zerocode learning

Learning Data Analytics Made Easy

USER GUIDE

RANDOM FOREST ANALYSIS

TABLE OF CONTENTS

INDEX

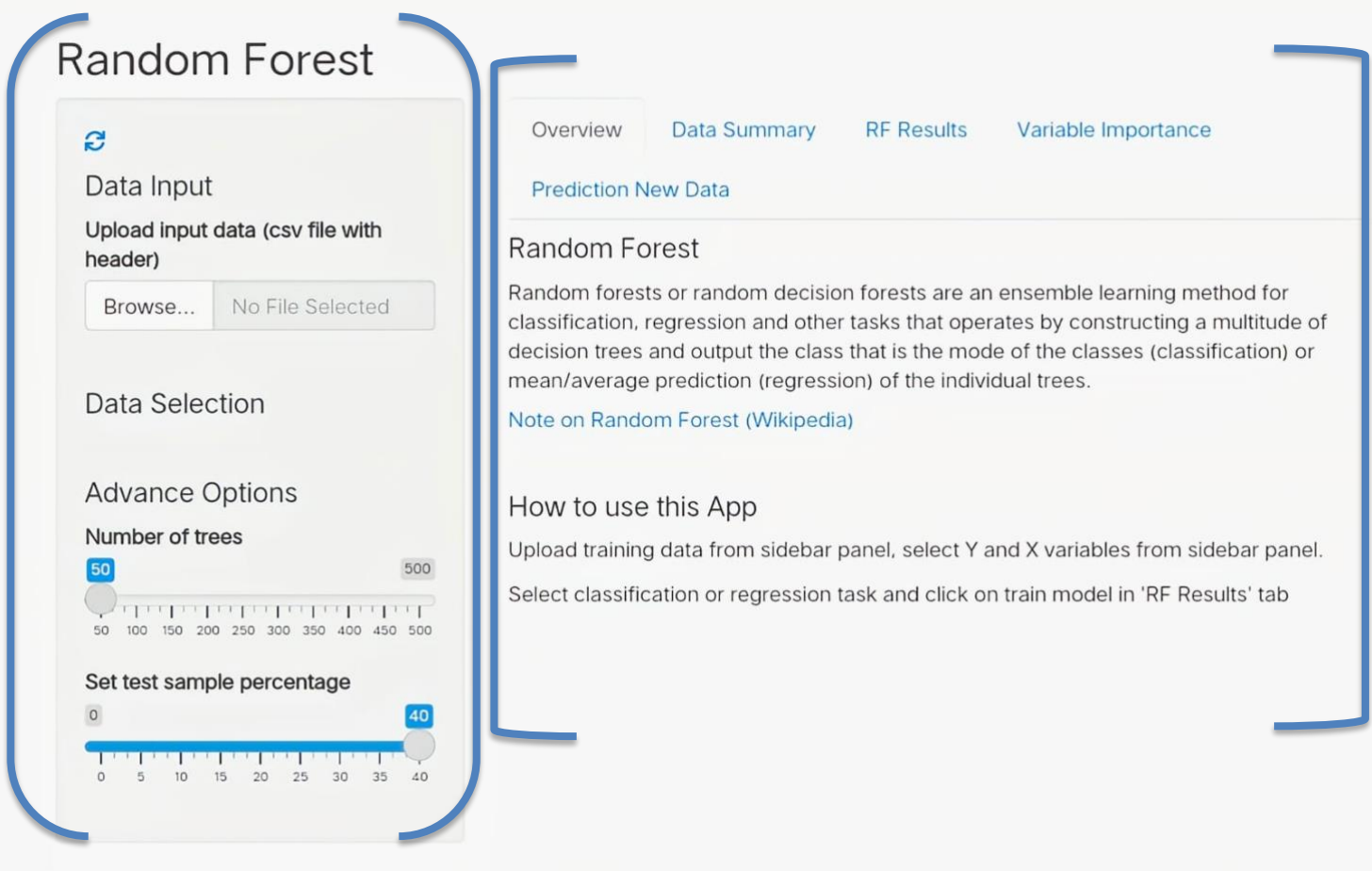
RANDOM FOREST ANALYSIS

01

Random forest, Random forests or random decision forests is a supervised machine learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. It is called a Random Forest because we use Random subsets of data and features and we end up building a Forest of decision trees. Random forest is a combination of many decision trees. A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)



LEFT PANEL (INP)

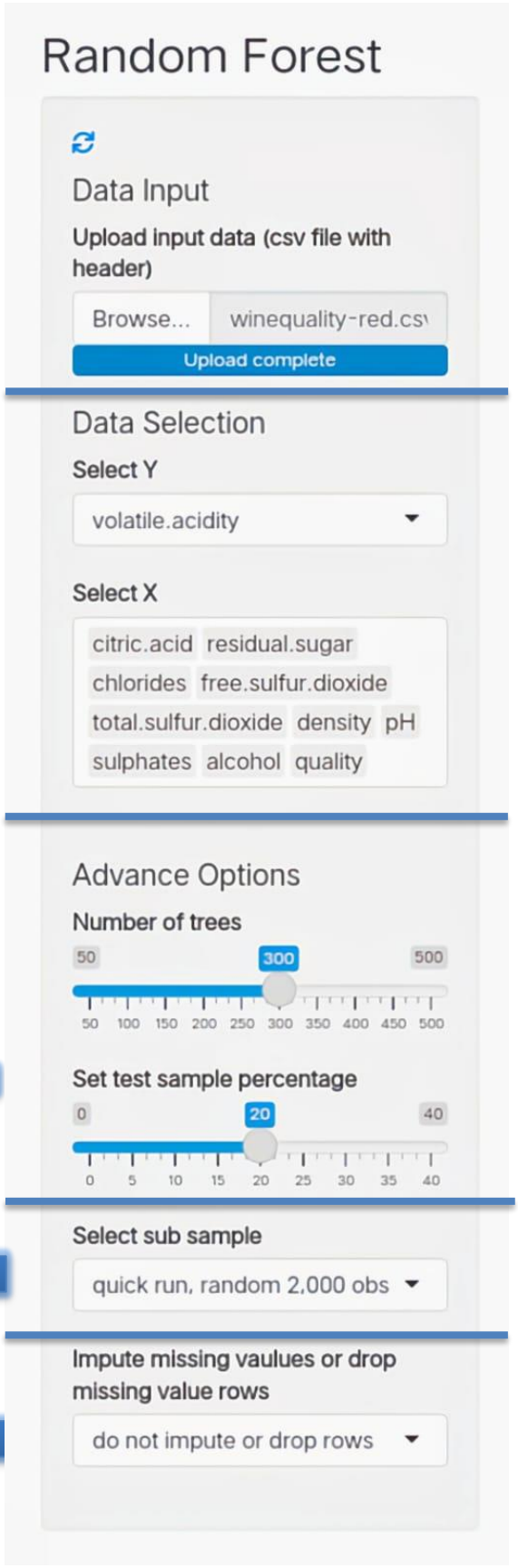
Upload your dataset here

Select your favorable variables required to base the analysis

Select the number of trees for analysis and can also vary the test sample percentage as per requirement

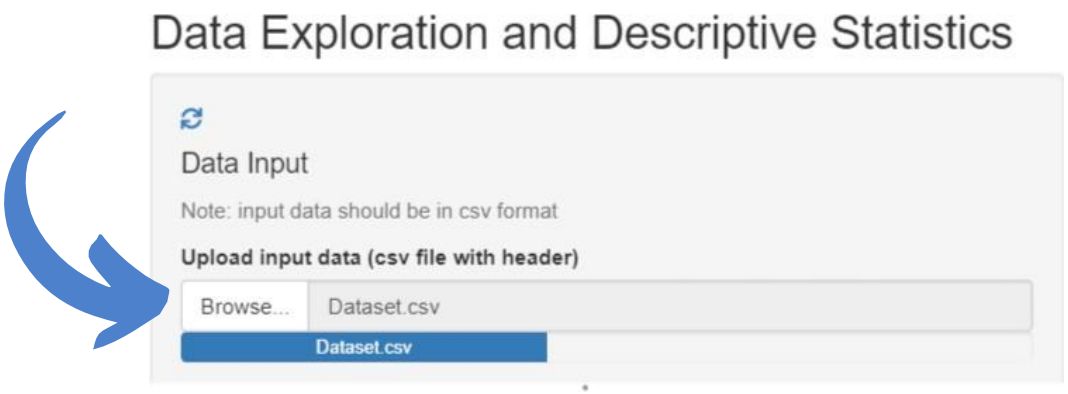
Select the subsamples or the whole data for testing.

Deal with missing values either drop or immute it.



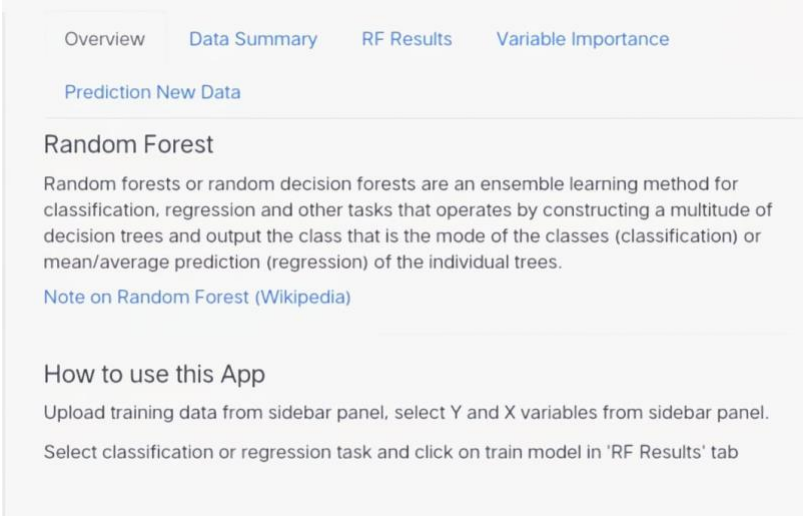
DATA INPUT
(UPLOADING DATASET)

- Click on browse
- Select the data file that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about random forest analysis and gives sample data and provides the description of analysis.



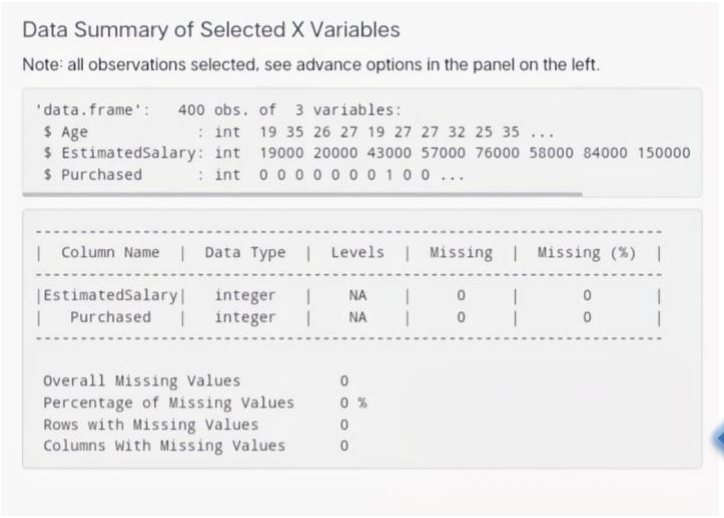
DATA SUMMARY TAB

It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **'Data Summary'** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' for all the data variables present in the dataset.

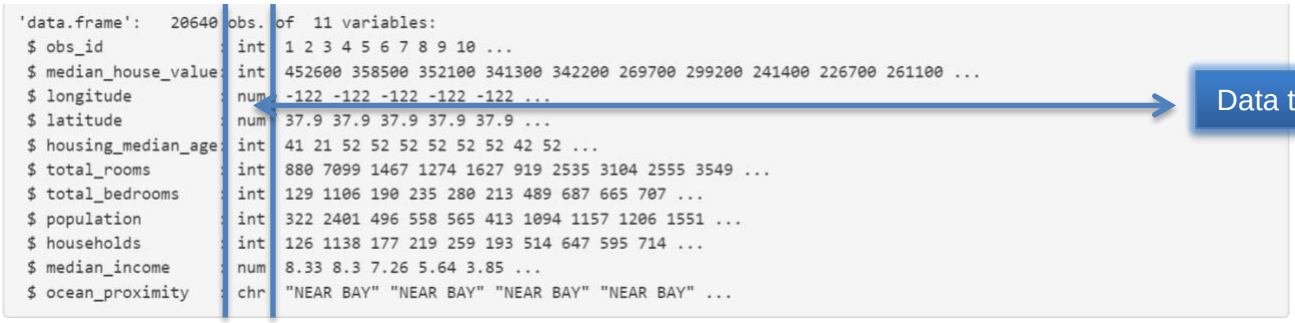
we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values, minimum and maximum value missing and null values etc.



This includes the minimum value maximum value , range between data values ,mean ,median ,mode with standard deviation that is the terms of statistics

Info about missing values

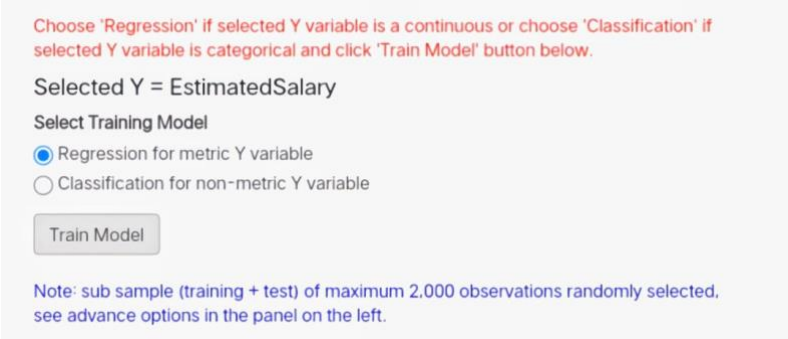
It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.



Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

RF RESULTS TAB

The RF(RANDOM FOREST) Result Tab allows the user to choose Either the RF regression or RF Classification for analysis based on the Y the predicting variable. If Y is continuous then it comes under regression else it will be under classification.



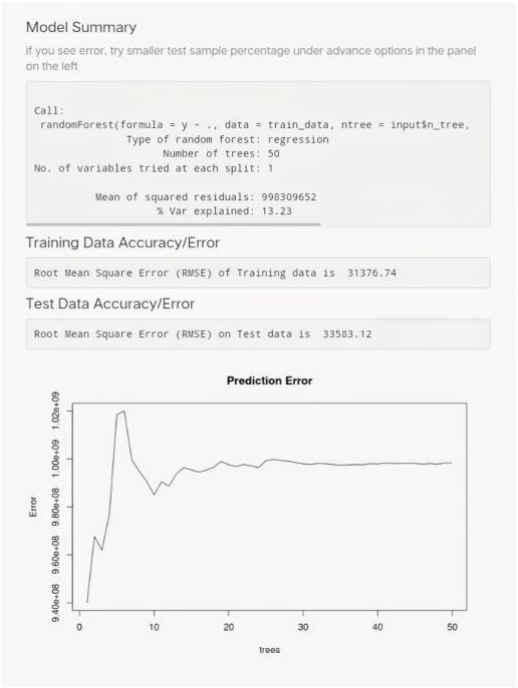
Select the regression model and train the model based on the requirement of Y prediction

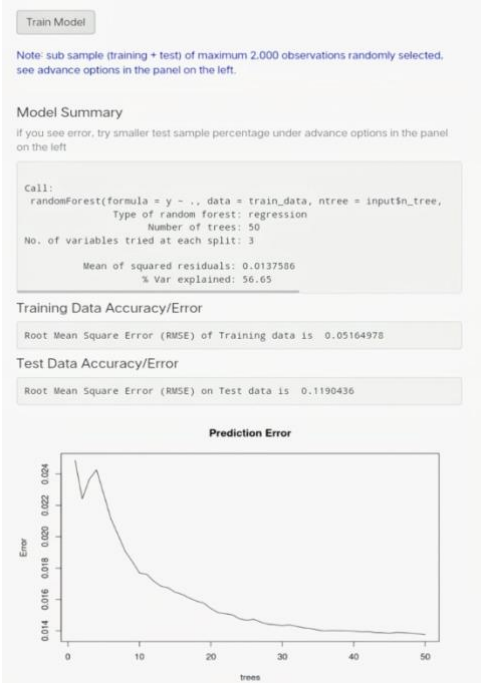
1) **RF REGRESSION:** Ensemble learning is the process of using multiple models, trained over the same data, averaging the results of each model ultimately finding a more powerful predictive/classification result.

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

The averaging makes a Random Forest better than a single Decision Tree hence improves its accuracy and reduces overfitting. A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest. A Random Forest's nonlinear nature can give it a leg up over linear algorithms, making it a great option. However, it is important to know your data and keep in mind that a Random Forest can't extrapolate.

2) **RF CLASSIFICATION :** The random forest is a classification algorithm consisting of many decisions trees.It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.





This is the result of the RF regression with the result and summary for given data where model is trained to perform regression.

This is the result of the RF classification with the result and summary for given data where model is trained to perform classification.

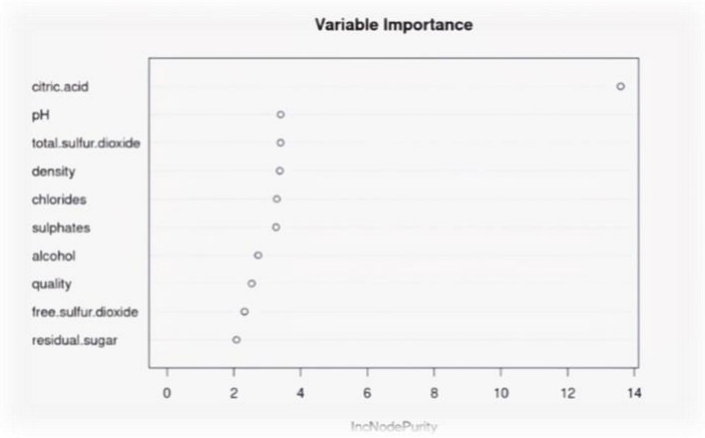
3) **Root Mean Square Error**: Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Based on a rule of thumb, it can be said that RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. In addition, Adjusted R-squared more than 0.75 is a very good value for showing the accuracy. In some cases, Adjusted R-squared of 0.4 or more is acceptable as well.

4) **PREDICTION GRAPH :** The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the Random Forest Classifier to be fit and validated whilst being trained 1.

Use the left panel to modify/deal with the outliers identified here.

VARIABLE IMPORTANCE TAB

The default method to compute variable importance is the mean decrease in impurity (or gini importance) mechanism: At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each. Variable importance refers to how much a given model "uses" that variable to make accurate predictions. The more a model relies on a variable to make predictions, the more important it is for the model. It can apply to many different models, each using different metrics. Variable importance is calculated by the sum of the decrease in error when split by a variable. Then, the relative importance is the variable importance divided by the highest variable importance value so that values are bounded between 0 and 1.



This is an example for showing the variable importance
For the dataset quality of red wine under classification analysis. Which gives the mean values by considering the attributes and their influence on the analysis.

Dataset References

https://www.kaggle.com/datasets/shub99/social-network-ads

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009