**HOMEWORK 3**

**Nikhil Sanjay Thorat**

**Spring 2020 ALY 6020**

**Predictive Analytics Instructor: Prof. Marco Montes de Oca**

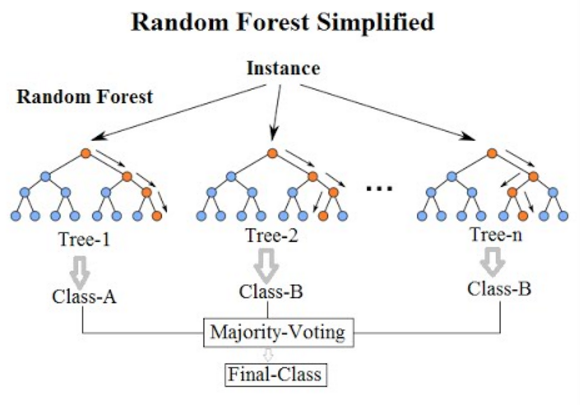
**Introduction to Random Forest:**

Random Forest consist of multiple decision trees. Each decision tree gives predictors which have the most number of votes. A random operation is performed which creates some trees. The robustness of the model depends on the number of trees, more the number of trees is the accuracy. The decision trees are trained with the bagging method.

**Bagging method:** Bagging method is used to increase the accuracy of the model and is mainly used in classification and regression problems. It is well defined averaging approach.

**Advantages of using Random Forest:**

* Using random forest, we can perform regression as well as classification.
* It can be used to handle missing values.
* Random forest models can’t be overfitted.
* It can also be operated on categorical variables.



**Fig.1**

Fig.1 Shows the simplified version of the Random Forest algorithm.

**Working of Random Forest:**

1. Selecting Random samples from the dataset.
2. Developing a decision tree for every sample and obtaining the result.
3. The result will be predicted using the number of votes.
4. The prediction results will be based on the most number of votes.

**Homework Outline:** The goal of the assignment is to classify how perfect a model can learn to compute statistics from a sample dataset.

**Dataset description:** The given dataset was downloaded from the link mentioned below:

<https://drive.google.com/file/d/1RSDumnSgINshUAu4zGjC2fsVIO6nOLXM/view>

The dataset is divided into two different parts Train and Test and the target variable is in the form of multiple columns which are quantile 0.1, 0.5, 0.9 including average and variance. The dataset consists of unknown probability distribution.

**Analysis:**

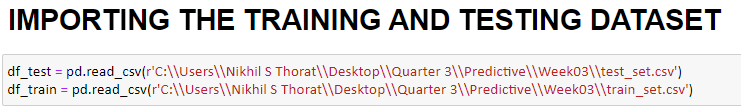
**Step1:**



**Fig.2**

From fig.2 we can see that we have inserted important packages required for performing the Random Forest machine learning algorithm.

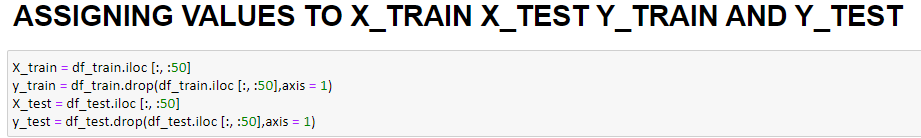
**Step 2:**



**Fig.3**

In Fig.3 we are importing dataset from a specific location present in the system. The test data is stored in df\_test and the training data is stored in df\_train.

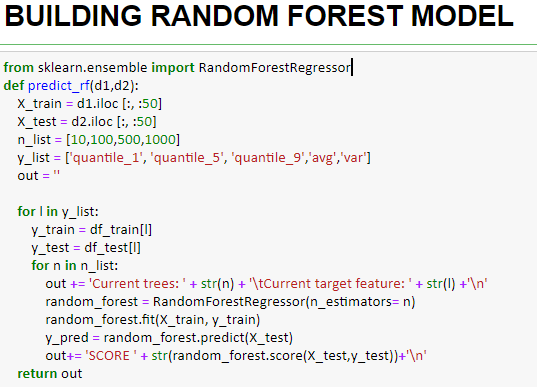
**Step 3:**



**Fig.4**

In fig.4 the values are assigned to X\_train, Y\_train, X\_test and Y\_test. X\_train and X\_test variables hold the values of the first 50 columns, which are considered to be the independent variables. This is done using the iloc function. Y\_Train and Y\_Test values hold the target variables which are the statistic values present in the dataset. This is done by dropping the 50 independent variables and assigning the statistics values using iloc function.

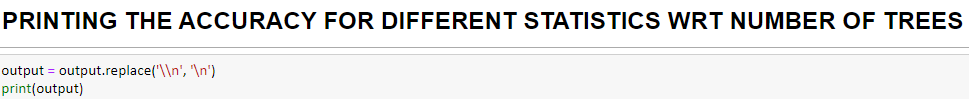
**Step 4:**



**Fig.5**

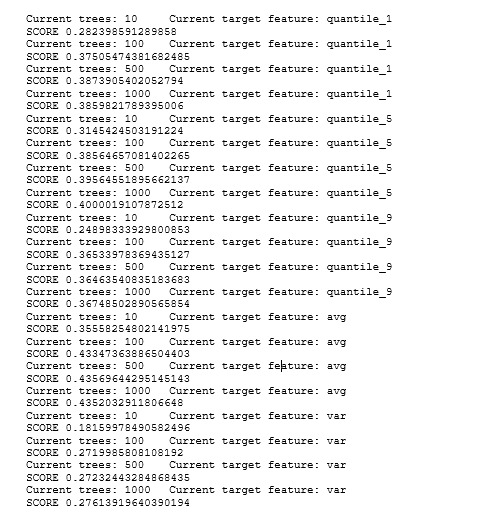
In fig.5 a random forest regressor model is built using sklearn.ensemble where X\_train and X\_test are defined with features variables. N\_list is declared as an array that holds the values of the number of trees for which the random forest model needs to be built. Y\_list is declared to be another array which contains all the five target variables. These arrays are later used in the loop to build the random forest model. Random forest regressor is used here to build the random forest model. Random forest regressor is implemented here as it can perform regression and classification. This is achieved by using the RandomForestRegressor function from sklearn package. This model is trained using X\_Train and Y\_Train values and X\_Test values are predicted using the model.predict() function. The accuracy of the model is predicted using the model.score() function, where X\_test and y\_test parameters are passed. This is executed in a loop, which traverses through the array to get the model for five statistics (Target variables) using four number of trees mentioned and the output is printed.

**Step 5:**



**Fig.6**

The above function is used to print the accuracy of different models that were built with different combinations of trees. The output obtained is shown in fig.7.



**Fig.7**

**Results:**

The accuracy obtained for different combinations is mentioned above. The highest accuracy obtained in this combination was with trees: 500, target variable: avg. The accuracy was around 43%. In the fifth model where the target variable is variance, the accuracy observed is very low. As we go from quantile 1 to quantile 9 the accuracy increases. The average column also shows high accuracy but the accuracy gets dropped in the variance model.

**References:**

[1] Understanding Random Forest

Tony Yiu - <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

**Dataset:**

<https://drive.google.com/file/d/1RSDumnSgINshUAu4zGjC2fsVIO6nOLXM/view>