CA 3 – CAP447

Section - D2112
Roll No – B81
Registration No – 12108348
Name – Aniket Kumar
Subject – Data Warehousing and Data Mining Laboratory

Task – Apply Classification Model on dataset

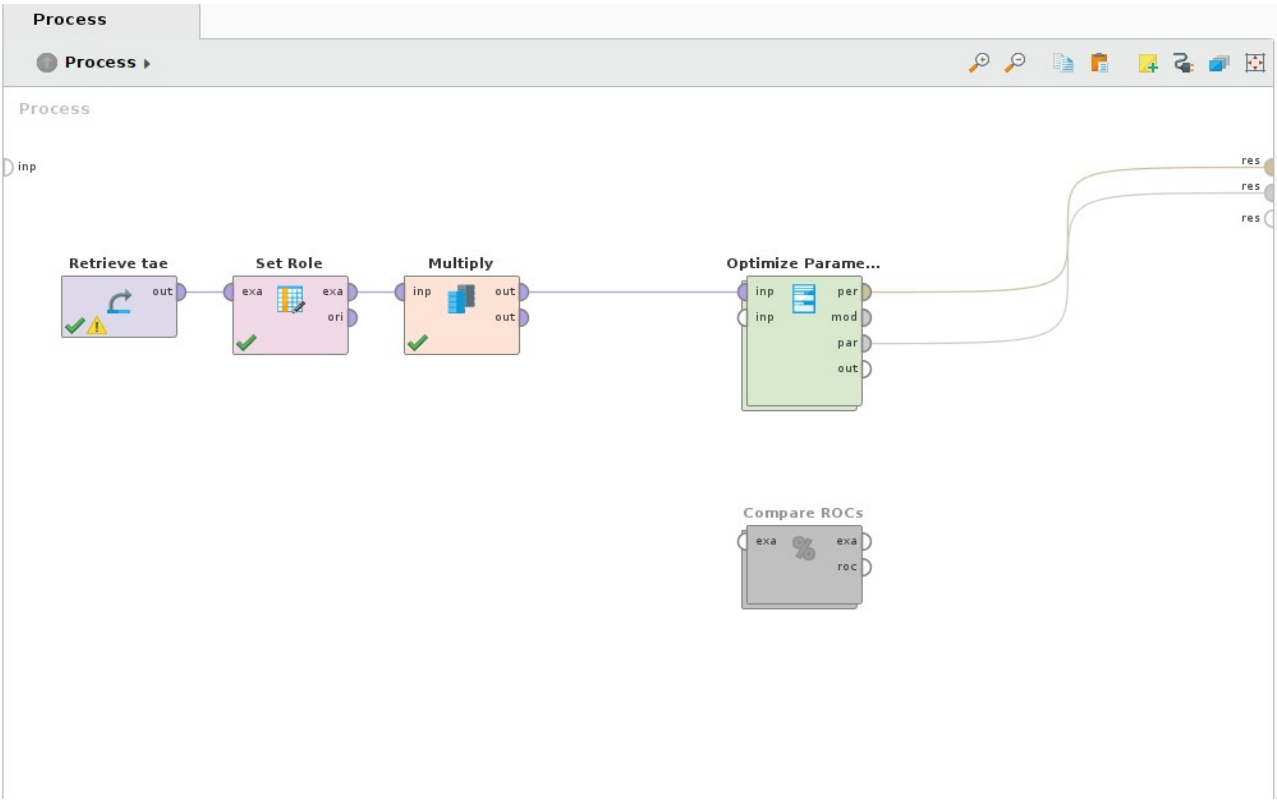
Dataset used – Teaching Assistant Evaluation (tae) dataset

Operators used -

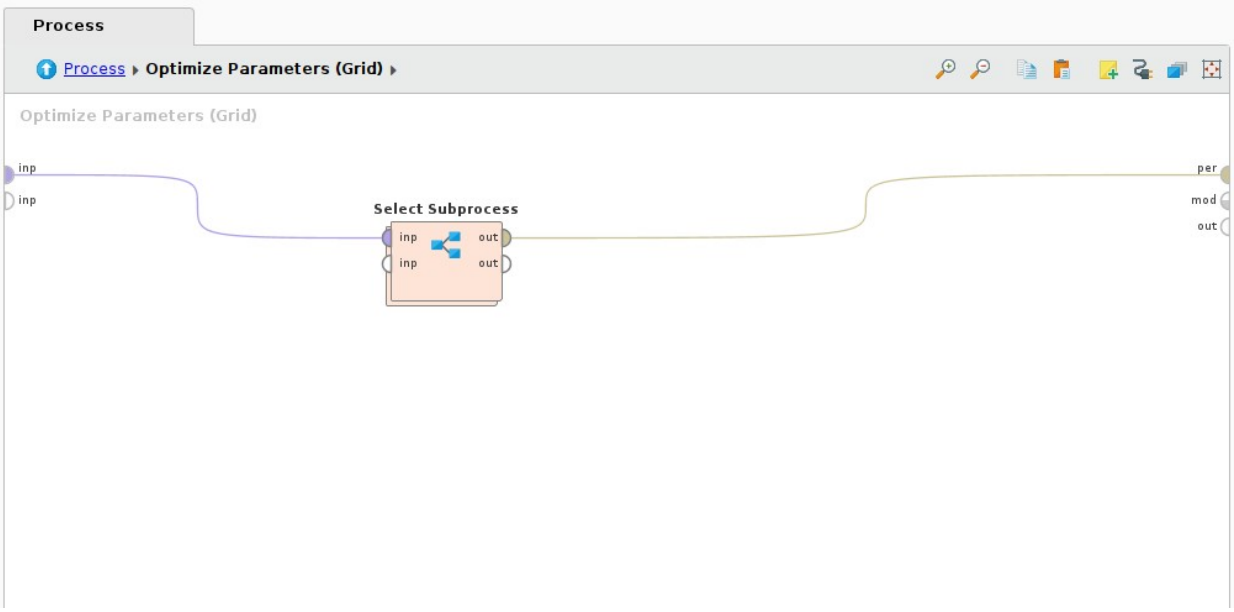
- **Set Role** – to change the role to label
- **Multiply** – to split the data in two parts, one is going to Optimize Parameters and other is going to Compare ROCs
- **Optimize Parameters (Grid)** – to find optimal values of selected parameters inside the subprocess
- **Select Subprocess –** This operator contains multiple subprocesses, but it executes only one subprocess at a time
- **Remember –** It stores the given object in the object store of the process
- **Cross Validation –** It performs a cross validation to estimate the statistical performance of a learning model
- **Decision Tree –** It generates a decision tree model, which can be used for classification
- **Random Forest –** It generates a random forest model, which can be used for classification
- **Rule Induction –** It learns a pruned set of rules with respect to the information gain from the given ExampleSet
- **Apply Model –** It applies a model on an ExampleSet
- **Performance (Classification) –** It is used for statistical performance evaluation of classification tasks
- **Compare ROCs –** It generates ROC charts for the models created by the learners in its subprocess and plots all the charts in the same plotter for comparison
- **Recall –** It retrieves the specified object from the object store of the process
- **Set Parameters –** It applies a set of parameters to the specified operators

Process

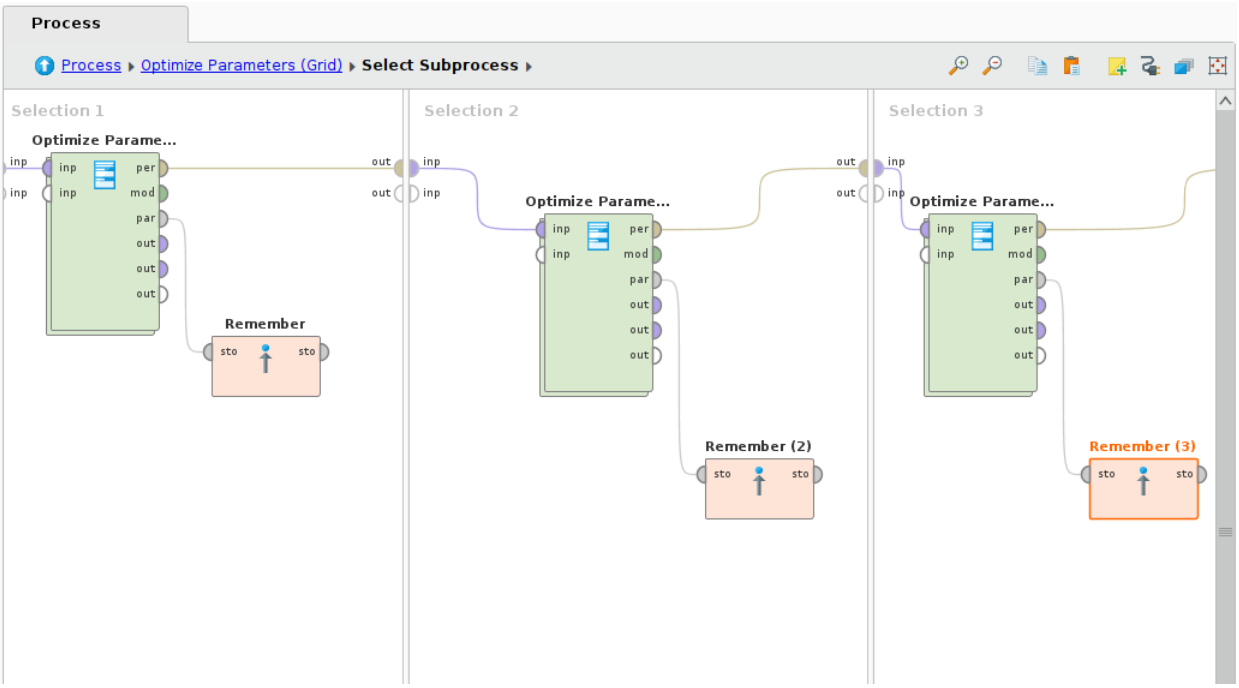
- Retrieve tae dataset
- connect Set Role Operator to it. In Parameters set attribute_name to Faculty_Performance and target role to label.
- Connect Multiply operator to Set Role.
- Connect Optimize Parameters (Grid) to the Multiply operator.
- In Parameters, go to Edit Parameter Settings. Here, Operator-Select Subprocess and Parameters-select_which.
- Connect per and par output ports of Optimize Parameters to the result port.



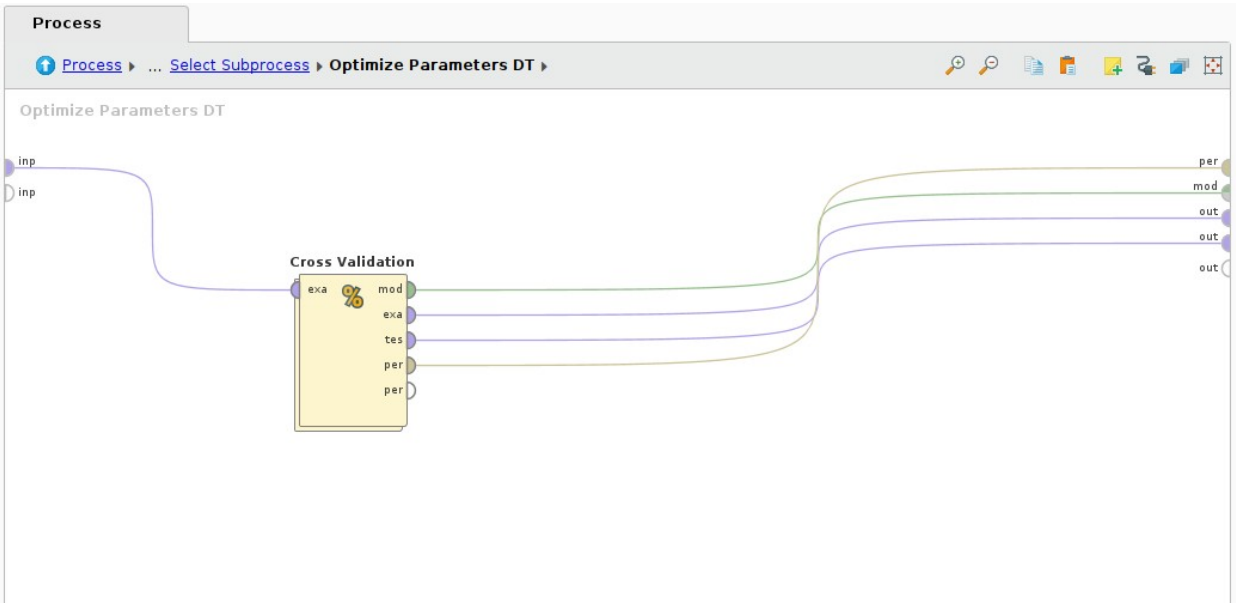
- Double Click on Optimize Parameters and add Select Subprocess Operator to it.



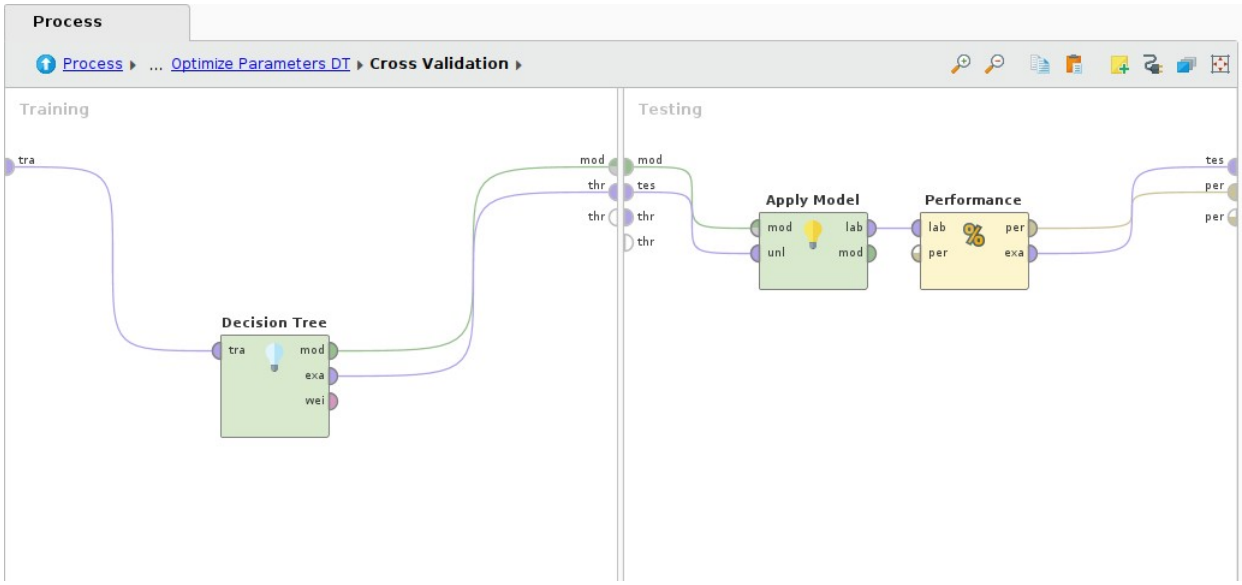
- Double Click on Select Subprocess operator.
- Here we have to add 3 different classification models using Optimize Parameters operator and connect Remember Operator to each of them.
- Parameters of Remember operator are: name-any name and io to ParameterSet.



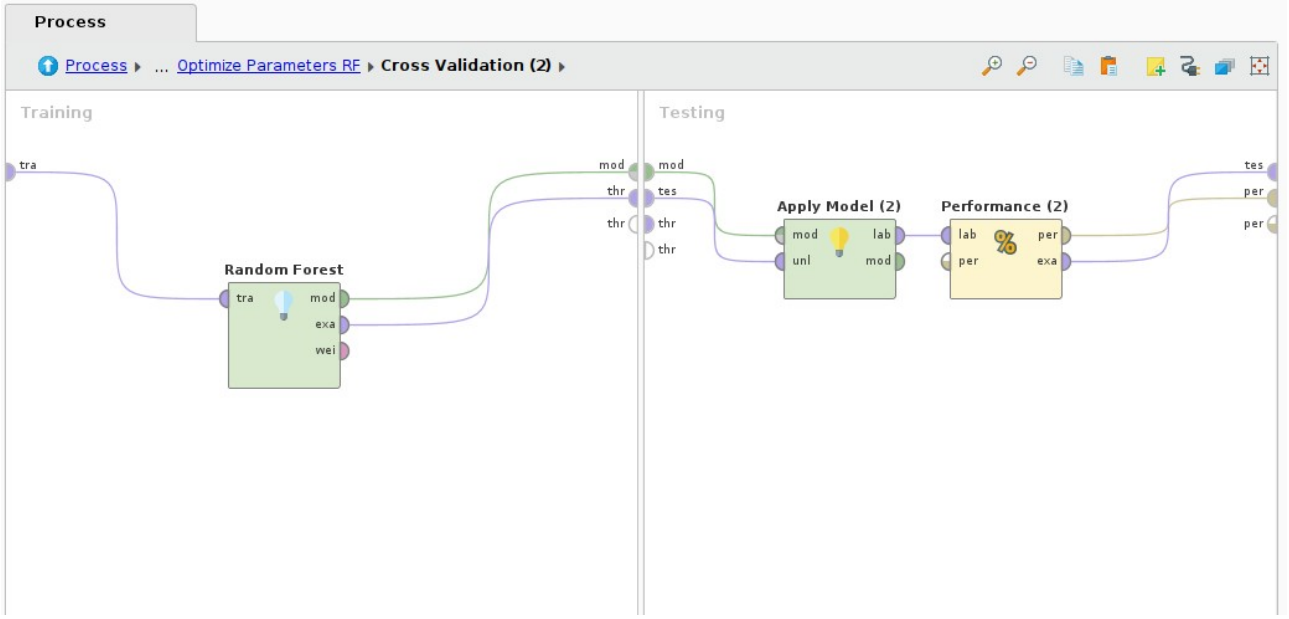
- Inside Optimize Parameters, we have to add Cross Validation operator.

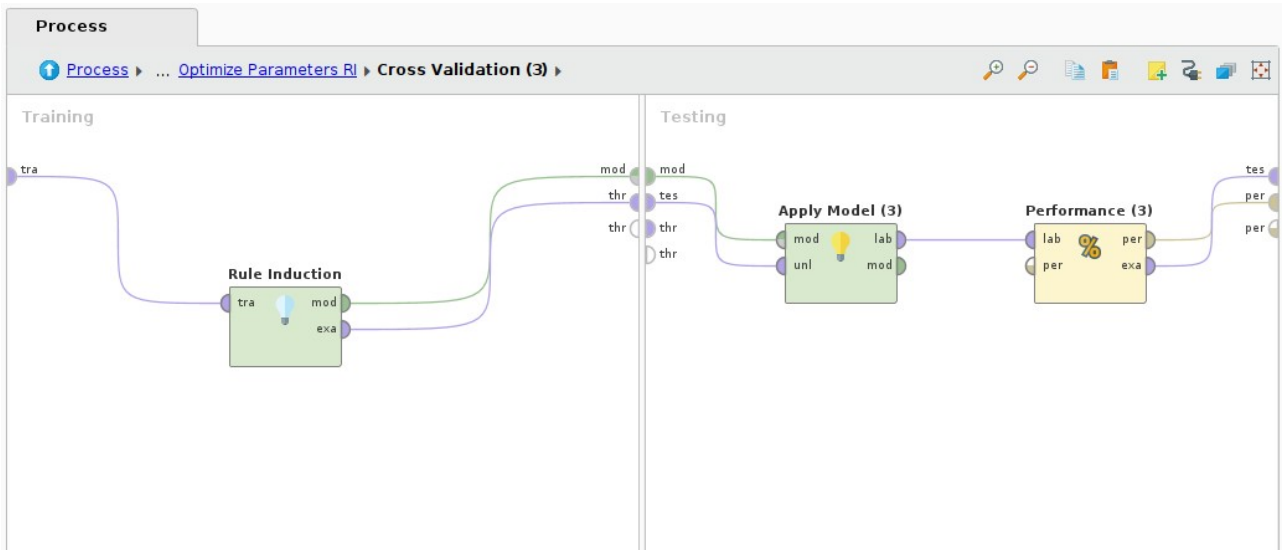


- Inside Cross Validation operator, add Decision Tree to Training part and Apply Model, Performance operators to Testing part.

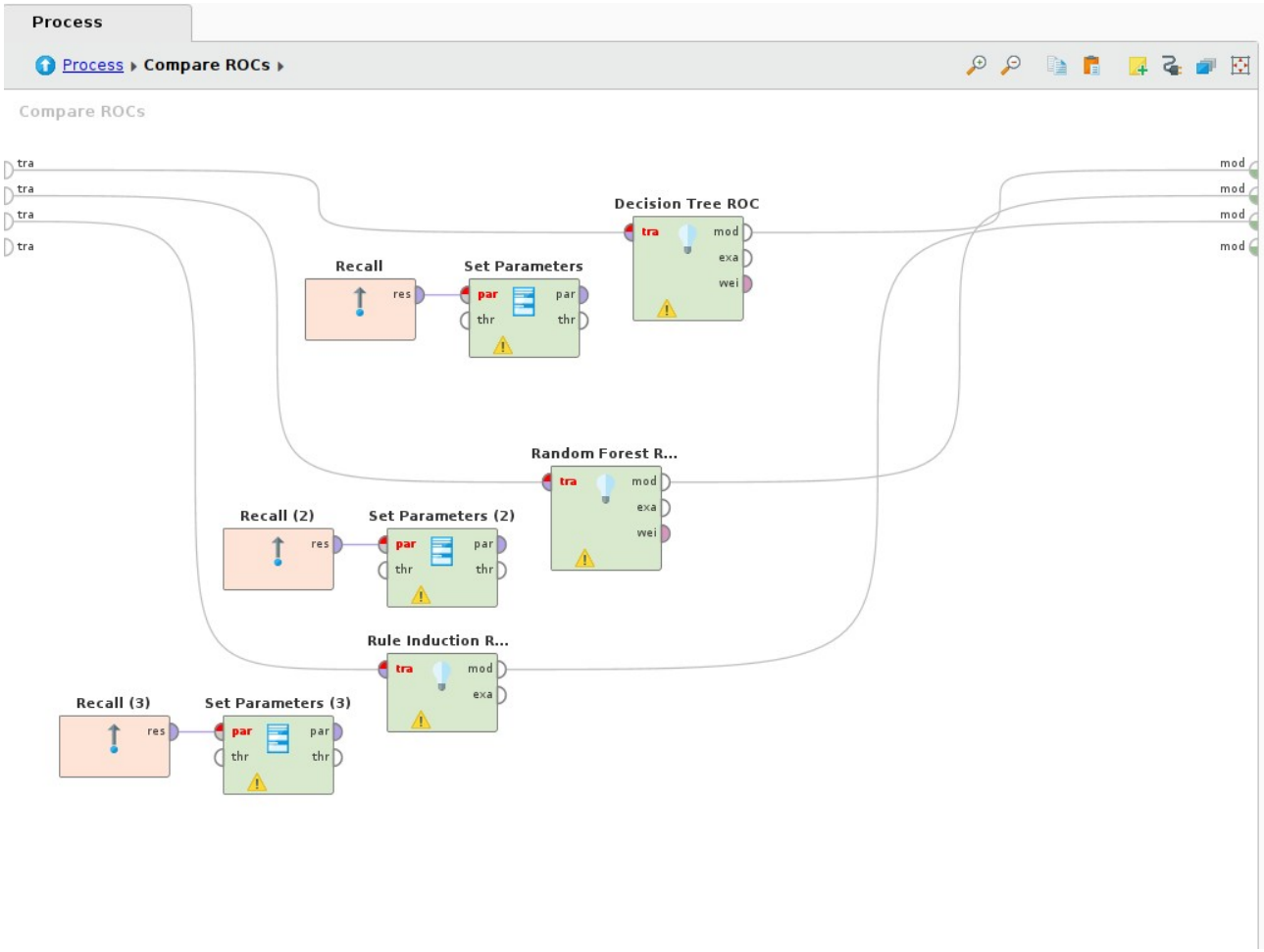


- Similarly add Random Forest and Rule Induction operators.





- Now, add Compare ROCs operator to the another output port of Multiply Operator.
- Inside Compare ROCs operator, add Decision Tree, Random Forest and Rule Induction operators one by one.
- Add 3 sets of Recall operator and Set Parameters Operators to inside the Compare ROCs operator one each for Decision Tree ROC, Random Forest ROC and Rule Induction ROC.
- In parameter of Recall operator, name is same as that of Remember operator.
- In parameter of Set Parameters, go to Edit List > set operator name as Optimize Parameters and Operator name as current operator name.



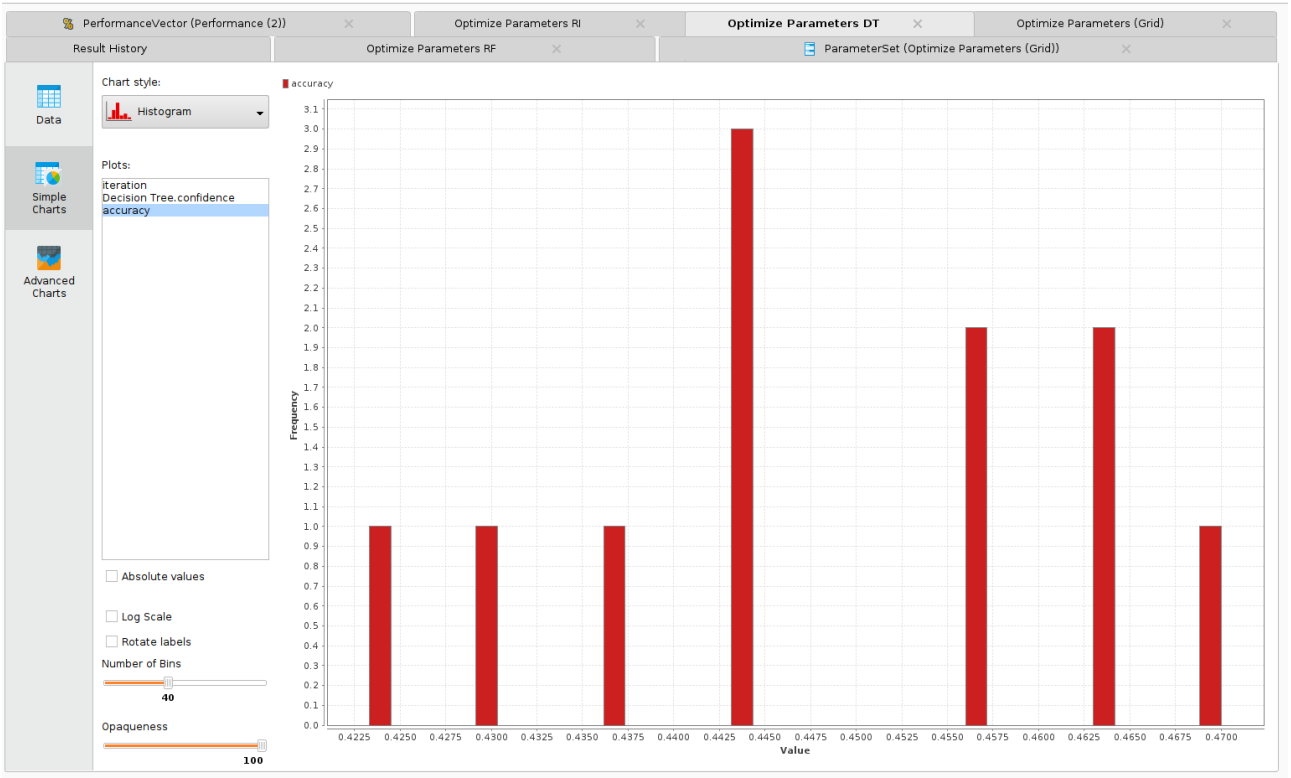
- Here, In this dataset, the attribute Faculty_Performance is not Binomial, hence, We are not going to use it as Compare ROCs operator only supports Binomial attributes.

Statistics

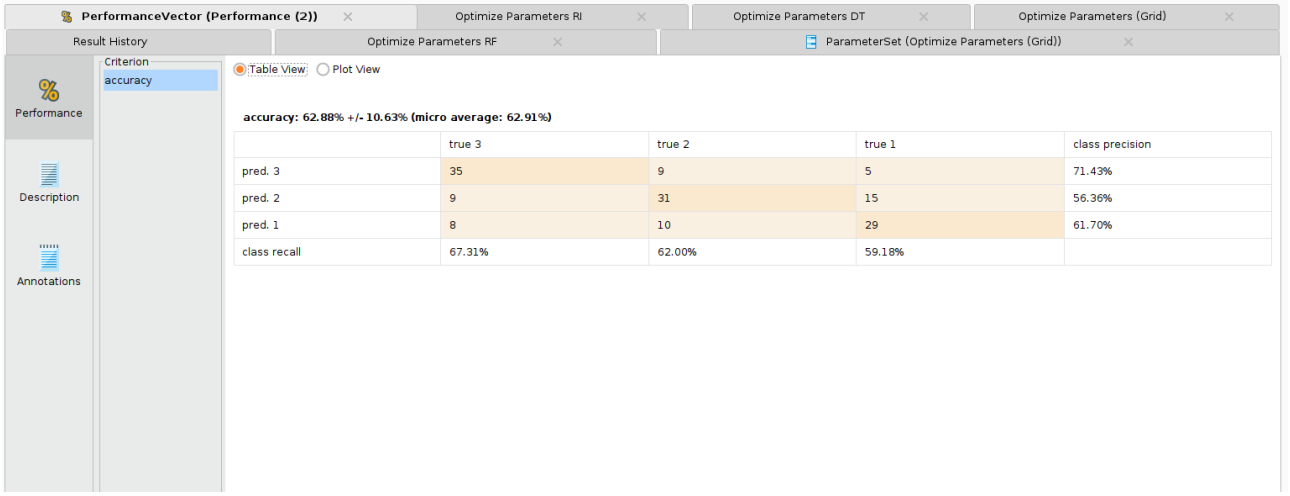
Name	Type	Missing	Statistics		
Filter (6 / 6 attributes): Search for Attributes					
ï»¿language	Binominal	0	Negative: 1	Positive: 2	Values: 2 (122), 1 (29)
faculty_id	Nominal	0	Least: 24 (1)	Most: 23 (17)	Values: 23 (17), 13 (14), ...[23 more]
course_id	Nominal	0	Least: 4 (1)	Most: 3 (45)	Values: 3 (45), 2 (16), ...[24 more]
summer_winter	Binominal	0	Negative: 1	Positive: 2	Values: 2 (128), 1 (23)
no_of_students	Integer	0	Min: 3	Max: 66	Average: 27.868
faculty_performance	Nominal	0	Least: 1 (49)	Most: 3 (52)	Values: 3 (52), 2 (50), ...[1 more]

Result

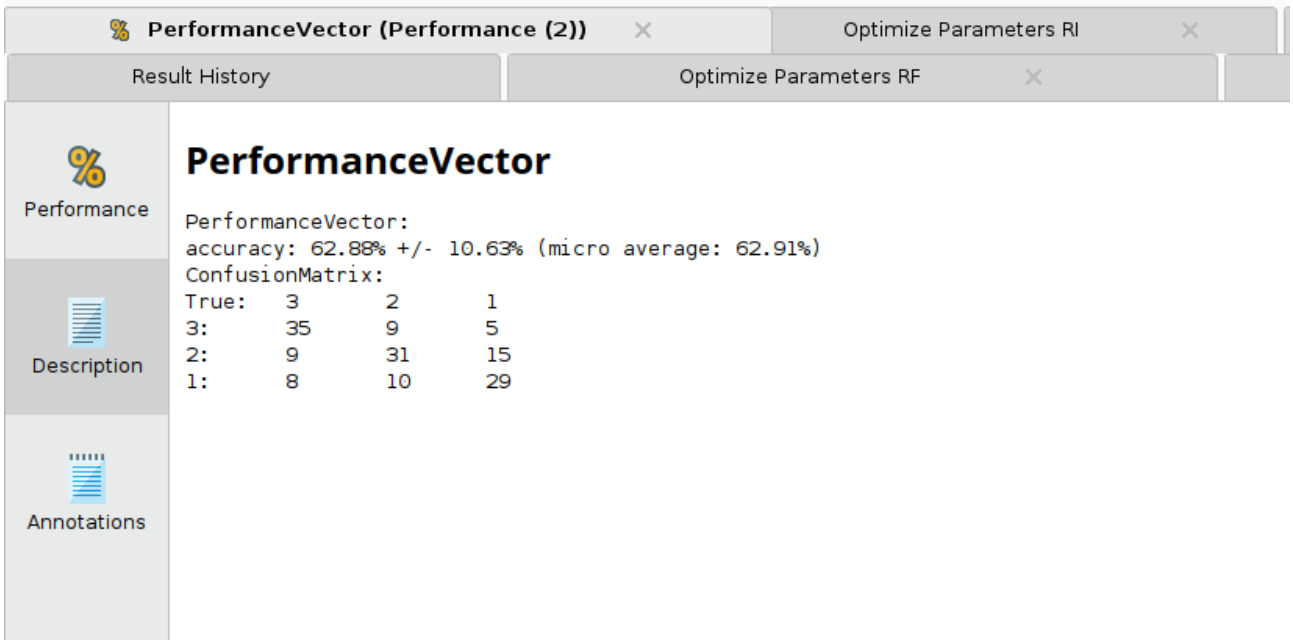
This graph shows the accuracy of Optimize Parameters DT.



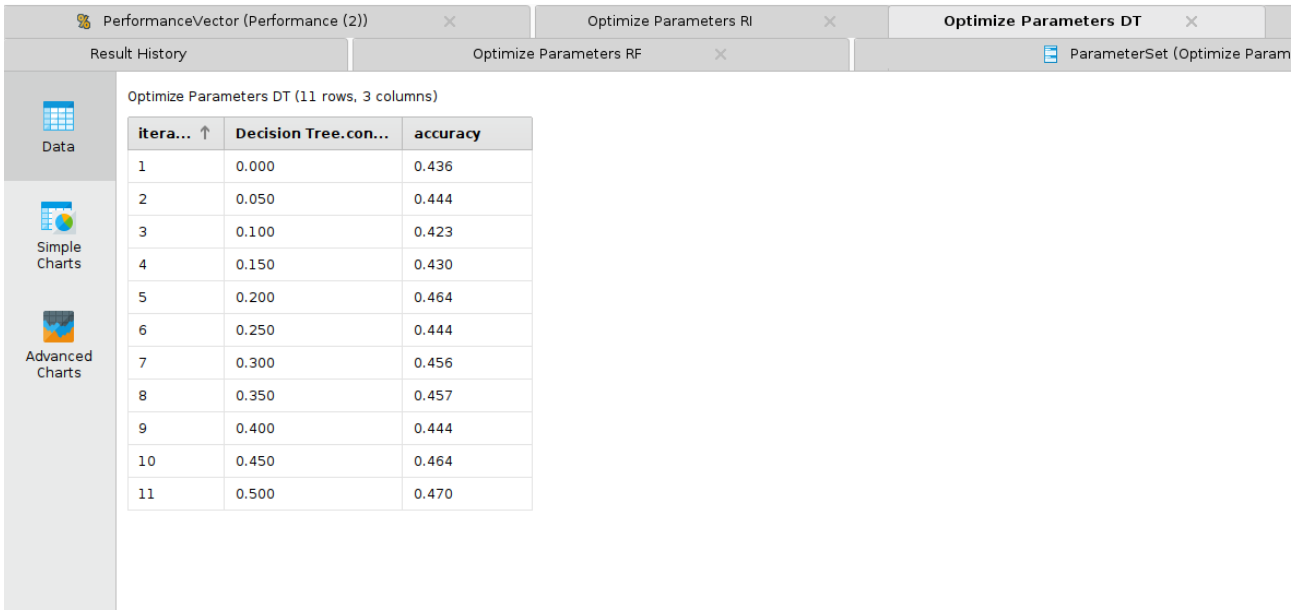
This table shows the accuracy of the classification model.
Here it is 62.88%.



This is the confusion matrix of the Performance Vector.

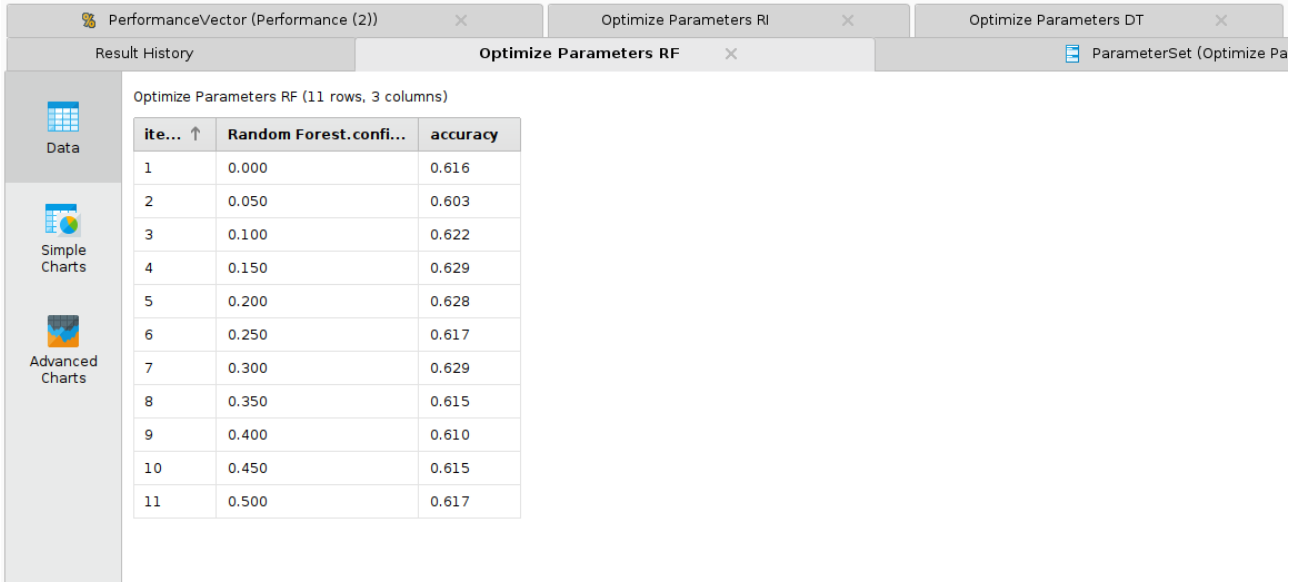


This is the value of accuracy and confidence of Decision Tree in each iteration.



itera...	Decision Tree.con...	accuracy
1	0.000	0.436
2	0.050	0.444
3	0.100	0.423
4	0.150	0.430
5	0.200	0.464
6	0.250	0.444
7	0.300	0.456
8	0.350	0.457
9	0.400	0.444
10	0.450	0.464
11	0.500	0.470

This is the value of accuracy and confidence of Decision Tree in each iteration.



ite...	Random Forest.confi...	accuracy
1	0.000	0.616
2	0.050	0.603
3	0.100	0.622
4	0.150	0.629
5	0.200	0.628
6	0.250	0.617
7	0.300	0.629
8	0.350	0.615
9	0.400	0.610
10	0.450	0.615
11	0.500	0.617