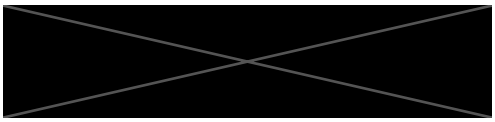
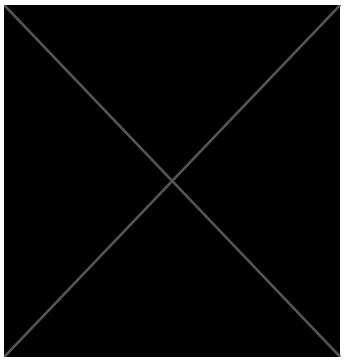


Predictive Analytics
MIS 5113-Introduction to Business Analytics



A report submitted in the fulfillment of
the requirement for the second group assignment.
March 27, 2023

I. Business Overview

The provided dataset is related to the assessment of credit risk for loan applicants. Each row in the dataset represents an individual loan application, containing attributes such as credit history, savings, account balance, employment status, and personal information of the loan applicant. Analyzing the dataset, several questions can be explored, such as identifying patterns or clusters within the dataset to improve credit standing, determining the distribution of credit standing across the dataset, and finding any correlation between numerical variables like saving account balance, months of account opened, and residence time with credit standing.

II. Data Analysis

The dataset contains 15 variables, including credit history, checking account, gender, and personal status. The variable "Credit Standing" is the target variable, which is a categorical variable with two possible values - "Good" and "Bad." It is a dependent variable, and our objective is to build a predictive model that classifies loan applicants as either "Good" or "Bad" standing, making it a supervised learning problem. All variables, except "Credit Standing," are independent variables that help to predict the target variable. The dataset comprises categorical, numerical, and binary variables, with 425 data records.

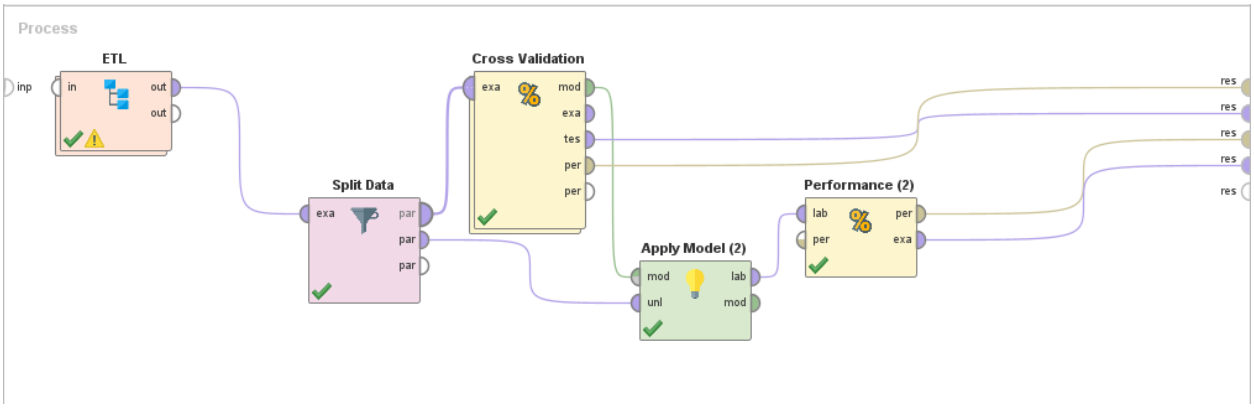
III. Data Preprocessing

During the data preprocessing stage, variable names were changed for better understanding, using the "Rename" operator in RapidMiner. For example, "Age Subtracted 1 from Original Age Variable" and "Months Acct" were renamed to "Age" and "Account Duration," respectively. One of the variables, "Gender," had missing values, and it was not included in the analysis due to its biased nature, as it is a protected characteristic under anti-discrimination law. Furthermore, some variables like "Months Acct" and "Residence Time" were transformed using the normalize operator through the z-score. All variables, except "Gender" and "Telephone," were included in the analysis.

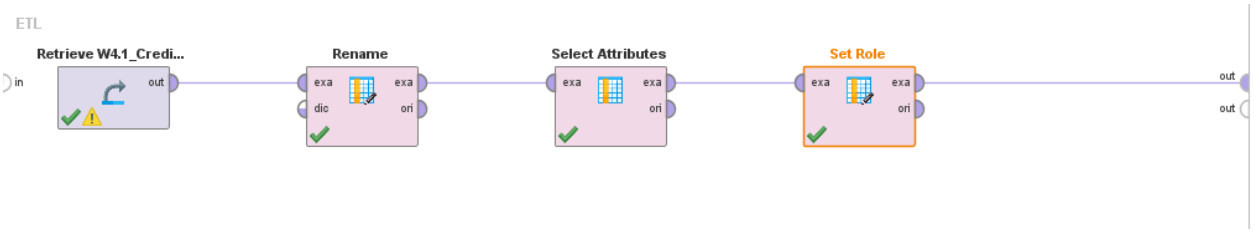
IV. Model Evaluation: Model Building & Testing

A. Decision Tree

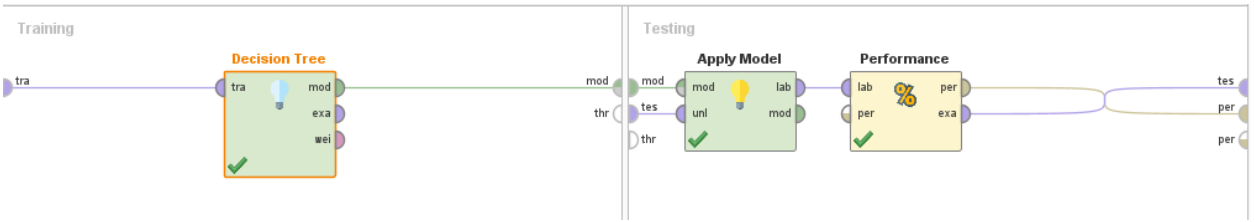
Process



ETL



Cross Validation



Training

accuracy: 69.24% +/- 5.55% (micro average: 69.23%)

	true Good	true Bad	class precision
pred. Good	139	60	69.85%
pred. Bad	64	140	68.63%
class recall	68.47%	70.00%	

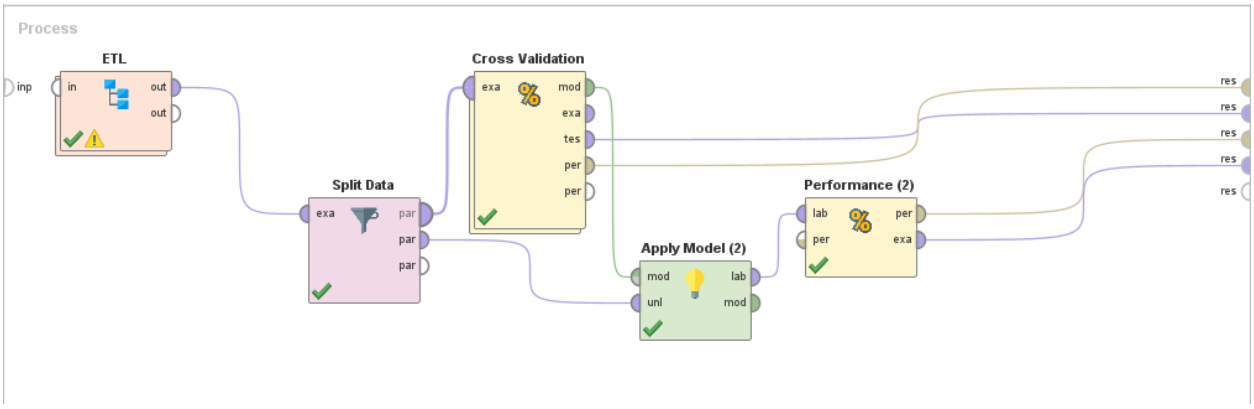
Learning

accuracy: 63.64%

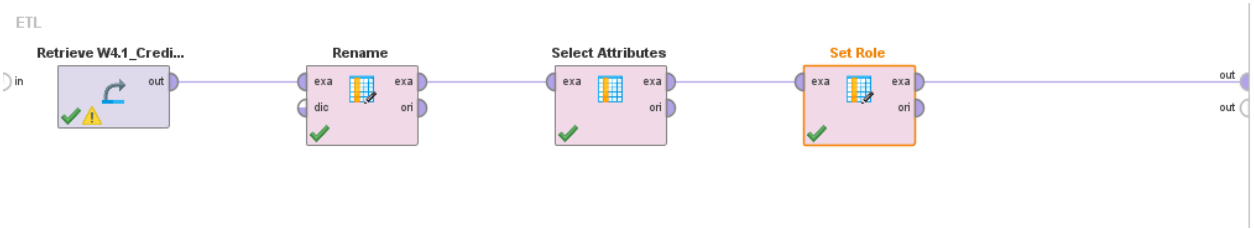
	true Good	true Bad	class precision
pred. Good	8	5	61.54%
pred. Bad	3	6	66.67%
class recall	72.73%	54.55%	

B. Naïve Bayes

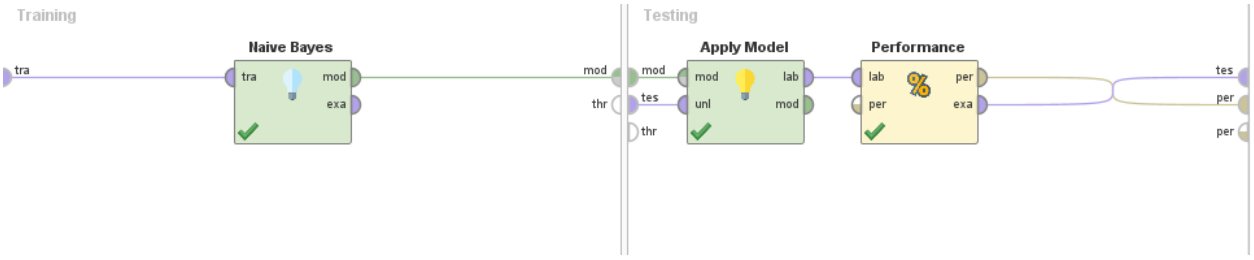
Process



ETL



Cross Validation



Training

accuracy: 69.77% +/- 7.31% (micro average: 69.73%)

	true Good	true Bad	class precision
pred. Good	137	56	70.98%
pred. Bad	66	144	68.57%
class recall	67.49%	72.00%	

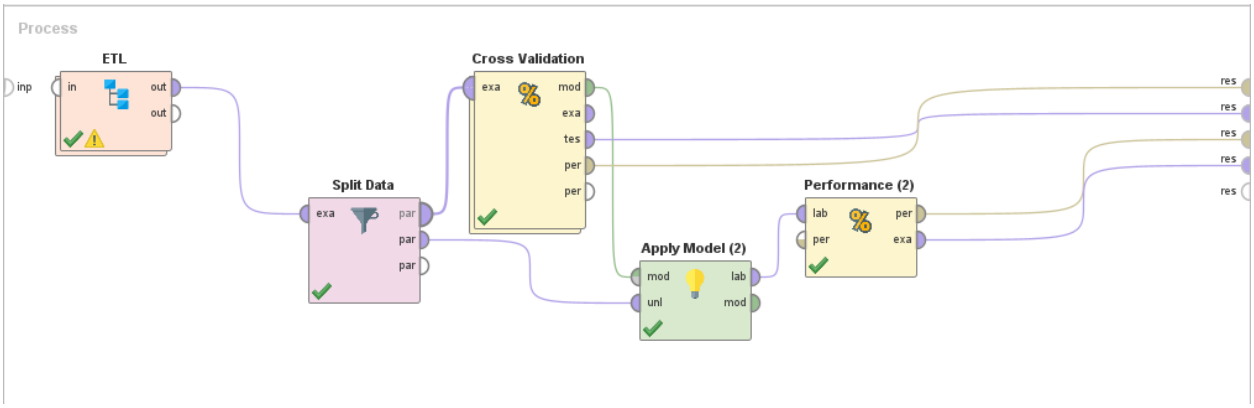
Learning

accuracy: 50.00%

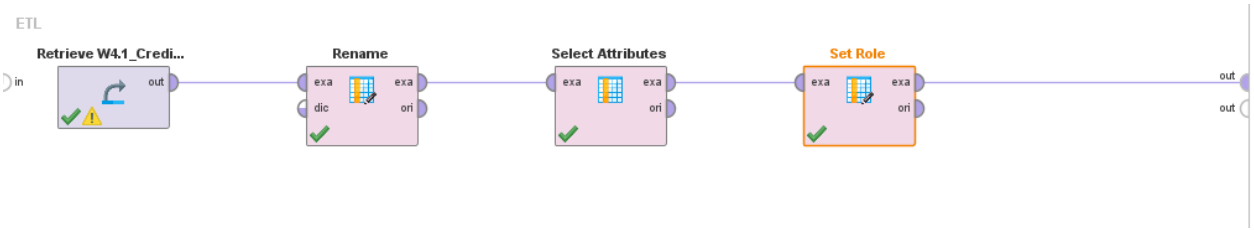
	true Good	true Bad	class precision
pred. Good	5	5	50.00%
pred. Bad	6	6	50.00%
class recall	45.45%	54.55%	

C. Random Forest

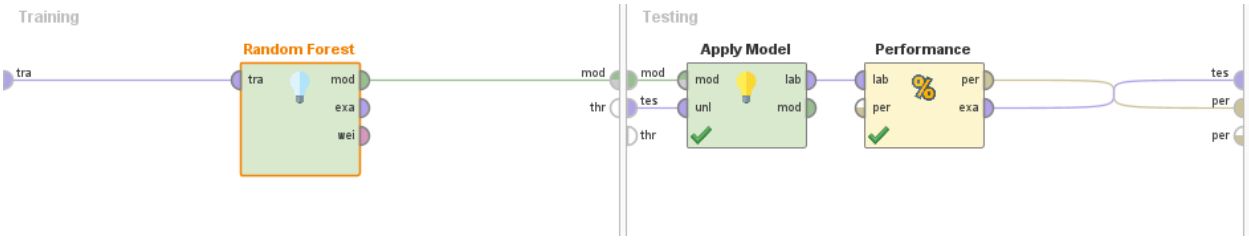
Process



ETL



Cross Validation



Training

accuracy: 70.74% +/- 5.39% (micro average: 70.72%)

	true Good	true Bad	class precision
pred. Good	143	58	71.14%
pred. Bad	60	142	70.30%
class recall	70.44%	71.00%	

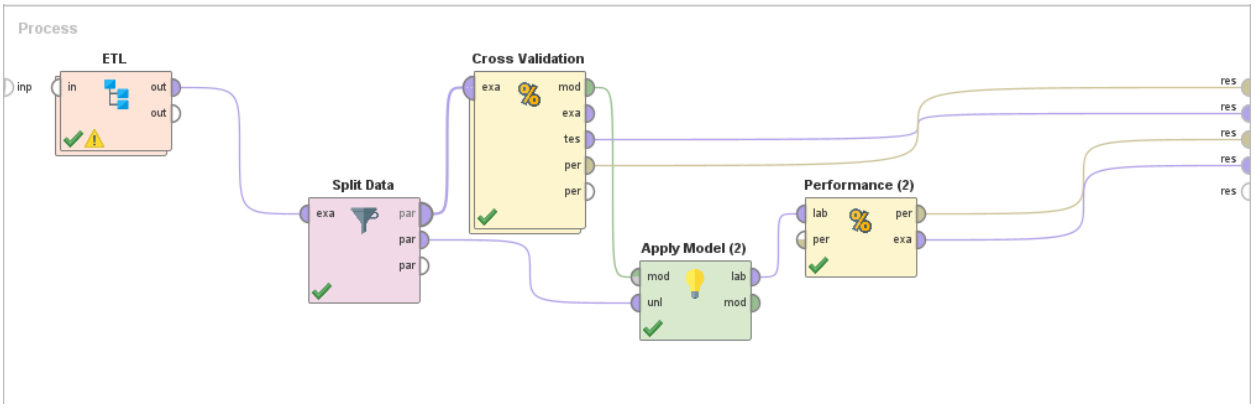
Learning

accuracy: 68.18%

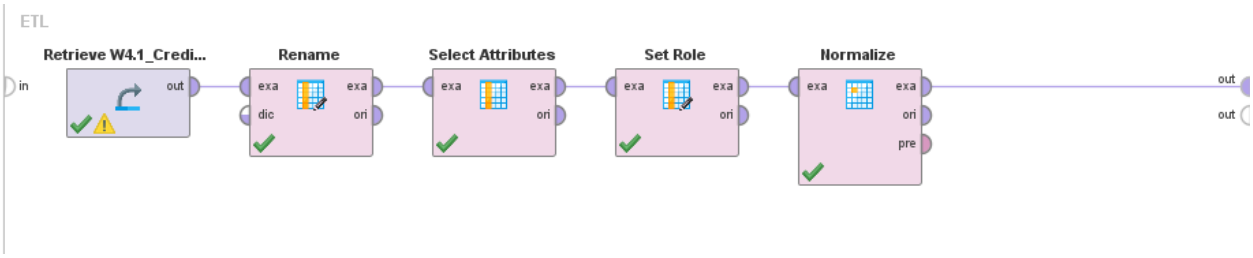
	true Good	true Bad	class precision
pred. Good	8	4	66.67%
pred. Bad	3	7	70.00%
class recall	72.73%	63.64%	

D. KNN

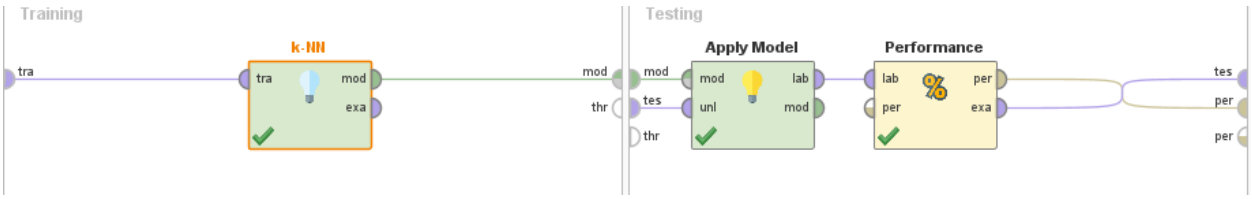
Process



ETL



Cross validation



Training

accuracy: 62.59% +/- 9.70% (micro average: 62.53%)

	true Good	true Bad	class precision
pred. Good	131	79	62.38%
pred. Bad	72	121	62.69%
class recall	64.53%	60.50%	

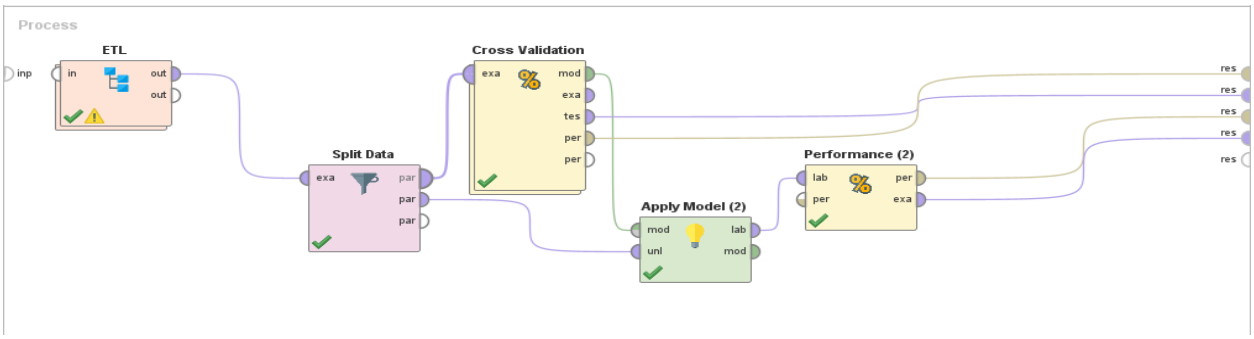
Learning

accuracy: 45.45%

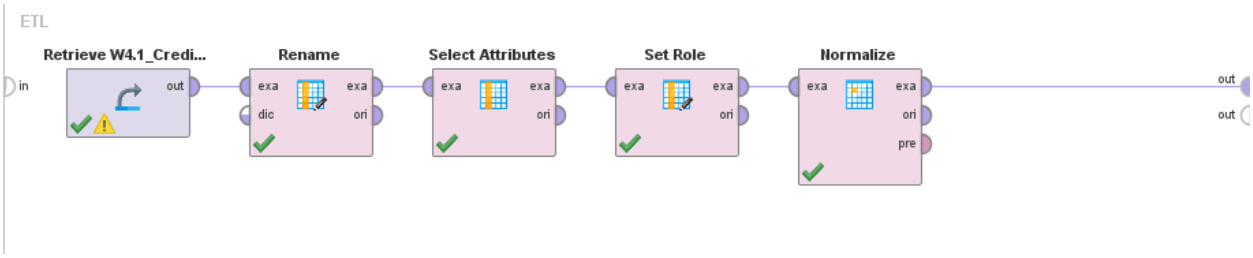
	true Good	true Bad	class precision
pred. Good	8	9	47.06%
pred. Bad	3	2	40.00%
class recall	72.73%	18.18%	

E. Logistic Regression

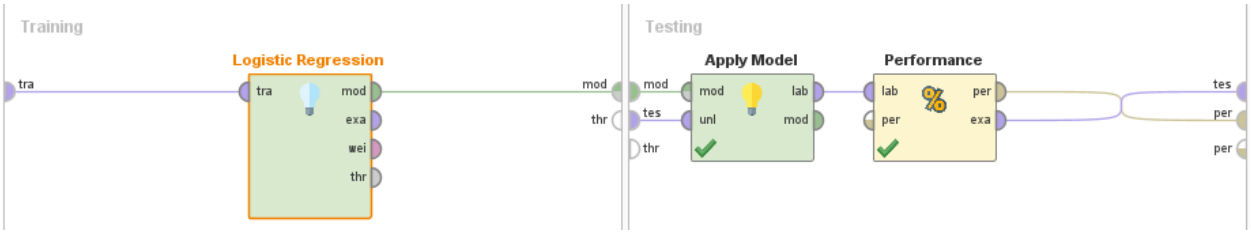
Process



ETL



Cross validation



Training

accuracy: 70.28% +/- 9.91% (micro average: 70.22%)

	true Good	true Bad	class precision
pred. Good	143	60	70.44%
pred. Bad	60	140	70.00%
class recall	70.44%	70.00%	

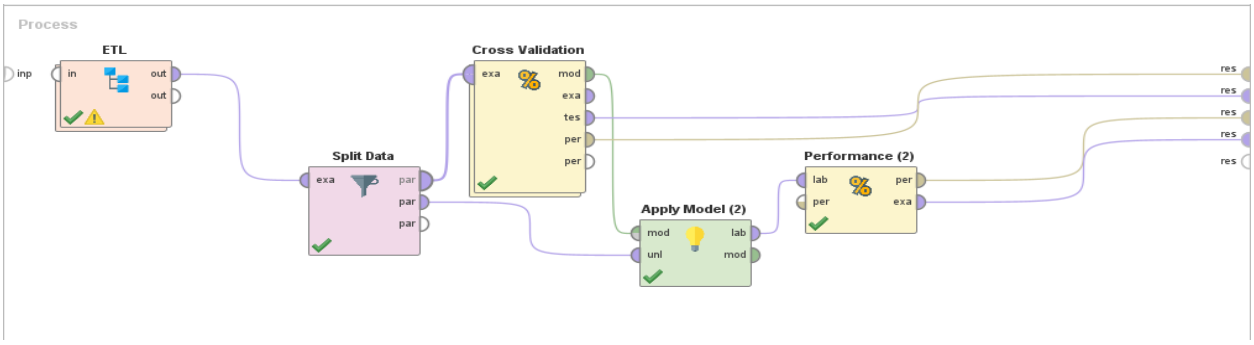
Learning

accuracy: 68.18%

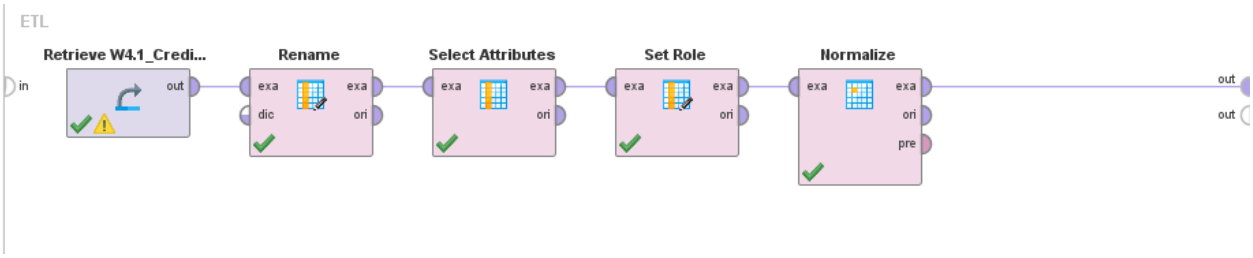
	true Good	true Bad	class precision
pred. Good	7	3	70.00%
pred. Bad	4	8	66.67%
class recall	63.64%	72.73%	

F. GLM

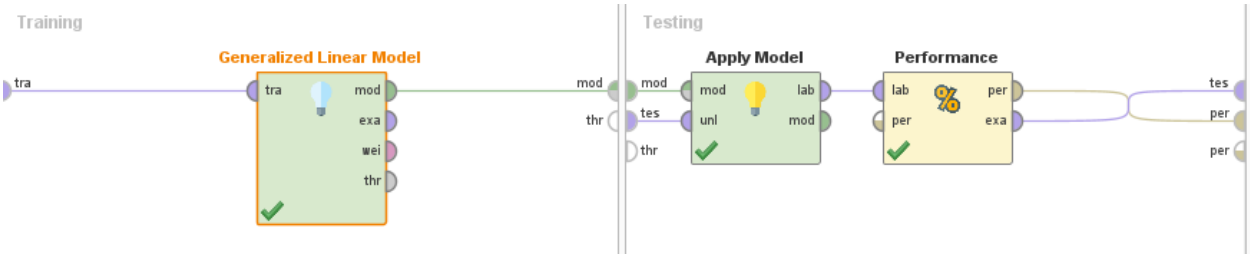
Process



ETL



Cross validation



Training

accuracy: 70.52% +/- 7.95% (micro average: 70.47%)

	true Good	true Bad	class precision
pred. Good	138	54	71.88%
pred. Bad	65	146	69.19%
class recall	67.98%	73.00%	

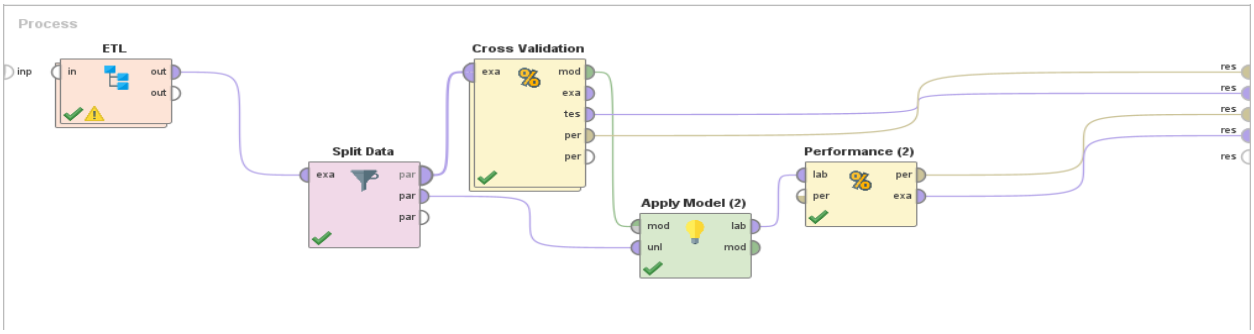
Learning

accuracy: 63.64%

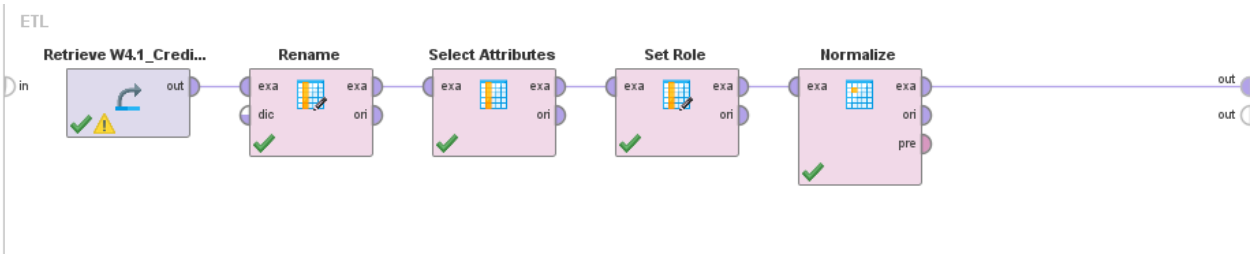
	true Good	true Bad	class precision
pred. Good	6	3	66.67%
pred. Bad	5	8	61.54%
class recall	54.55%	72.73%	

G. Neural Net

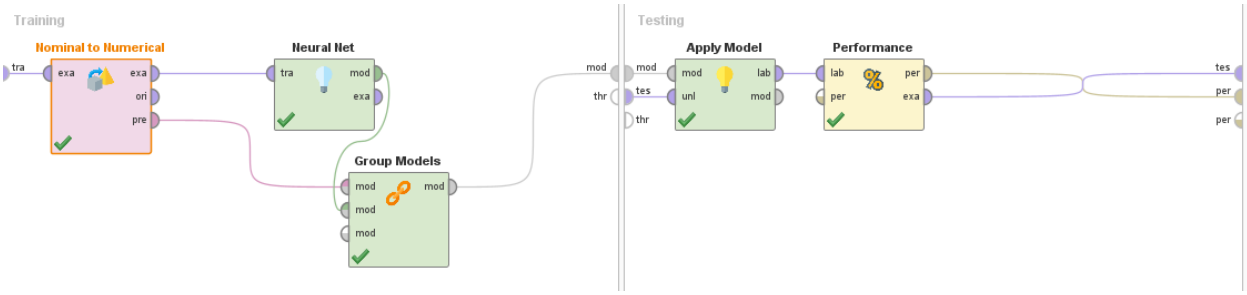
Process



ETL



Cross validation



Training

accuracy: 62.29% +/- 5.05% (micro average: 62.28%)

	true Good	true Bad	class precision
pred. Good	127	76	62.56%
pred. Bad	76	124	62.00%
class recall	62.56%	62.00%	

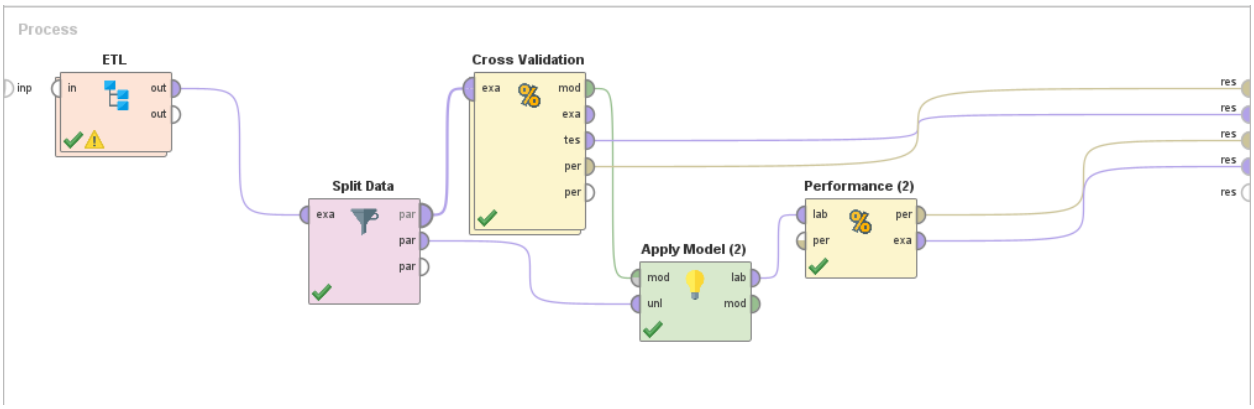
Learning

accuracy: 63.64%

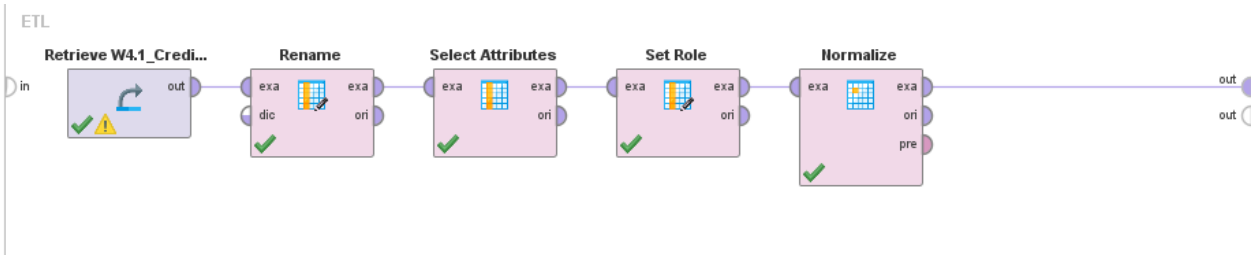
	true Good	true Bad	class precision
pred. Good	8	5	61.54%
pred. Bad	3	6	66.67%
class recall	72.73%	54.55%	

H. SVM

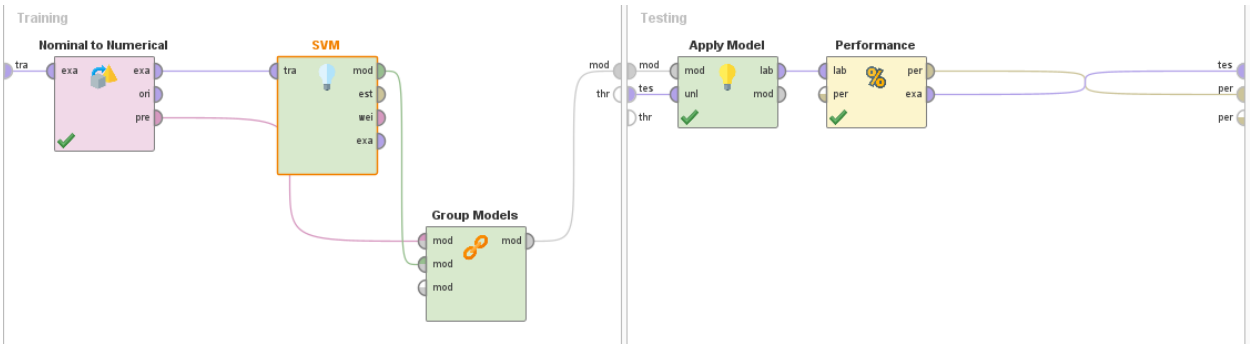
Process



ETL



Cross validation



Training

accuracy: 70.02% +/- 8.84% (micro average: 69.98%)

	true Good	true Bad	class precision
pred. Good	142	60	70.30%
pred. Bad	61	140	69.65%
class recall	69.95%	70.00%	

Learning

accuracy: 59.09%

	true Good	true Bad	class precision
pred. Good	5	3	62.50%
pred. Bad	6	8	57.14%
class recall	45.45%	72.73%	

V. Final Selection

The decision of which model to choose depends on specific problem we ae trying to solve and the requirement of the project. The below table shows the accuracy of models on training and learning model.

Model	**Accuracy (Train)**	**Accuracy (Learn)**	**Difference**
Decision Tree	69.23	63.64	5.59
Naïve Bayes	69.73	50.0	19.73
Random Forest	70.72	68.18	2.54
KNN	62.53	45.45	17.08
Logistic Regression	70.22	68.18	2.04
GLM	70.47	63.64	6.83
Neural Net	62.28	63.64	-1.36
SVM	69.98	59.09	10.89

Selection Criteria 1: Choosing a model that performs better on the test side. The training set accuracy may not always be a good indicator of how well the model will perform in new data. The model can be overfitting to the training set. Based on criteria, the following table is computed.

Model	**Accuracy (Train)**	**Accuracy (Learn)**	**Difference**
Decision Tree	69.23	63.64	5.59
Random Forest	70.72	68.18	2.54
Logistic Regression	70.22	68.18	2.04
GLM	70.47	63.64	6.83
Neural Net	62.28	63.64	-1.36

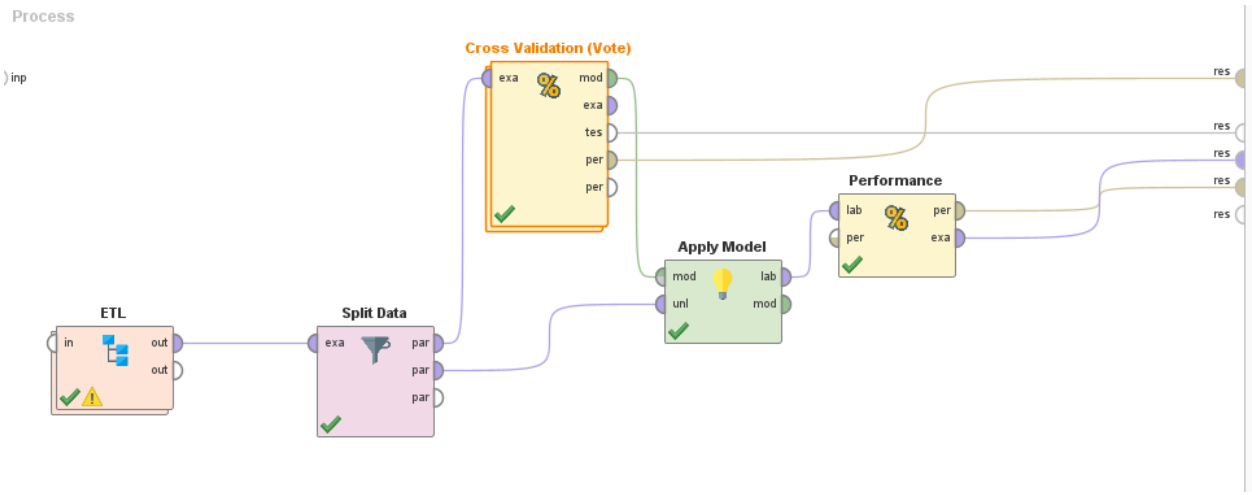
Selection Criteria 2: Figuring out the difference between the training and learning accuracy. This gives the idea of how much the model id overfitting. If the difference is large, this leads to overfitting to the training data that lead to poor generalization performance on new data. Based on this, below table is computed.

Model	**Accuracy (Train)**	**Accuracy (Learn)**	**Difference**
Logistic Regression	70.22	68.18	2.04
Neural Net	62.28	63.64	-1.36

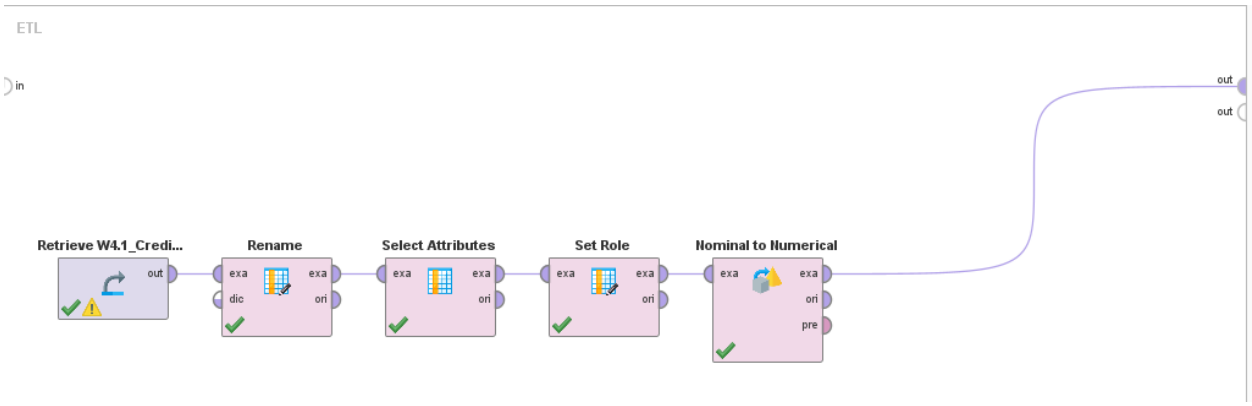
Selection Criteria 3: Considering the complexity of the model. Based on this, logistic regression is generally simple and more interpretable although neural network has higher learning accuracy and less difference.

VI. Ensemble Model

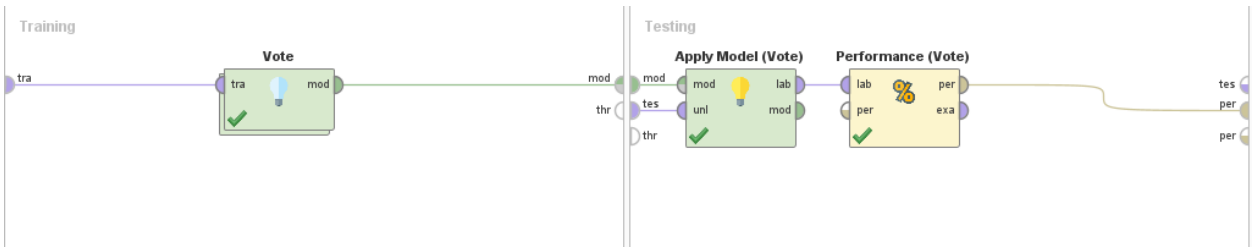
Process



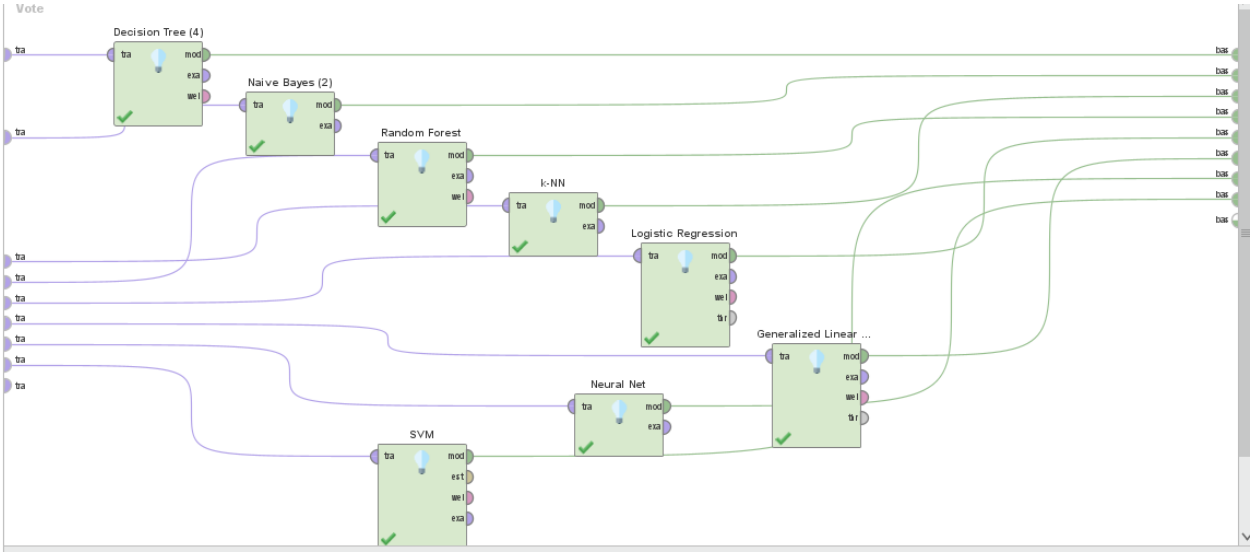
ETL



Cross validation (Vote)



Vote



Training

accuracy: 68.52% +/- 8.44% (micro average: 68.49%)

	true Good	true Bad	class precision
pred. Good	141	65	68.45%
pred. Bad	62	135	68.53%
class recall	69.46%	67.50%	

Learning

accuracy: 63.64%

	true Good	true Bad	class precision
pred. Good	6	3	66.67%
pred. Bad	5	8	61.54%
class recall	54.55%	72.73%	