HW1- CSE 291: Pattern Recognition

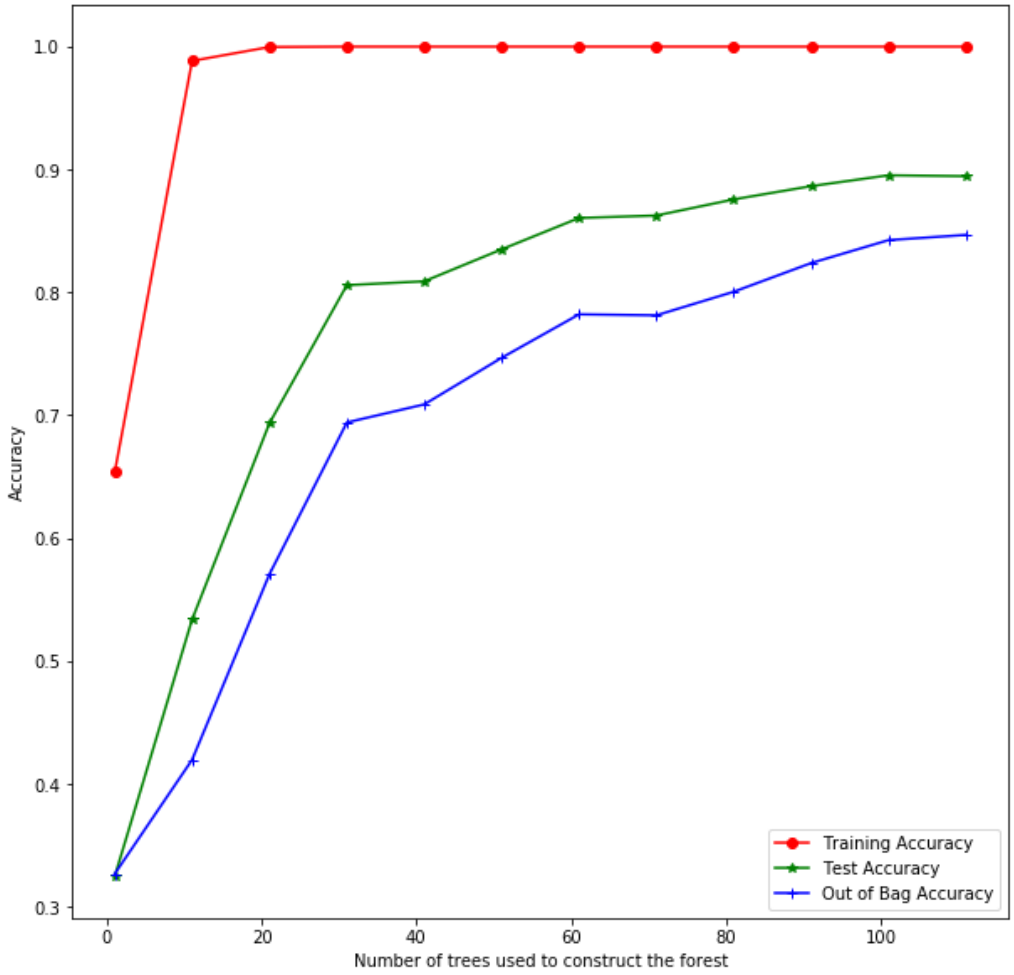
Ronit Shaw A53220859
Sumedha Khatter A53094878

Random Forest

ON MNIST DATA

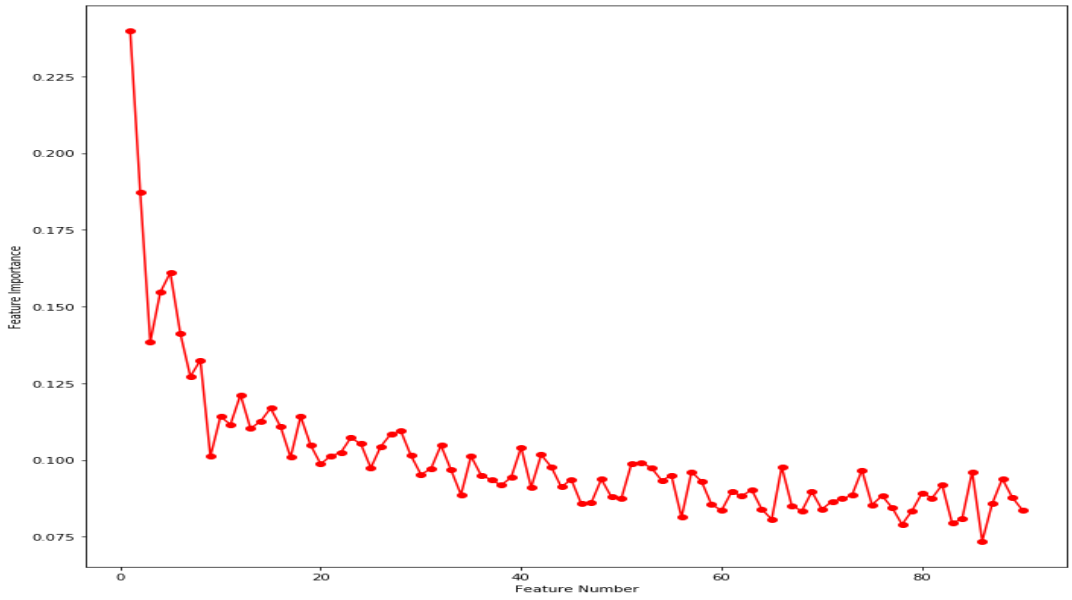
Initially PCA is done on the data set to find out the relevant components of the data and then out of those 90 is chosen principal components are chosen and the data is projected into those components. Then using them as features Random forest algorithm is built on the data.

With decision trees grown to the full extent

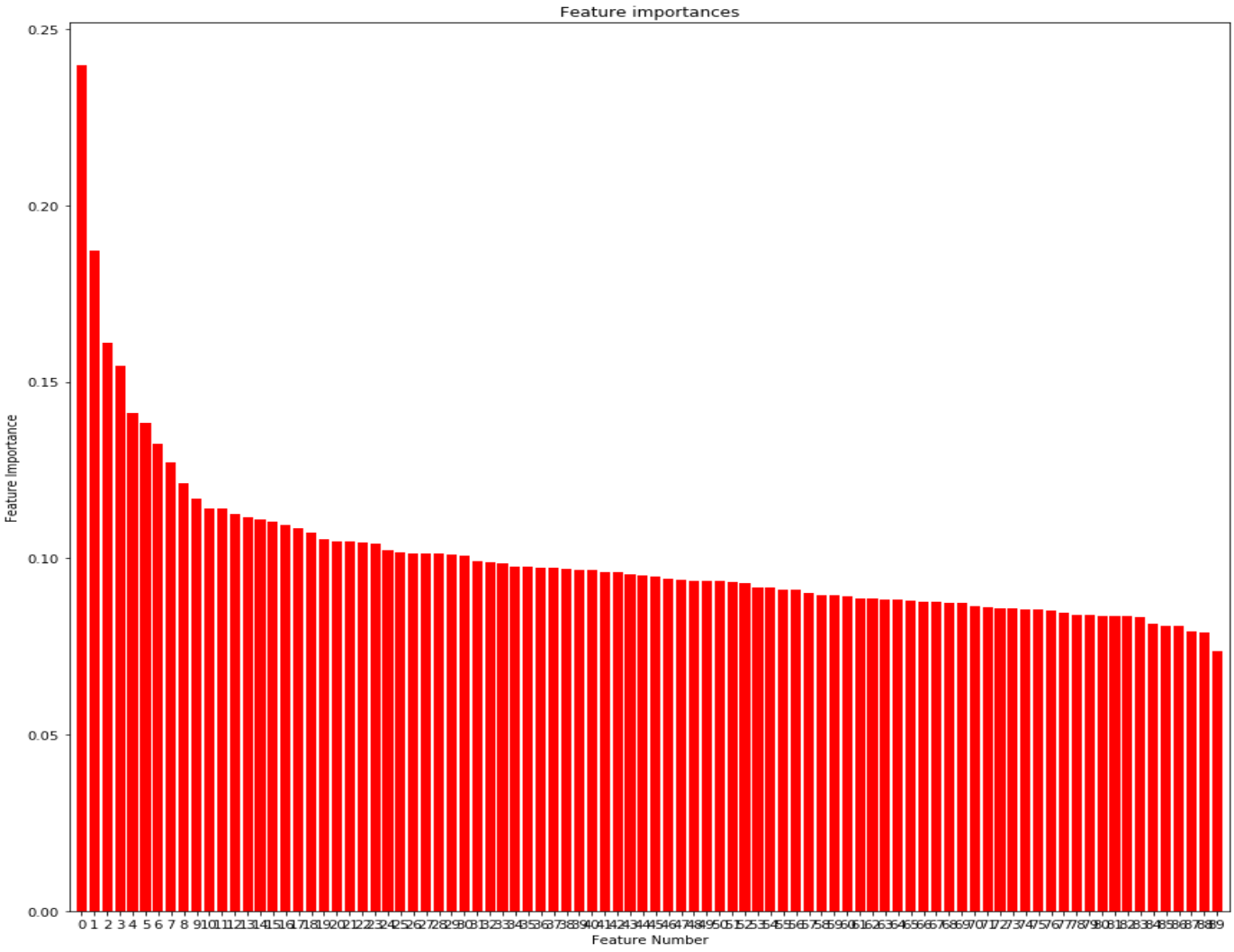


Graph of Variation of different accuracies (Training, Test and Out of Bag) with the number of Trees used in building the forest

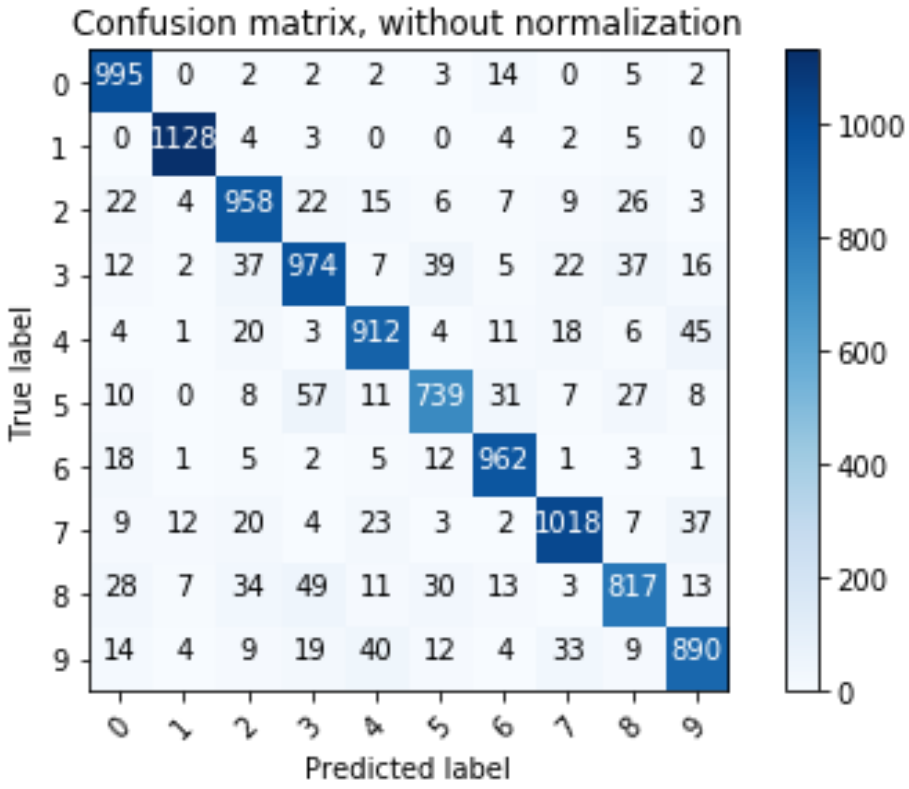
Using the above graph a good value of B=111 trees was chosen



Feature Importance of the respective features found in the selected trained forest

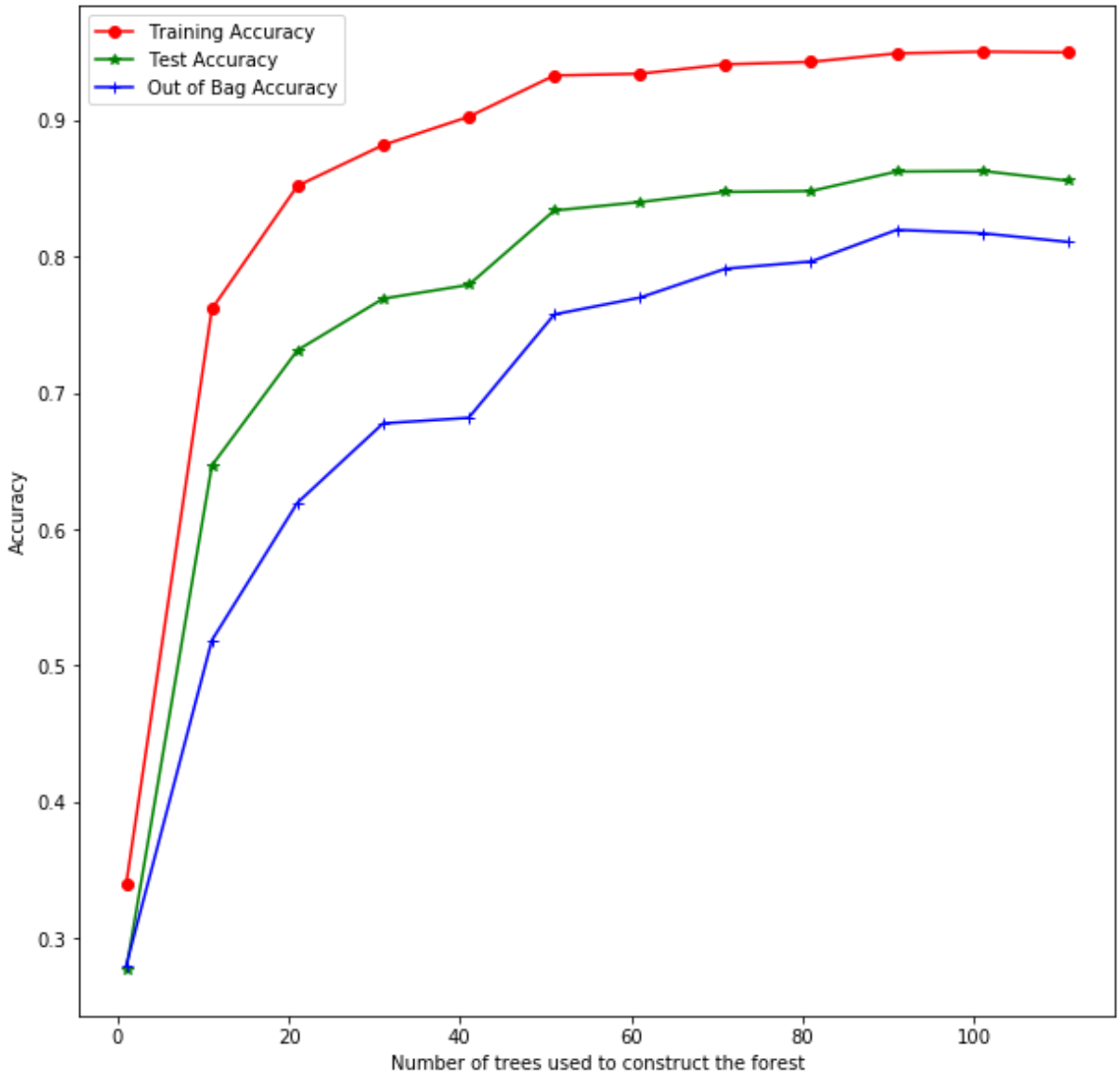


Feature Importance of the respective features sorted according to their Importance



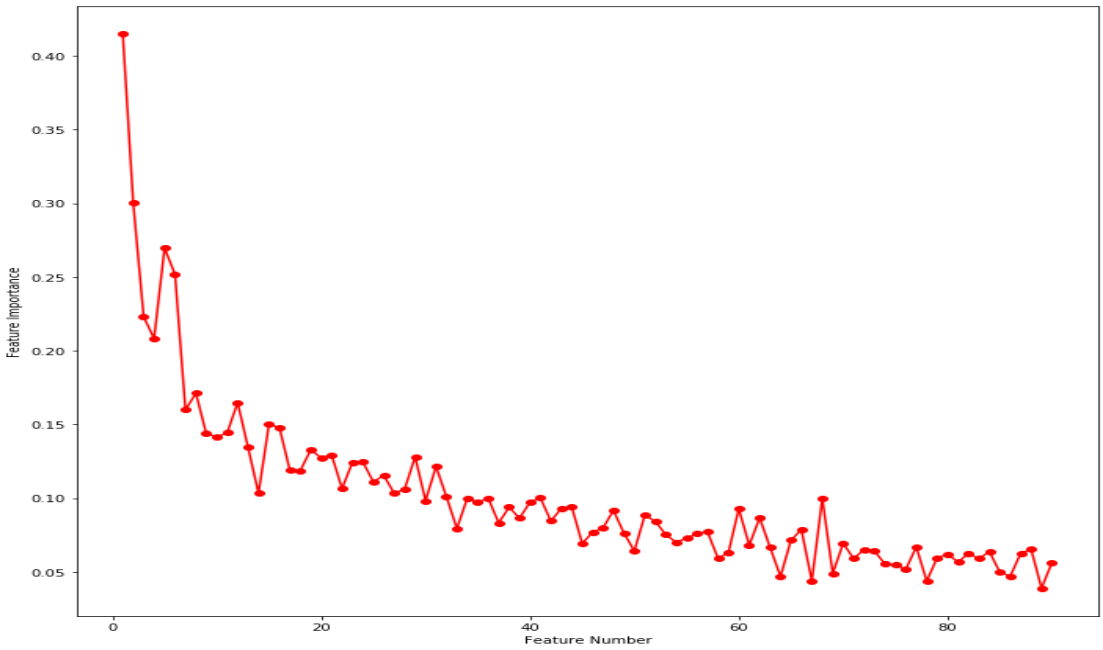
Confusion Matrix of the Selected Trained Forest on the Test Set

With decision tree grown to the maximum depth of 10

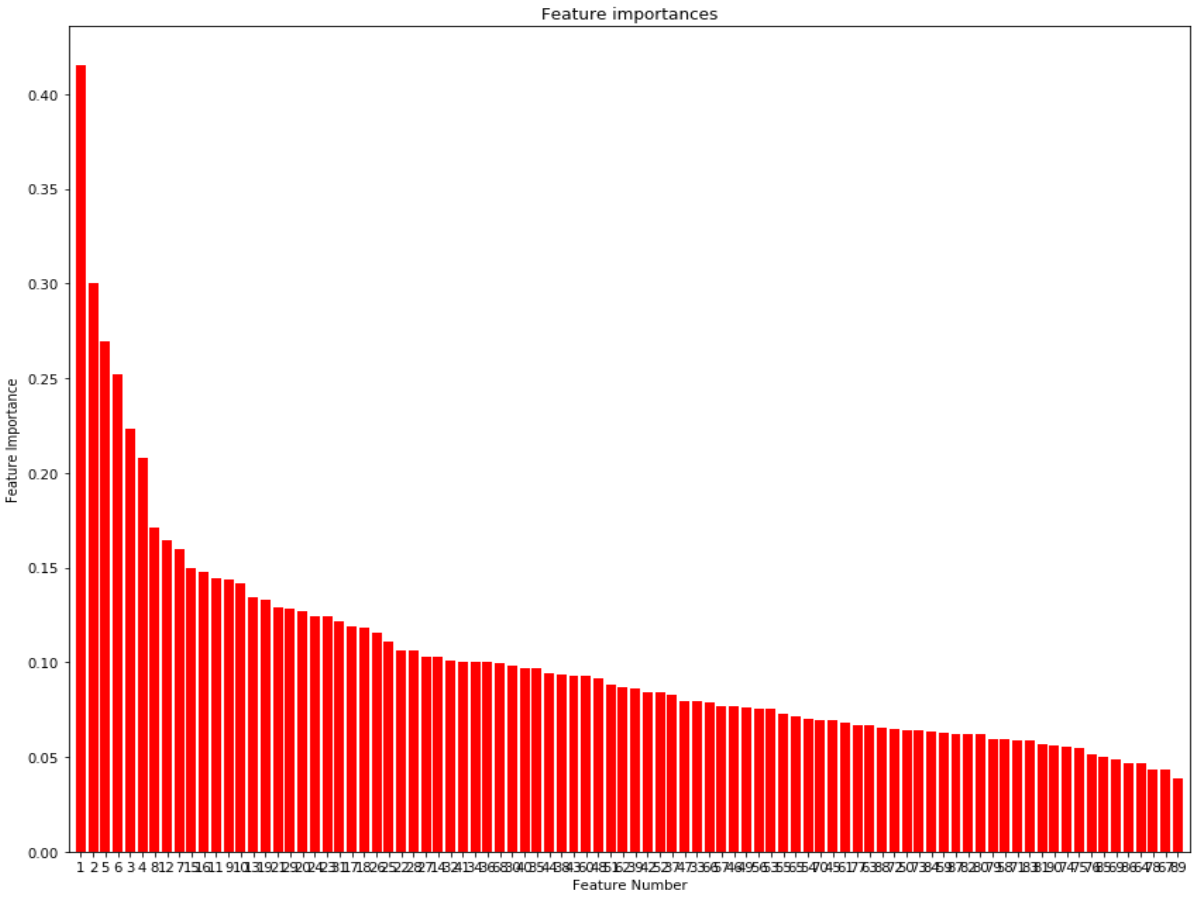


Graph of Variation of different accuracies (Training, Test and Out of Bag) with the number of Trees used in building the forest

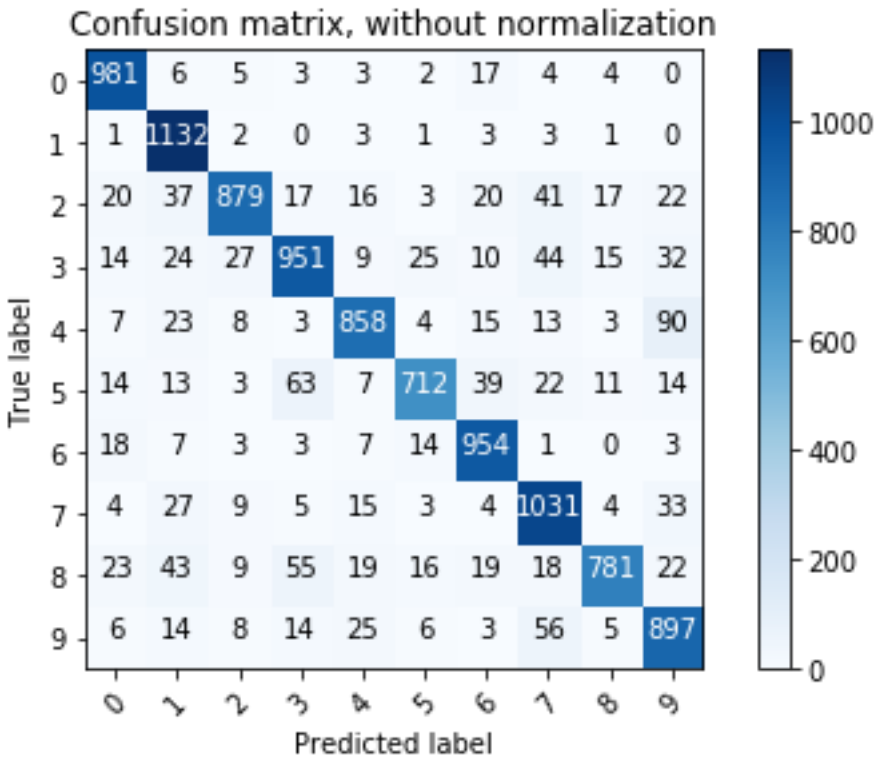
Using the above graph a good value of B=111 trees was chosen



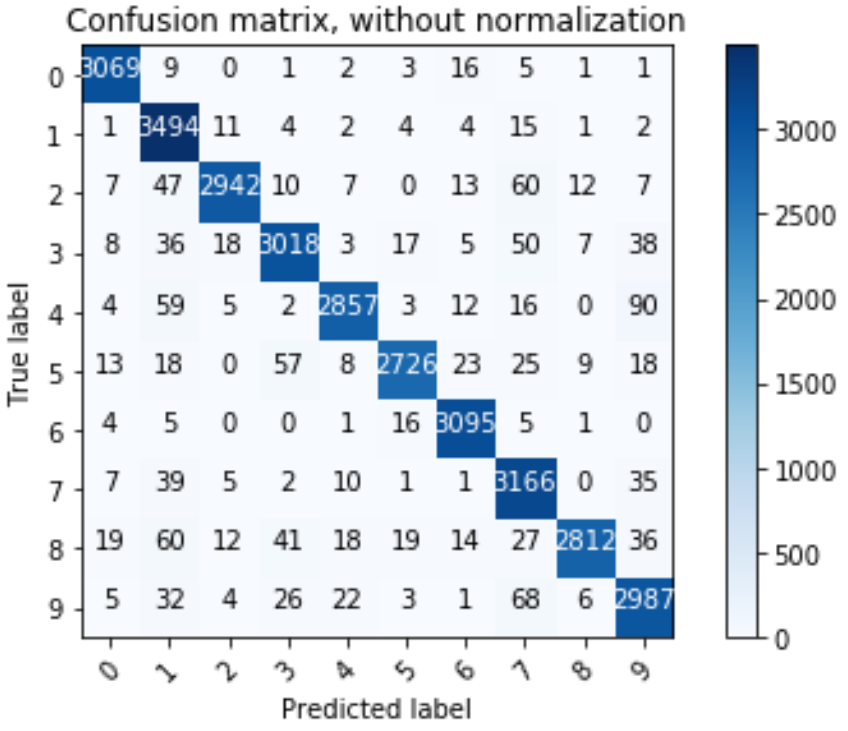
Feature Importance of the respective features found in the selected trained forest



Feature Importance of the respective features sorted according to their importance



Confusion Matrix of the Selected Trained Forest on the Test Set



Confusion Matrix of the Selected Trained Forest on the Training Set

As the number of trees used to build the forest was increased the Accuracies values started increasing. With the forests where trees were grown to the full extent Training Accuracy reached 100% mainly because of over fitting. We see that with the increase in the number of trees test performance improves and gradually stabilizes. We also observe that with the increase in the number of trees Out of Bag Accuracy approaches Test Accuracy.

The importance values where reflecting the work of PCA as most important Principal component's significance was more!

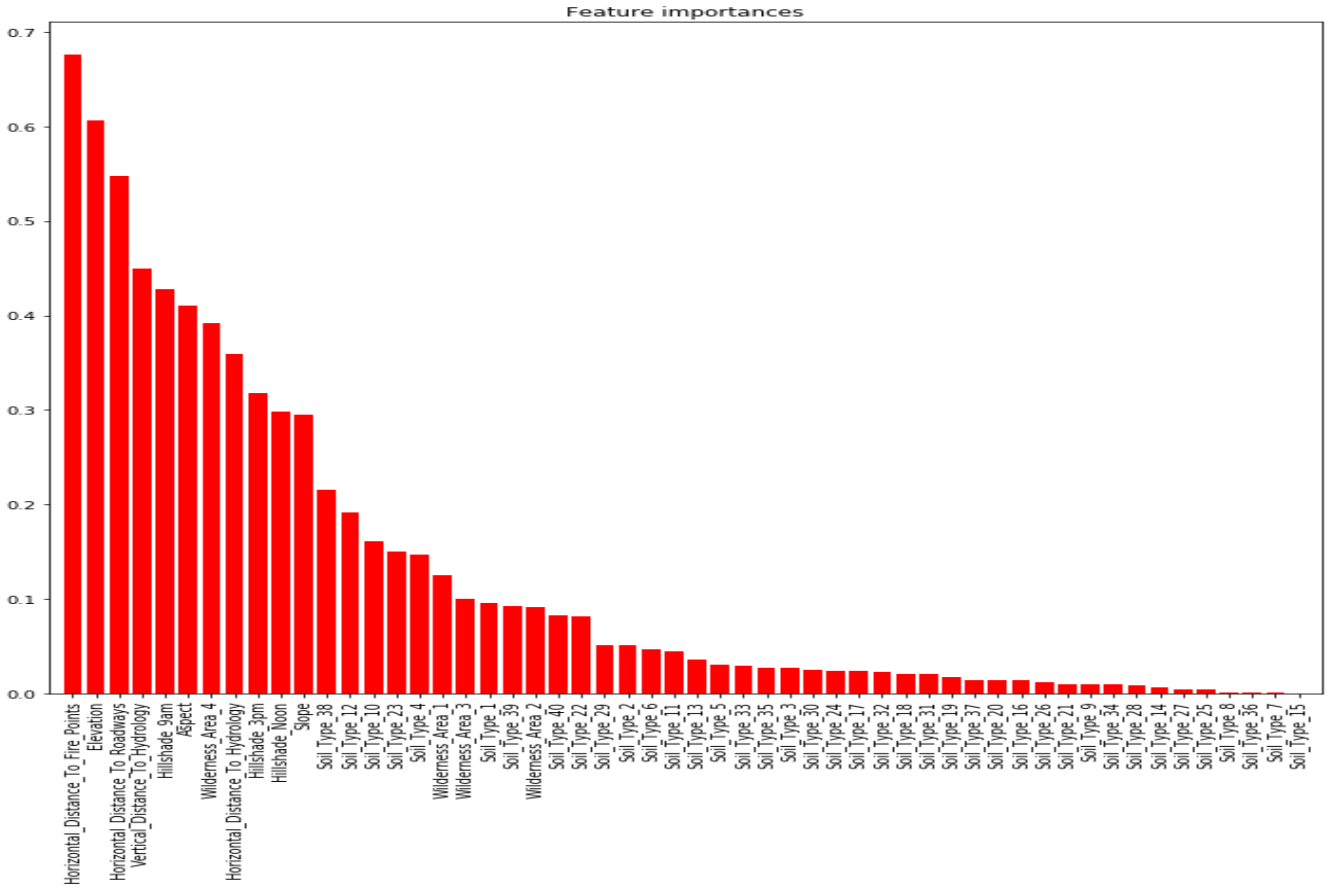
The Accuracy of the forest with trees max depth of 10 was almost same but slightly lower than the forests were trees were allowed to grow for the fullest range.

The Training accuracy also did not went high almost instantly signifying less over fitting as compared to the forests where the trees were grown to the fullest extent.

Though the Training Accuracy was very much high in forests allowed to grown to the fullest extent signifying over fitting.

COVERTYPE DATASET

In the cover type dataset initially uniform sampling was done at a gap of 20 and random forest with both trees grown to the fullest extent and with maximum depth was built. From these forests, the features that were the most significant/important were found and out of those 25 most important features were later used to with the forests again but this time with a uniform sampling at a gap of 10.

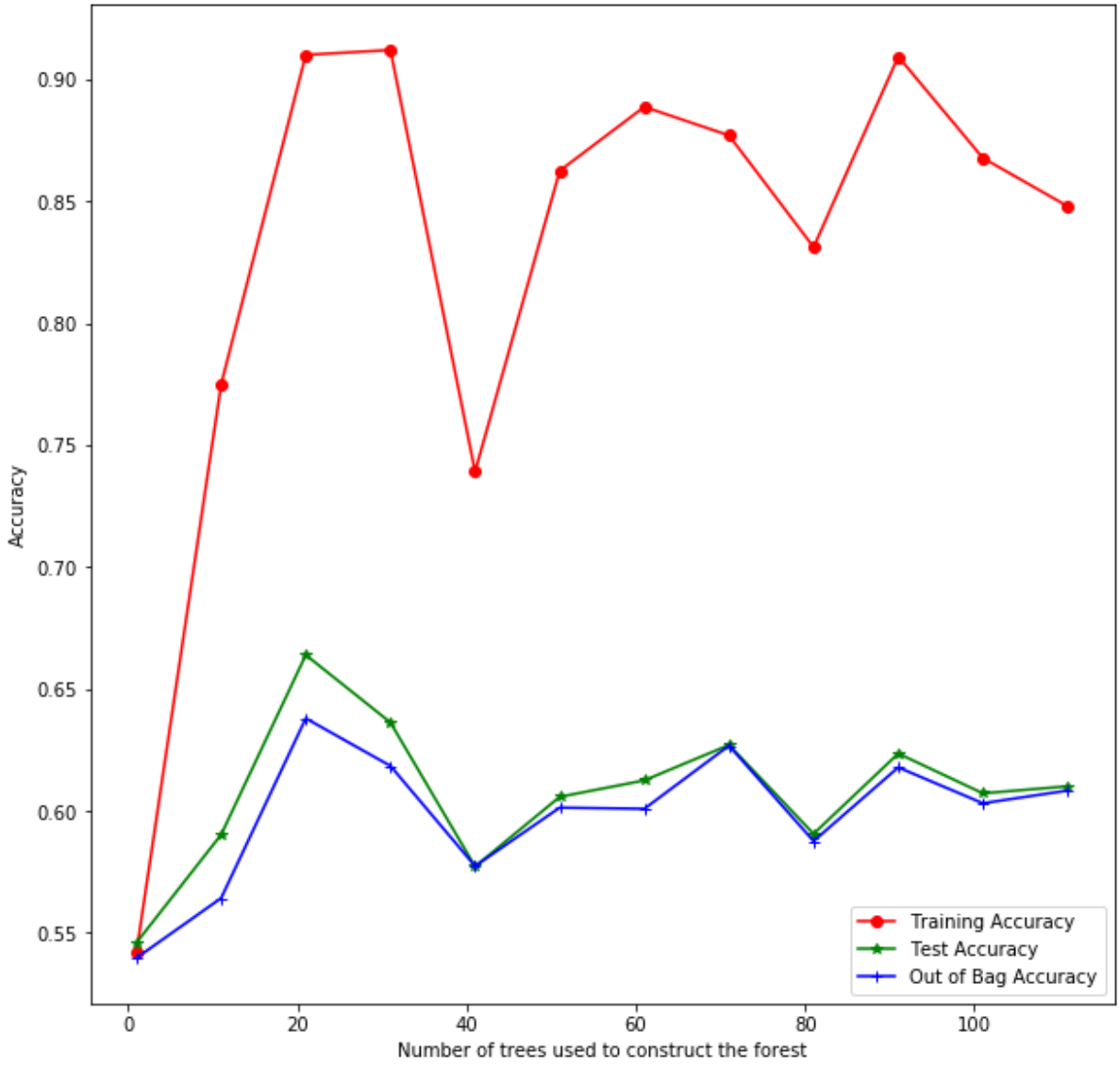


Sorted Feature Importance of the respected features found from the initial construction of the forests.

25 Important features selected were:

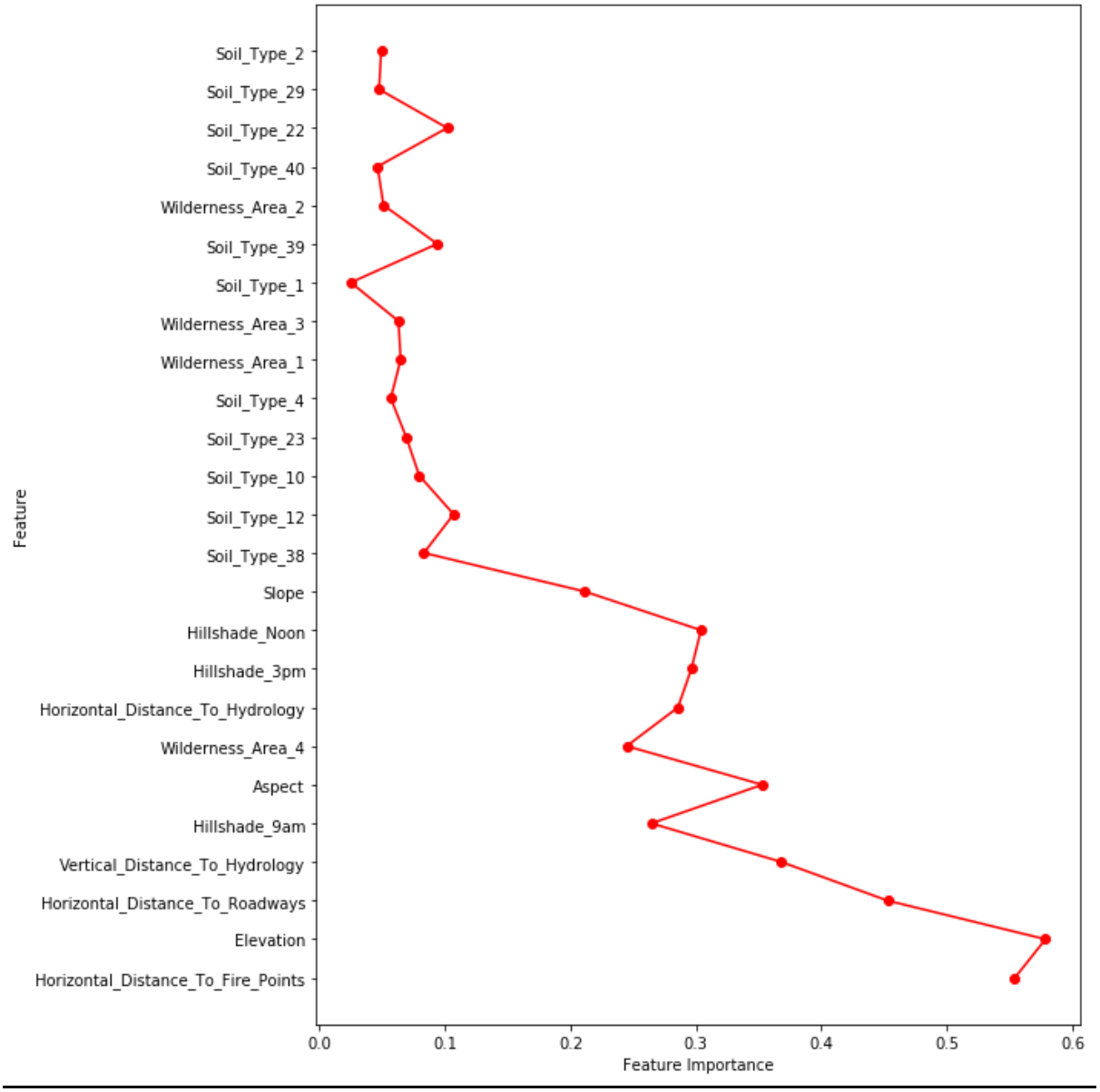
['Horizontal_Distance_To_Fire_Points','Elevation','Horizontal_Distance_To_Roadways','Vertical_Distance_To_Hydrology','Hillshade_9am','Aspect','Wilderness_Area_4','Horizontal_Distance_To_Hydrology','Hillshade_3pm','Hillshade_Noon','Slope','Soil_Type_38','Soil_Type_12','Soil_Type_10','Soil_Type_23','Soil_Type_4','Wilderness_Area_1','Wilderness_Area_3','Soil_Type_1','Soil_Type_39','Wilderness_Area_2','Soil_Type_40', 'Soil_Type_22', 'Soil_Type_29', 'Soil_Type_2']

With decision trees grown to the full extent

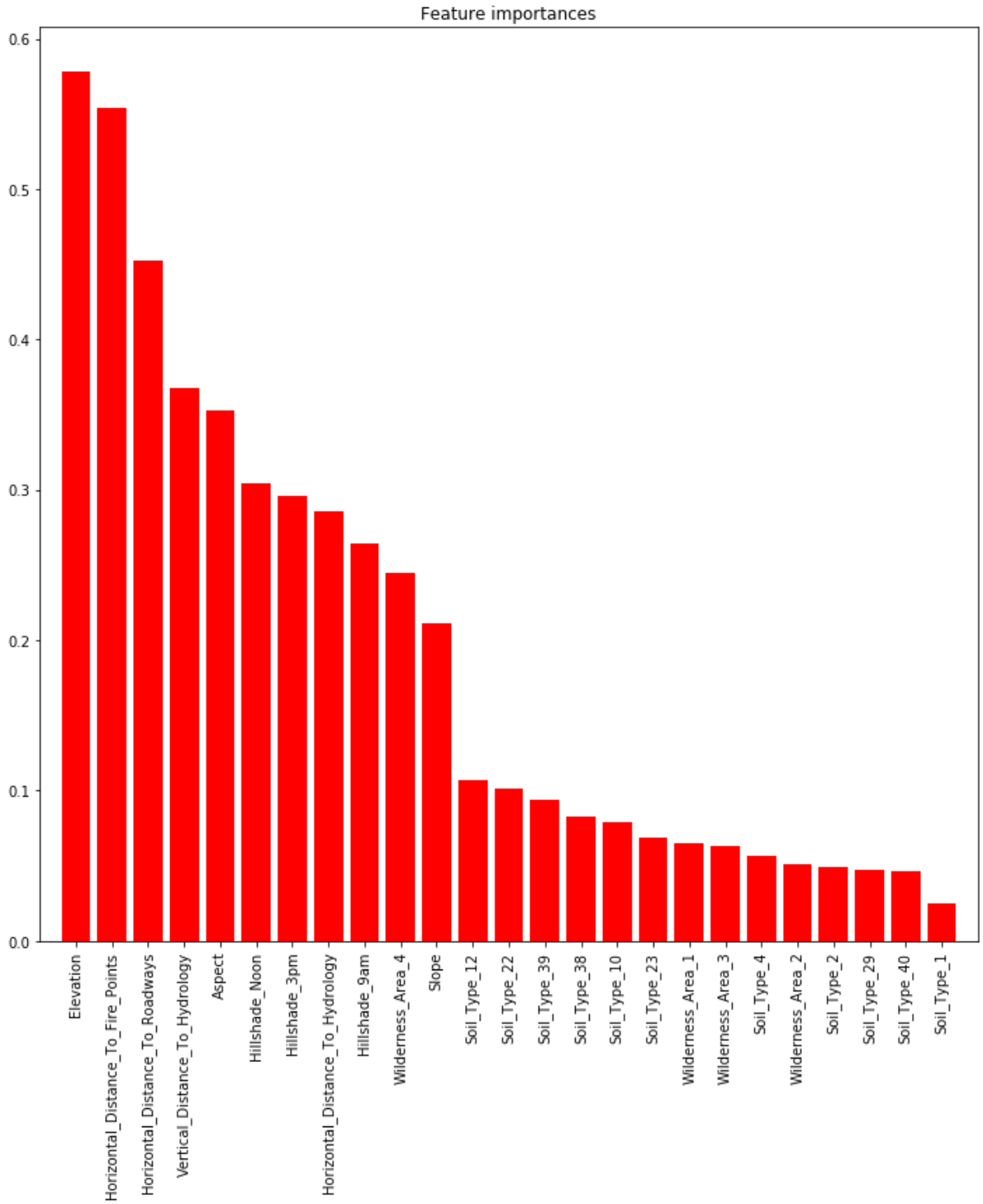


Graph of Variation of different accuracies (Training, Test and Out of Bag) with the number of Trees used in building the forest using only the 25 most important features.

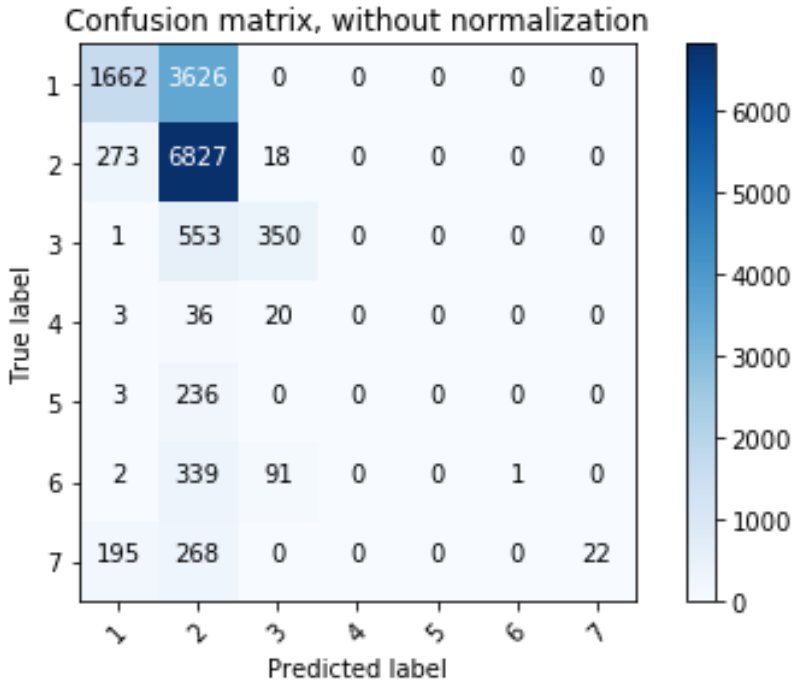
Using the above graph a good value of B=111 trees was chosen



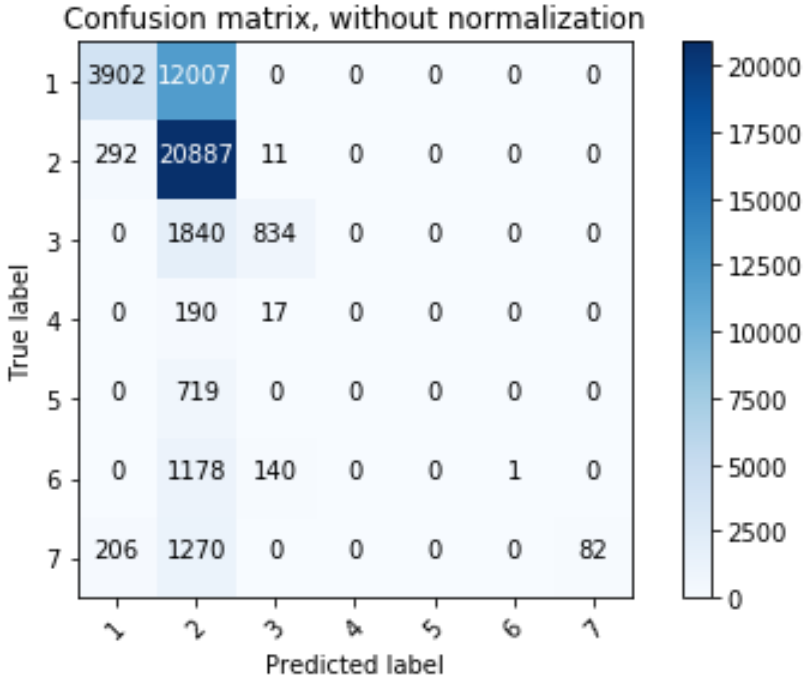
Feature Importance of the respective features found in the selected trained forest



Feature Importance of the respective features sorted according to their importance

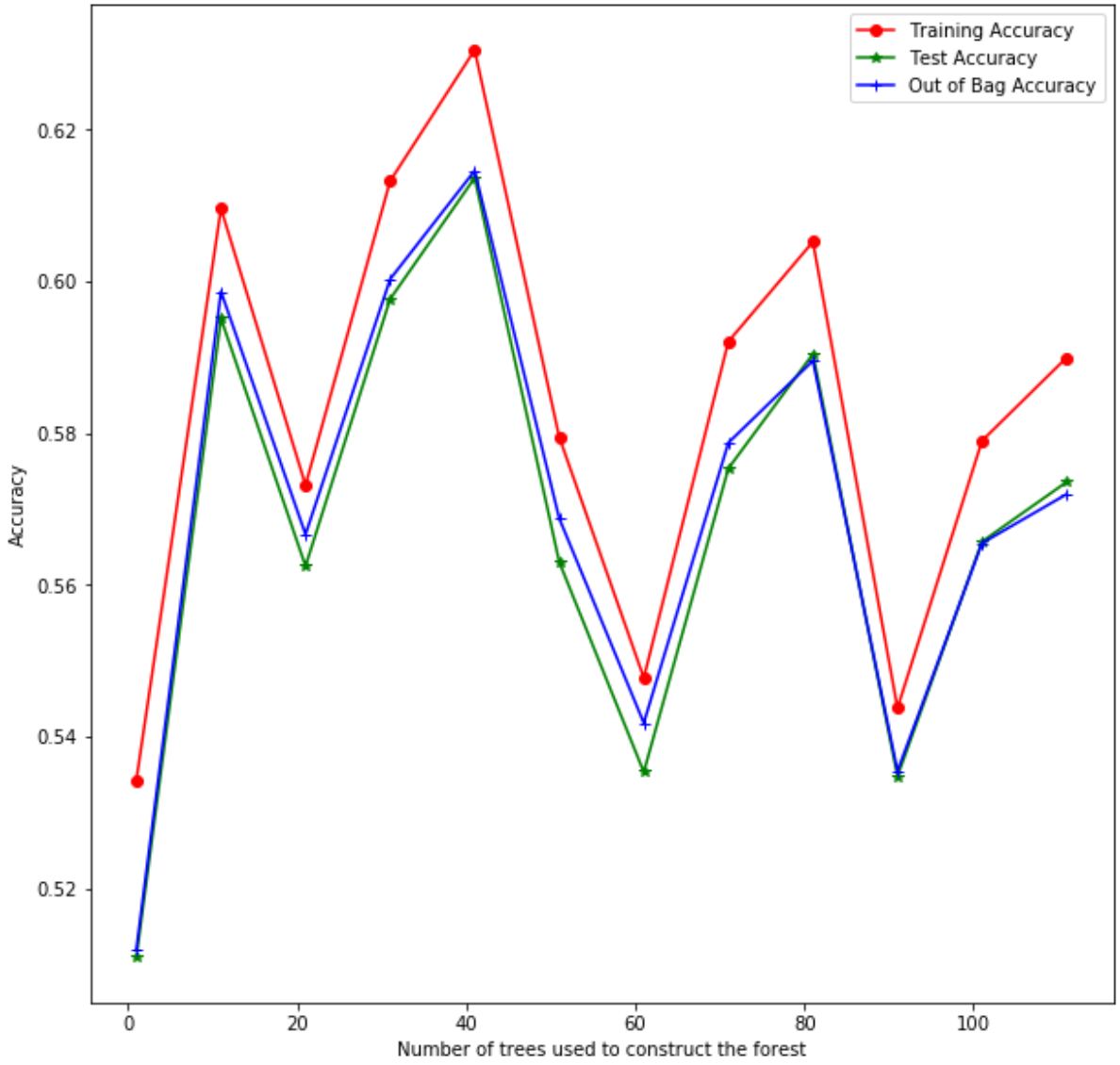


Confusion Matrix of the Selected Trained Forest on the Test Set



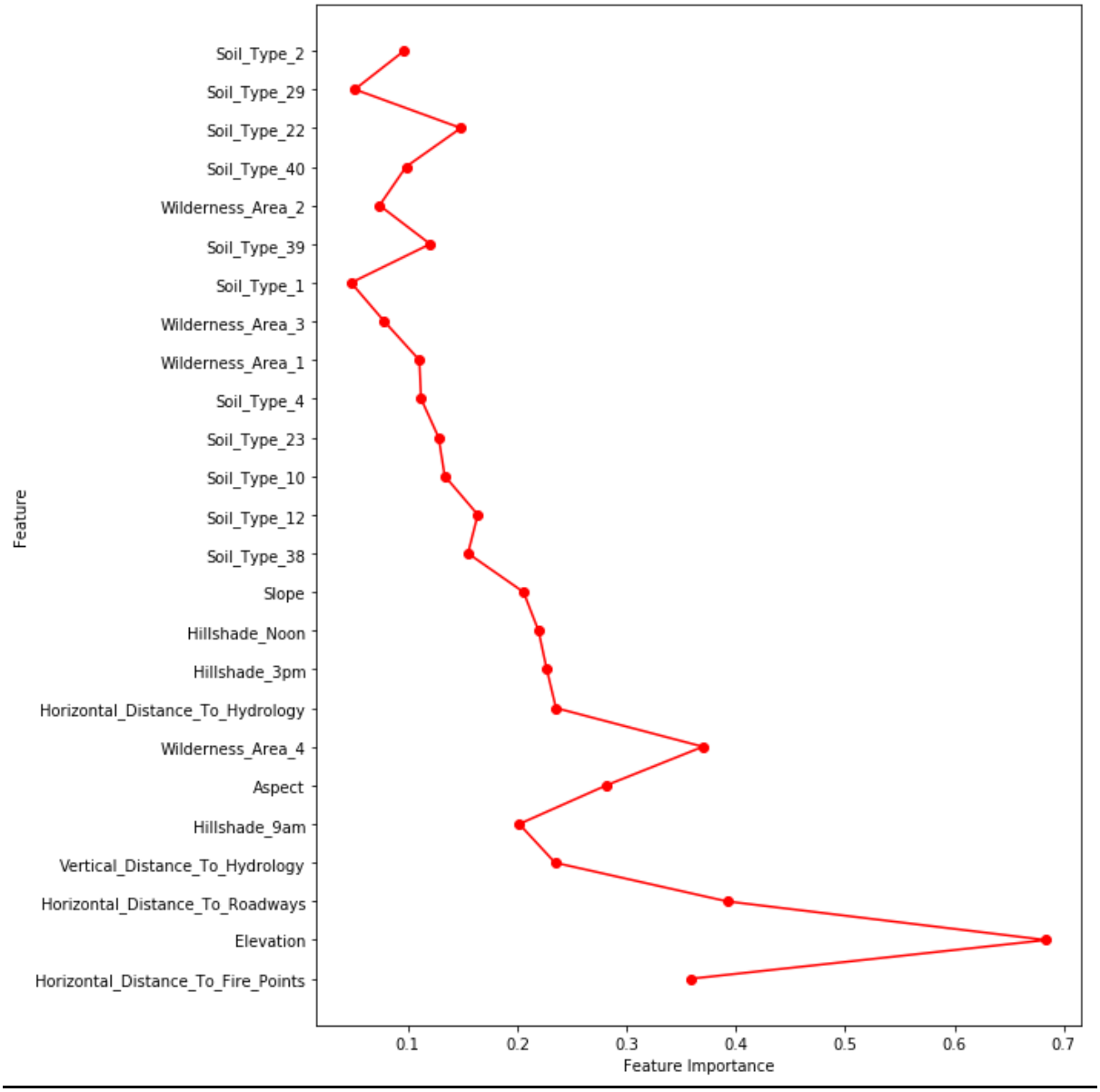
Confusion Matrix of the Selected Trained Forest on the Training Set

With decision tree grown to the maximum depth of 10

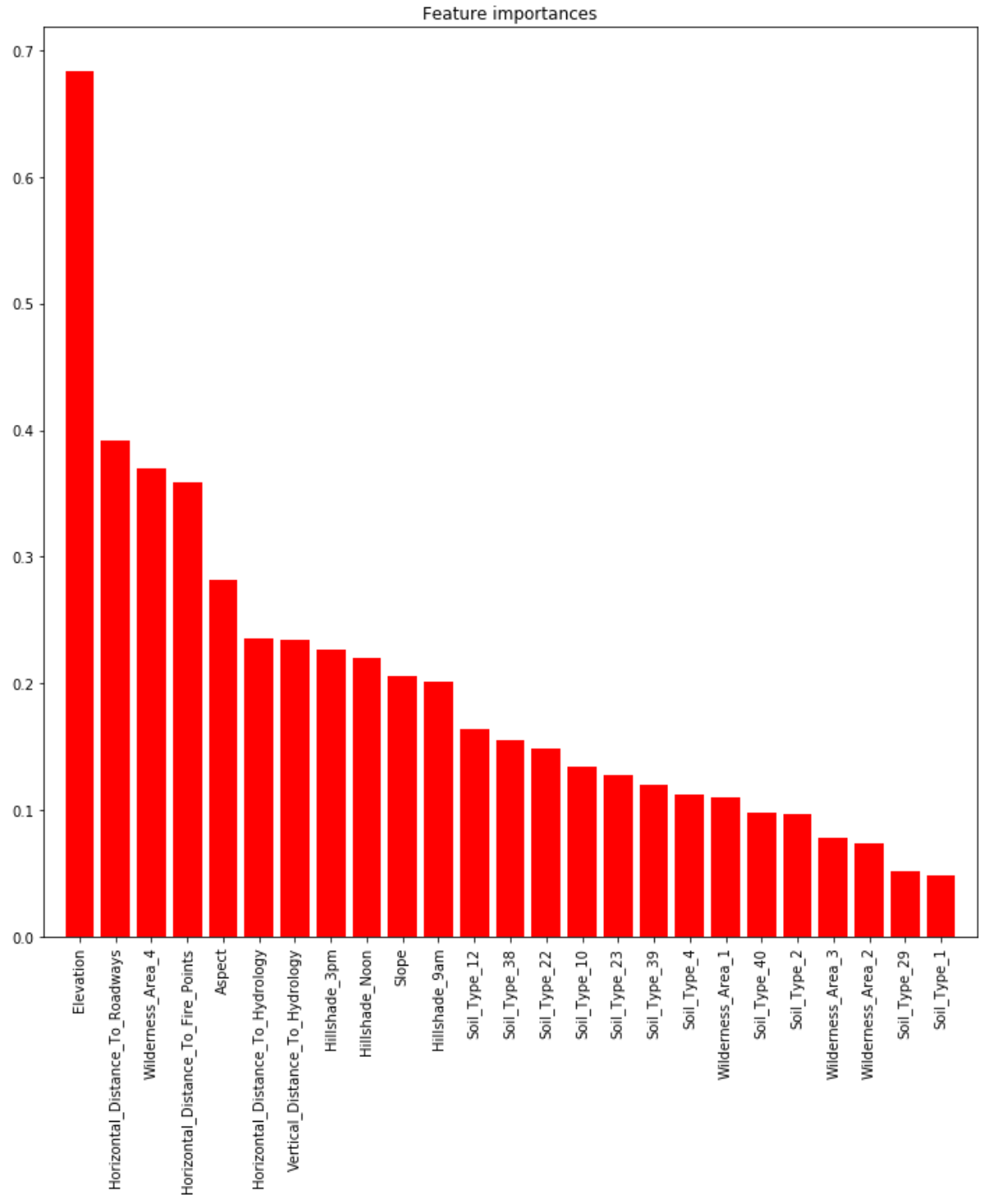


Graph of Variation of different accuracies (Training, Test and Out of Bag) with the number of Trees used in building the forest using only the 25 most important features.

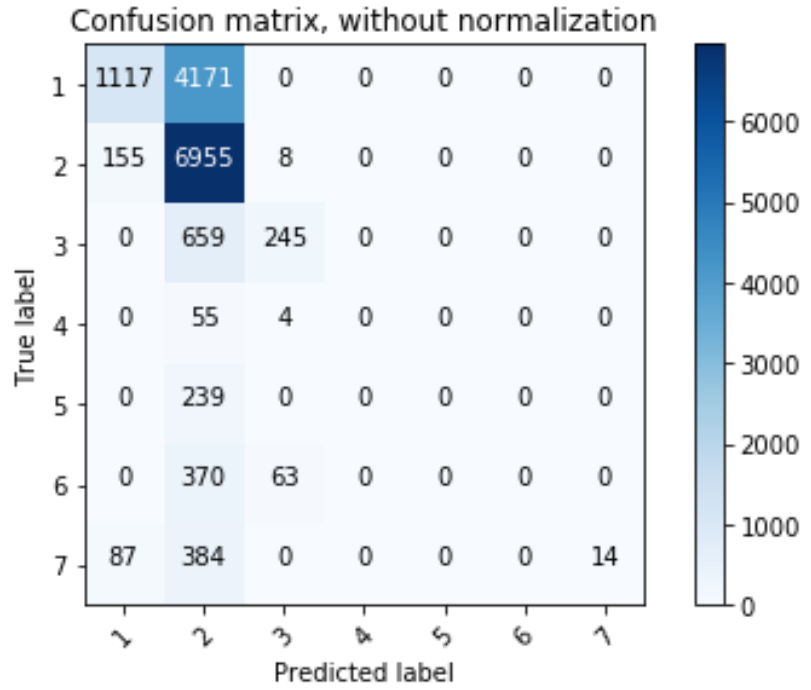
Using the above graph a good value of B=111 trees was chosen



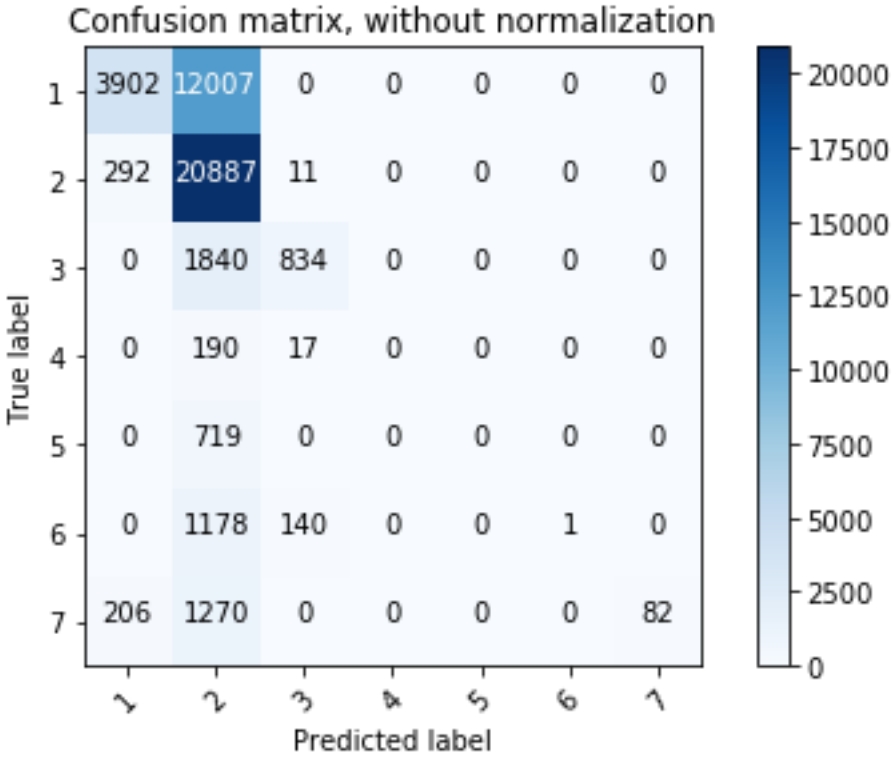
Feature Importance of the respective features found in the selected trained forest



Feature Importance of the respective features sorted according to their importance



Confusion Matrix of the Selected Trained Forest on the Test Set



Confusion Matrix of the Selected Trained Forest on the Training Set

As the number of trees used to build the forest was increased the Accuracies values started increasing. We also observe that with the increase in the number of trees Out of Bag Accuracy approaches Test Accuracy.

Few of the most important features of Cover type data found from the forests according their importance values were:

['Horizontal_Distance_To_Fire_Points','Elevation','Horizontal_Distance_To_Roadways','Vertical_Distance_To_Hydrology','Hillshade_9am','Aspect','Wilderness_Area_4','Horizontal_Distance_To_Hydrology','Hillshade_3pm','Hillshade_Noon','Slope',]

The Test Accuracy of the forest with trees max depth of 10 was almost same as compared to the forests where trees were allowed to grow for the fullest range.

The Training accuracy was much higher for forests grown over the fullest extent mainly because of over fitting as compared to the forests where the trees were grown to max depth of 10.