MINI PROJECT REPORT

BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)

SUBMITTED BY

Chinmay Nikam	41248
Shantanu Mangale	41242
Pratik Patil	41251
Gaurav Fugat	41219

DEPARTMENT OF COMPUTER ENGINEERING
P.E.S MODERN COLLEGE OF ENGINEERING
PUNE - 411005.

SAVITRIBAI PHULE PUNE UNIVERSITY

[2020 - 21]

Progressive Education Society's Modern College of Engineering
Department of Computer Engineering Shivajinagar, Pune - 411005.

CERTIFICATE

This is to certify that the following students of Final Year Computer Engineering of PES's, Modern College of Engineering have successfully completed the Laboratory Practice II (Mini Project) under the guidance of Mr. Dattatray Modani.

The Group Members are:

Chinmay Nikam	41248
Shantanu Mangale	41242
Pratik Patil	41251
Gaurav Fugat	41219

This is in partial fulfillment of the award of the degree Bachelor of Computer Engineering of Savitribai Phule Pune University.

Date:

Internal Guide Mr. Dattatray Modani	(Prof. Dr. Mrs. S. A. Itkar) HOD of Computer Engineering	External Examiner

Title:

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

Problem Definition:

Heart Failure Prediction using classification techniques such as Decision Tree, Logistic regression & Random Forest using Rapid Miner.

Prerequisite:

Basic Concepts of ETL

Software Requirements:

Rapid Miner

Outcome:

Rightly predicted Heart failure probabiltiy.

Data Set Description (Test and Training Data):

1] The data were taken from the Kaggle. There are 299 samples of Heart failure patients in the data sets. Each Heart failure patient sample (row) has the following characteristics (columns):

- Row No.
- Age
- Anemia
- Creatinine_Phosphokinase.
- Diabetes
- Ejection_fraction
- High_Blood_Pressure
- Platelets
- Serum_Creatinine
- Sex
- Smoking
- Time
- DEATH_EVENT

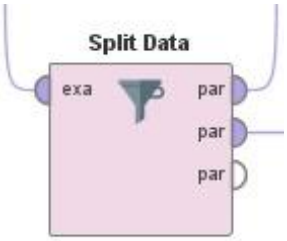
2] Wine quality dataset don't have any missing values. Correlation of alcohol, sulphates, density, citric acid, volatile acidity with quality of wine is greater so we use this attribute for prediction purposes.
3] The classification goal is to predict excellent or poor quality of wine.

Operators used:

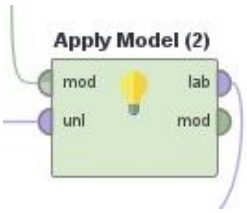
1. **Retrieve:** This Operator reads stored information in the Repository and loads them into the Process.

2. **Set Role:** This operator is used to set the role of one or more attributes to label, prediction, cluster, id, regular, weight or batch.

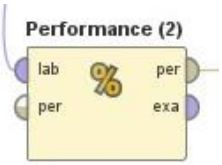
3. **Split Data:** This operator is used to split a dataset into parts. And then use training and testing.



4. **Apply Model:** This operator is used to run (i.e. to apply) train model on provided dataset.

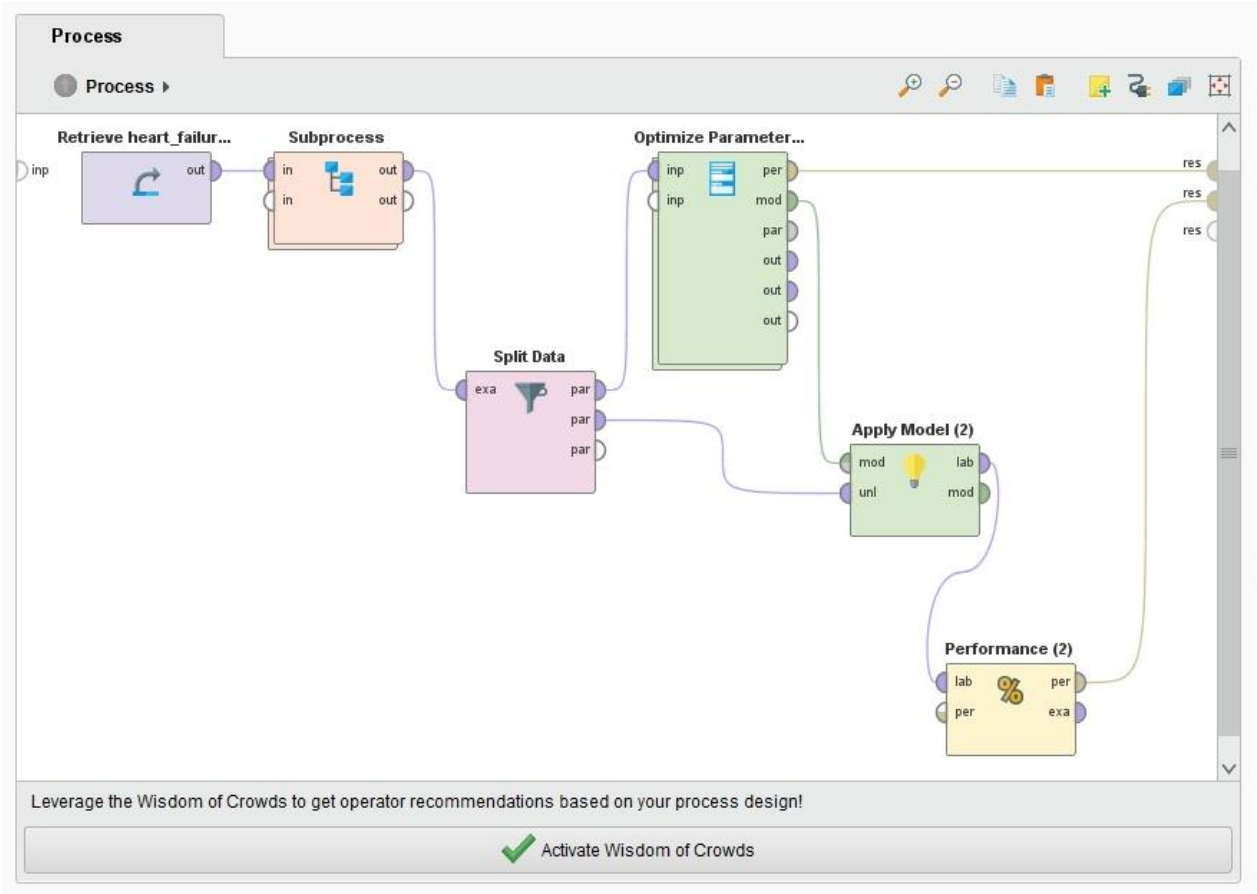


5. **Performance:** This operator gives us accuracy or performance of a trained model on the basis of the provided dataset by model.



Process:

In the following picture we design an approach for prediction of outcomes using different algorithms.



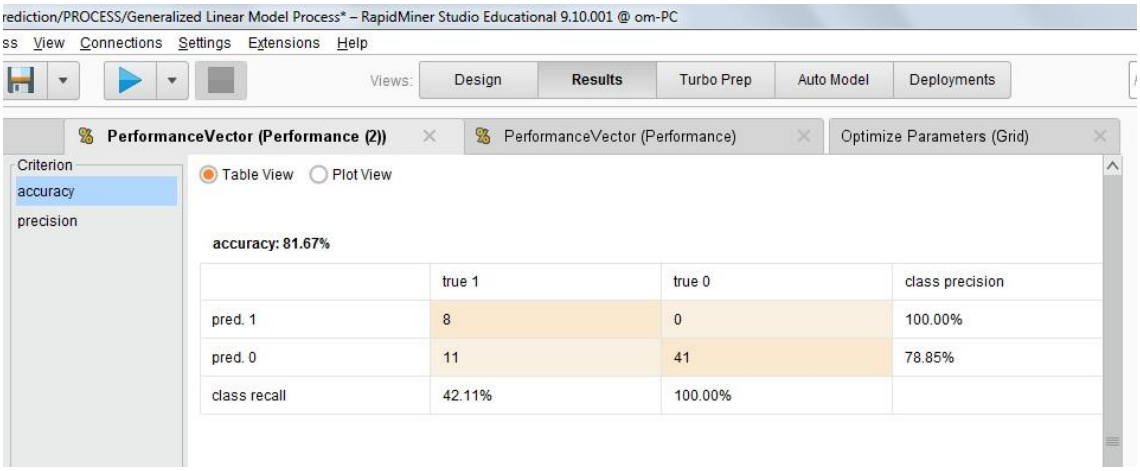
a) Generalized Linear Classification Model:

Generalized linear models (GLMs) are an extension of traditional linear models. This algorithm fits generalized linear models to the data by maximizing the log-likelihood. The elastic net penalty can be used for parameter regularization. The model fitting computation is parallel, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients.

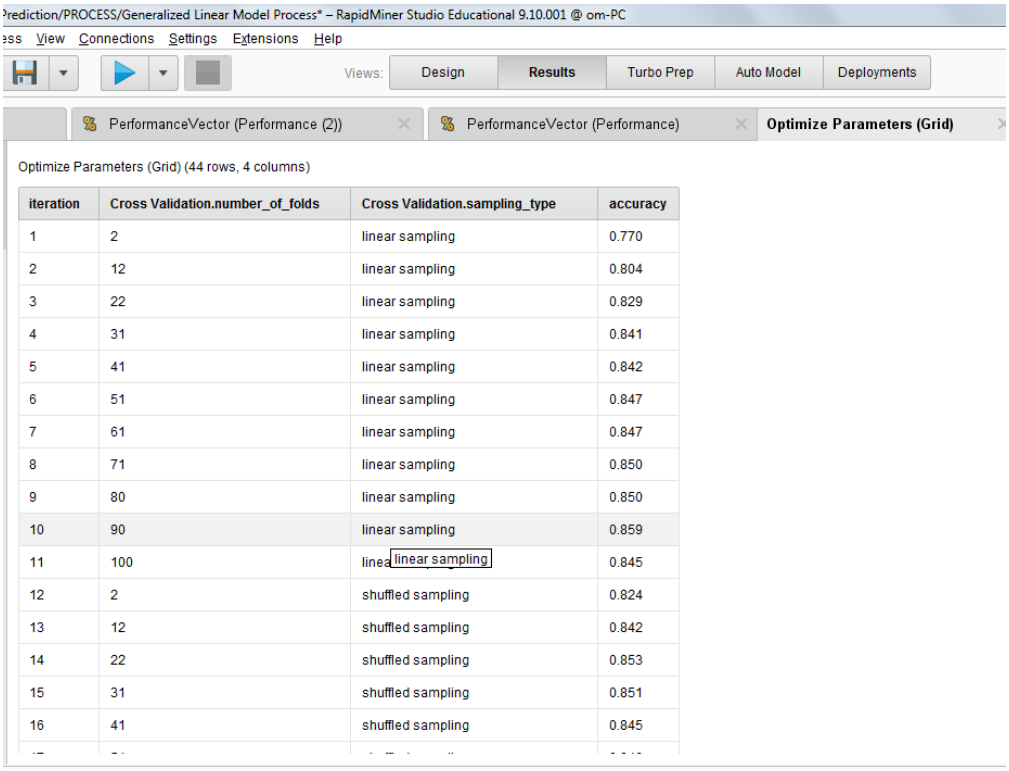
The operator starts a 1-node local H2O cluster and runs the algorithm on it. Although it uses one node, the execution is parallel. You can set the level of parallelism by changing the Settings/Preferences/General/Number of threads setting. By default it uses the recommended number of threads for the system. Only one instance of the cluster is started and it remains running until you close Rapid Miner Studio.

The GLM operator is used to predict the Future customer attribute of the Deals sample data set. All parameters are kept at the default value in the GLM. This means that because of the binominal label the Family parameter will be set automatically to "binominal", and the corresponding Link function to "logit". The resulting model is connected to an Apply Model operator that applies the Generalized Linear model on the Deals_Testset sample data. The labeled Example Set is connected to a Performance (Binominal Classification) operator that calculates the Accuracy metric.

Performance:



Example Set:



iteration	Cross Validation.number_of_folds	Cross Validation.sampling_type	accuracy
1	2	linear sampling	0.770
2	12	linear sampling	0.804
3	22	linear sampling	0.829
4	31	linear sampling	0.841
5	41	linear sampling	0.842
6	51	linear sampling	0.847
7	61	linear sampling	0.847
8	71	linear sampling	0.850
9	80	linear sampling	0.850
10	90	linear sampling	0.859
11	100	linear sampling	0.845
12	2	shuffled sampling	0.824
13	12	shuffled sampling	0.842
14	22	shuffled sampling	0.853
15	31	shuffled sampling	0.851
16	41	shuffled sampling	0.845

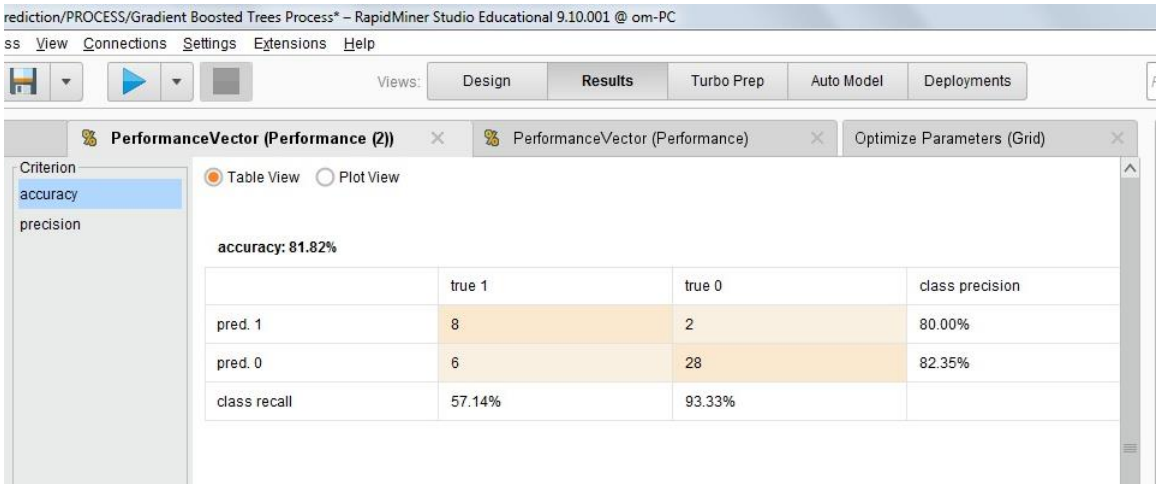
b) Gradient Boosted Trees Classification Model:

A gradient boosted model is an ensemble of either regression or classification tree models. Both are forward-learning ensemble methods that obtain predictive results through gradually improved estimations. Boosting is a flexible nonlinear regression procedure that helps improving the accuracy of trees. By sequentially applying weak classification algorithms to the incrementally changed data, a series of decision trees are created that produce an ensemble of weak prediction

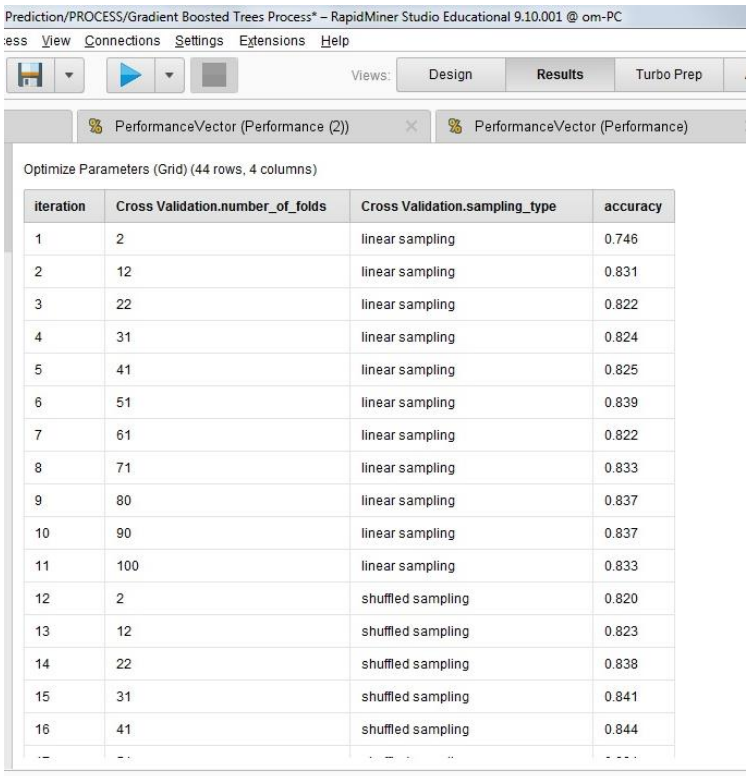
models. While boosting trees increases their accuracy, it also decreases speed and human interpretability. The gradient boosting method generalizes tree boosting to minimize these issues.

The operator starts a 1-node local H2O cluster and runs the algorithm on it. Although it uses one node, the execution is parallel. You can set the level of parallelism by changing the Settings/Preferences/General/Number of threads setting. By default it uses the recommended number of threads for the system. Only one instance of the cluster is started and it remains running until you close Rapid Miner Studio.

Performance:

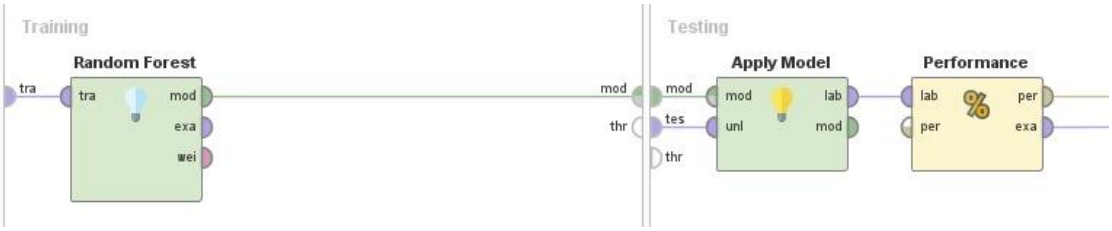


Example Set:



iteration	Cross Validation.number_of_folds	Cross Validation.sampling_type	accuracy
1	2	linear sampling	0.746
2	12	linear sampling	0.831
3	22	linear sampling	0.822
4	31	linear sampling	0.824
5	41	linear sampling	0.825
6	51	linear sampling	0.839
7	61	linear sampling	0.822
8	71	linear sampling	0.833
9	80	linear sampling	0.837
10	90	linear sampling	0.837
11	100	linear sampling	0.833
12	2	shuffled sampling	0.820
13	12	shuffled sampling	0.823
14	22	shuffled sampling	0.838
15	31	shuffled sampling	0.841
16	41	shuffled sampling	0.844

c) Random Forest Classification Model:

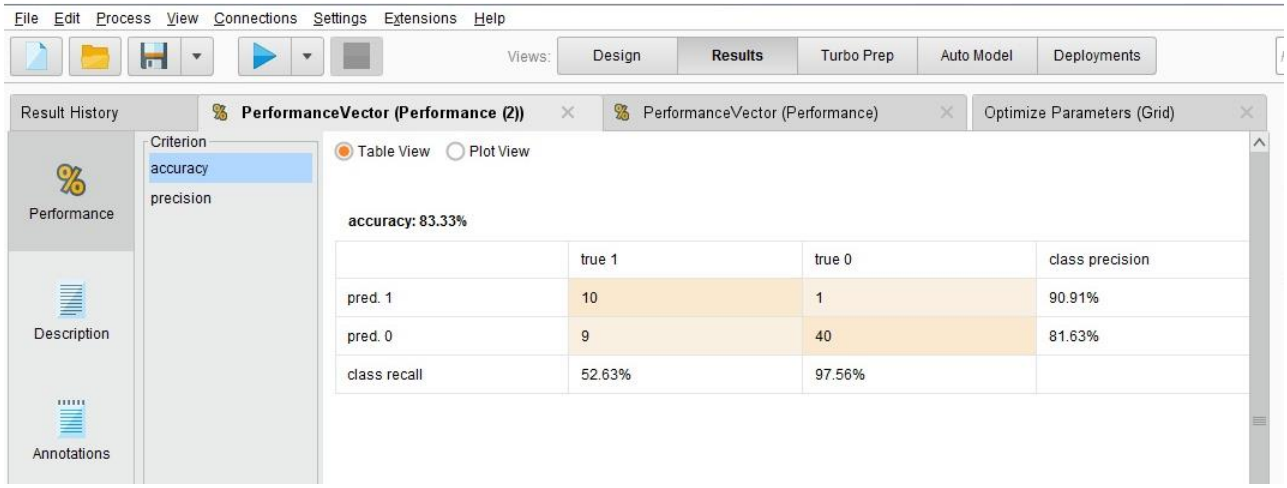


Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

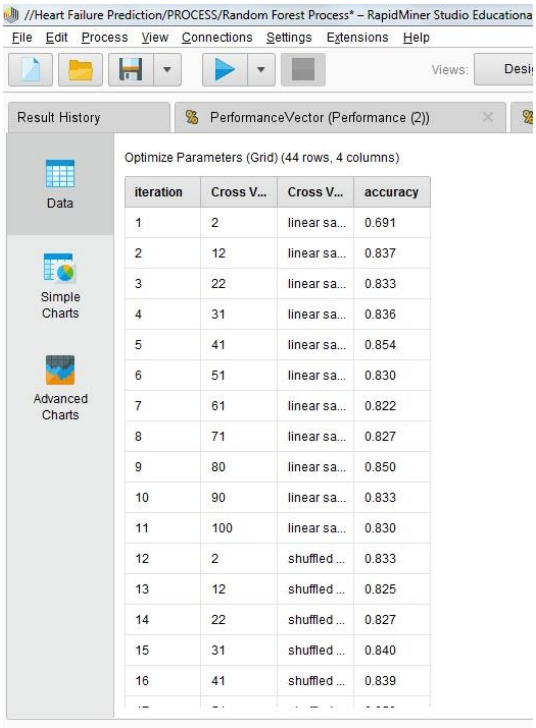
Standard decision tree classifiers have the disadvantage that they are prone to over fitting to the training set. The random forest's ensemble design allows the random forest to compensate for this and generalize well to unseen data, including data with missing values. Random forests are also good at handling large datasets with high dimensionality and heterogeneous feature types (for example, if one column is categorical and another is numerical).

Random forests are very good for classification problems but are slightly less good at regression problems.

Performance:



Example Set:



Accuracy Table:

Algorithmic Approach	Accuracy
Generalized Linear	81.67%
Gradient Boosted Trees	81.82%
Random Forest	83.33%

According to the above table Random Forest give the best accuracy.

Conclusion:

Heart Failure prediction using classification techniques generating results Generalized Linear, Random Forest & Gradient Boosted Trees using Rapid Miner.