**Final Project Documentation**

**Forest Type Prediction**

**Project :**  In this project I had to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (as opposed to remotely sensed data). The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables such as wilderness areas and soil type.

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

**Data :** The data available for us in the training set is

**Elevation** - Elevation in meters  
**Aspect** - Aspect in degrees azimuth  
**Slope** - Slope in degrees  
**Horizontal\_Distance\_To\_Hydrology** - Horz Dist to nearest surface water features  
**Vertical\_Distance\_To\_Hydrology** - Vert Dist to nearest surface water features  
**Horizontal\_Distance\_To\_Roadways** - Horz Dist to nearest roadway  
**Hillshade\_9am** (0 to 255 index) – Hill shade index at 9am, summer solstice  
**Hillshade\_Noon** (0 to 255 index) – Hill shade index at noon, summer solstice  
**Hillshade\_3pm** (0 to 255 index) – Hill shade index at 3pm, summer solstice  
**Horizontal\_Distance\_To\_Fire\_Points** - Horz Dist to nearest wildfire ignition points  
**Wilderness\_Area** (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation  
**Soil\_Type** (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation  
**Cover\_Type** (7 types, integers 1 to 7) - Forest Cover Type designation

And in the testing set we have 565892 observations for which we have to sort out forest cover type

**EDA :** I have checked my testing and training set for missing values and in the testing set there were a few categorical variables like soil type and wilderness I had produced a graphical plot for understanding the percentage of those categorical variables against other data and for the continuous variables like hill shade, elevation, horizontal distance to road ways and other variables I have found out the max min and median values for them.

**Machine learning :**  I used cross validation to calculate the score and achieved an accuracy of .79 which is not bad. And I have built my model using Random Forest classifier with 600 n\_estimators I have also built another model with Extra Tree Classifier but I could come up with an accuracy of .78 even though the accuracy difference is minute in prediction we opt for the optimal accuracy that is the reason I chose to consider Random Forest model and predicted the cover type of all the forest areas in the test set then stored the whole log information in the result file which can be found in the folder after the execution is completed

**Post data analysis :** After predicting the cover types for all the areas in the test file I have checked for the missing values in the result, Which I did not find any. And produced a detailed graphical view of all the forest cover types in the result file.

