**Forest Cover Type Prediction**

For the last task, I attempted to foresee the forest spread type from entirely cartographic variables. The genuine woodland spread sort for a given 30 x 30 meter cell was resolved from US Forest Service (USFS) Region 2 Resource Information System information. Autonomous variables were acquired from the USGS and USFS. The information is in crude structure (not scaled) and contains double sections of information for subjective autonomous variables, for example, wild ranges and soil sort. This study region incorporates four wild ranges situated in the Arapaho & Roosevelt National Forests. These ranges speak to woodlands with insignificant human-created aggravations, so existing backwoods spread sorts are more an aftereffect of natural methodologies instead of timberland administration hones.

The training set consists of the following

**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) – Hillshade index at 9am, summer solstice  
**Hillshade\_Noon** (0 to 255 index) – Hillshade index at noon, summer solstice  
**Hillshade\_3pm** (0 to 255 index) – Hillshade 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

The testing set has 565892 observations for which we have to sort out forest cover type.

**Exploratory Data analysis**

After checking 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 I have calculated the max min and median values for them continuous variables like hill shade, elevation, horizontal distance to road ways and other variables.

**Machine learning**

I achieved an accuracy of .79 using cross validation. 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.

**Post EDA**

After predicting the cover types for all the areas in the test file I couldn't find any missing values in the result. And produced a detailed graph of all the forest cover types in the result file.

