	3	female	27	1	0