INTERIM PROJECT REPORT



Forest Cover type Prediction



**CHAPTER 1:INDUSTRY REVIEW**

* 1. **Literature Survey:**
* 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.

Original Owners of Database:

Jock A. Blackard and Dr. Denis J. Dean.

Remote Sensing and GIS Program

Department of Forest Sciences

College of Natural Resources

Colorado State University

Fort Collins, CO 80523

Donors of database: **3**

1. Jock A. Blackard Email

2. Dr. Denis J. Dean Email

3. Dr. Charles W. Anderson Email

**Relevant Papers**:

Blackard, Jock A. and Denis J. Dean. 2000. "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from

Cartographic Variables." Computers and Electronics in Agriculture 24(3):131-151. Visit Site

Blackard, Jock A. and Denis J. Dean. 1998. "Comparative Accuracies of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables." Second Southern Forestry GIS Conference. University of Georgia. Athens, GA. Pages 189-199.

**CHAPTER 2: DATASET AND DOMAIN**

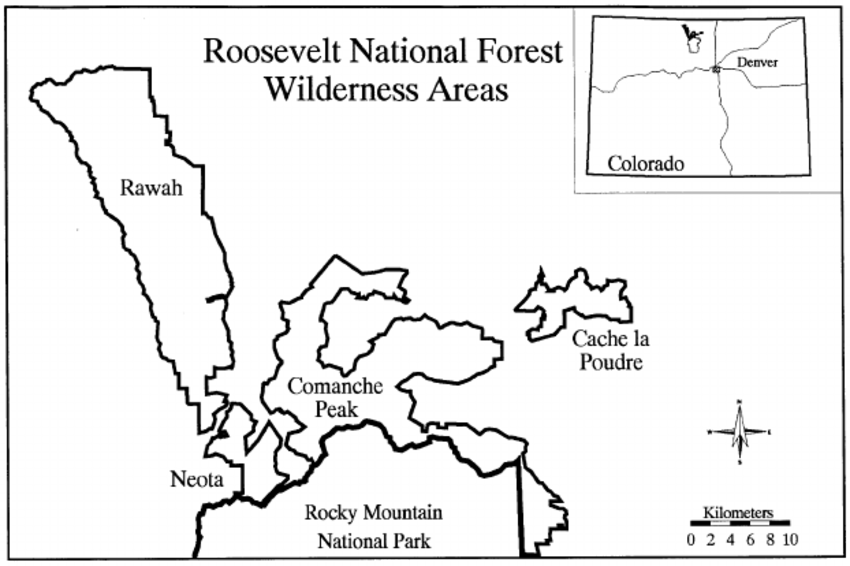
**2.1 Domain Background**:

* Given elevation, hydrologic, soil, and sunlight data can we predict what type of tree would be in a small patch of forest? This project attempts to predict the predominant type of tree in sections of wooded area.

This study area includes 4 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 process rather than forest management practices.
* Each observation is 30m x 30m forest cover type determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from the data originally obtained from US Geological Survey (USGS) and USFS data.



**2.2 Problem Statement:**

* We have been given a total of 54 attributes, these attributes contain Binary and Quantitative attributes, and we need to predict which *Forest Cover-Type* is it from the given features.

**2.3 NEED FOR STUDY:**

* Understanding forest composition is a valuable aspect of managing the health and vitality of our wilderness areas. Classifying cover type can help further research regarding forest fire susceptibility and de/reforestation concerns.
* Forest cover type data is often collected by hand or computed using remote sensing techniques, e.g. satellite imagery. Such processes are both time and resource intensive. In this project, we aim to predict forest cover type using cartographic data and a variety of classification algorithms.

**2.4 Data Dictionary:**

This dataset has been taken from UCI Machine Learning Repository.

|  |  |
| --- | --- |
| Dataset Info*: No. of Instances* | 581,012 |
| *No. of Attributes (Features)* | 54 |
| *Associate Task* | Classification |
| *Dataset Characteristic* | Multivariate |
| *Attribute Characteristic* | Categorical, Integer |
| *Missing Value* | None |
| *Area* | Life |
| *Target Variable* | Forest Cover Type |

* Our Dataset consists of more than half a million observation and has 54 attributes or features to help us predict type of Forest Cover which is our target or output variable. The target variable has 7 different classes hence making this a Multi-Class Classification problem.

**2.4.1 Variable categorization**:

10 Quantitative variable, 4 Binary Variable (*Wilderness Area*) and other 40 Binary Variable (*Soil Type*). Which makes a total of 54 Variable/Features/Attributes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Name** | **Feature Description** | **Min Value** | **Max**  **Value** | **Std Dev.** |
| Elevation | Elevation in metres | 1859 | 3858 | 279 |
| Aspect | Aspect in degrees azimuth | 0 | 360. | 111. |
| Slope | Slope in degrees | 0 | 66. | 7.48 |
| Horizontal\_Distance\_To\_Hydrology | Horizontal distance to the nearest surface water features | 0 | 1397 | 212.5 |
| Vertical\_Distance\_To\_Hydrology | Vertical distance to the nearest surface water features | -173 | 601 | 58.2 |
| Horizontal\_Distance\_To\_Roadways | Horizontal distance to the nearest roadway | 0 | 7117 | 1559 |
| Hillshade\_9am | Hillshade index at 9 AM, summer solstice | 0 | 254 | 26 |
| Hillshade\_noon | Hillshade index at 12 PM (noon, summer solstice | 0 | 254 | 19 |
| Hillshade\_3pm | Hillshade index at 3 AM, summer solstice | 0 | 254 | 38 |
| Horizontal\_Distance\_To\_Fire\_Points | Horizontal distance to the nearest wildfire ignition points | 0 | 7173. | 1324 |

***Table 1*** shows the numerical features along with their statistical parameters.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Number of Categorical Values** |
| Wilderness\_Area (4) | Wilderness area designation | 4 |
| Soil\_Type (40) | Soil type designation | 40 |
| CoverType(target) | Forest cover type (7) | 7 |

***Table 2:***  shows the Categorical features.

* As we can see from the above table, the *wilderness area* has 4 columns, and these columns are binary meaning it can only be present one for each observation. Here we have 4 types of Wilderness area, and it can have only one of these in each observation/instance that are given to us. Same goes for *Soil Type* feature.

***2.4.2 Wilderness areas* details:**

|  |  |
| --- | --- |
| *Wilderness\_Area1* | Rawah Wilderness Area |
| *Wilderness\_Area2* | Neota Wilderness Area |
| *Wilderness\_Area3* | Comanche Wilderness Area |
| *Wilderness\_Area4* | Cache La Poudre Wilderness Area |

* Above is the table, on the left side of it are the column names in the dataset which gives us the information of wilderness and to the right of it are the names of the wilderness areas.

Some Background Information for these 4 *Wilderness Area*:

* *Neota* probably has the highest mean elevation values of the 4 Wilderness Areas. *Rawah* and *Comanche* would have a lower mean elevation value, while *Cache la Poudre* would have the lowest mean elevation value.

***2.4.3 Soil Type* Feature Details:**

* Soil Type feature has 40 columns of it, meaning there are 40 types of Soils collected from 4 Wilderness Areas in the *Roosevelt National Forest.*

|  |  |
| --- | --- |
| 1 | Cathedral family - Rock outcrop complex, extremely stony |
| 2 | Vanet - Ratake families complex, very stony |
| 3 | Haploborolis - Rock outcrop complex, rubbly |
| 4 | Ratake family - Rock outcrop complex, rubbly |
| 5 | Vanet family - Rock outcrop complex, rubbly |
| 6 | Vanet - Wetmore families - Rock outcrop complex, stony |
| 7 | Gothic family |
| 8 | Supervisor - Limber families complex |
| 9 | Troutville family, very stony |
| 10 | Bullwark - Catamount families - Rock outcrop complex, rubbly |
| 11 | Bullwark - Catamount families - Rock land complex, rubbly |
| 12 | Legault family - Rock land complex, stony |
| 13 | Catamount family - Rock land - Bullwark family complex, rubbly |
| 14 | Pachic Argiborolis - Aquolis complex |
| 15 | *unspecified in the USFS Soil and ELU Survey* |
| 16 | Cryaquolis - Cryoborolis complex |

|  |  |
| --- | --- |
| 17 | Gateview family - Cryaquolis complex |
| 18 | Rogert family, very stony |
| 19 | Typic Cryaquolis - Borohemists complex |
| 20 | Typic Cryaquepts - Typic Cryaquolls complex |
| 21 | Typic Cryaquolls - Leighcan family, till substratum complex |
| 22 | Leighcan family, till substratum, extremely bouldery |
| 23 | Leighcan family, till substratum, - Typic Cryaquolls complex. |
| 24 | Leighcan family, extremely stony |
| 25 | Leighcan family, warm, extremely stony |
| 26 | Granile - Catamount families complex, very stony |
| 27 | Leighcan family, warm - Rock outcrop complex, extremely stony |
| 28 | Leighcan family - Rock outcrop complex, extremely stony |
| 29 | Como - Legault families complex, extremely stony |
| 30 | Como family - Rock land - Legault family complex, extremely stony |
| 31 | Leighcan - Catamount families complex, extremely stony |
| 32 | Catamount family - Rock outcrop - Leighcan family complex, extremely stony |
| 33 | Leighcan - Catamount families - Rock outcrop complex, extremely stony |
| 34 | Cryorthents - Rock land complex, extremely stony |
| 35 | Cryumbrepts - Rock outcrop - Cryaquepts complex |
| 36 | Bross family - Rock land - Cryumbrepts complex, extremely stony |
| 37 | Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony |
| 38 | Leighcan - Moran families - Cryaquolls complex, extremely stony |
| 39 | Moran family - Cryorthents - Leighcan family complex, extremely stony |
| 40 | Moran family - Cryorthents - Rock land complex, extremely stony |

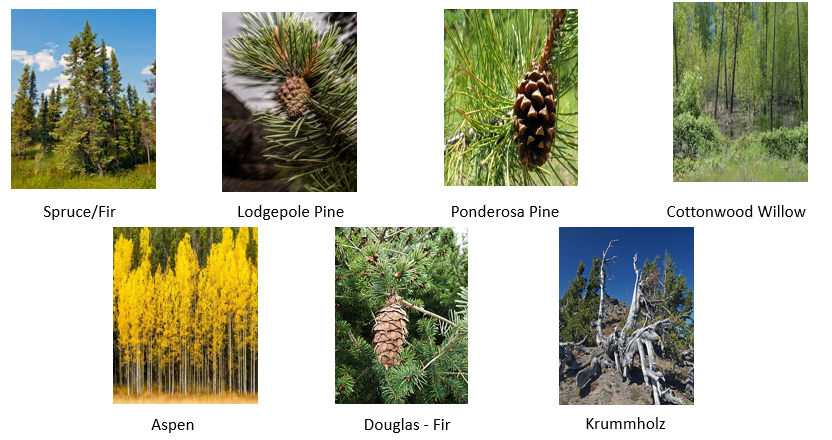
**2.4.4 Forest Cover Type**

**(The variable for our prediction):**

* This is the variable which we are going to predict and it has only one column which represents integer values from 1 to 7, where these digits represent type of forest cover for the observations.
* This is the variable which is not *one-hot encoded* like *Soil Type* and *Wilderness Area*, that’s why it doesn’t have 7 columns to represent each class for the observations.

Now let’s look at the names of these types of Forest Cover:

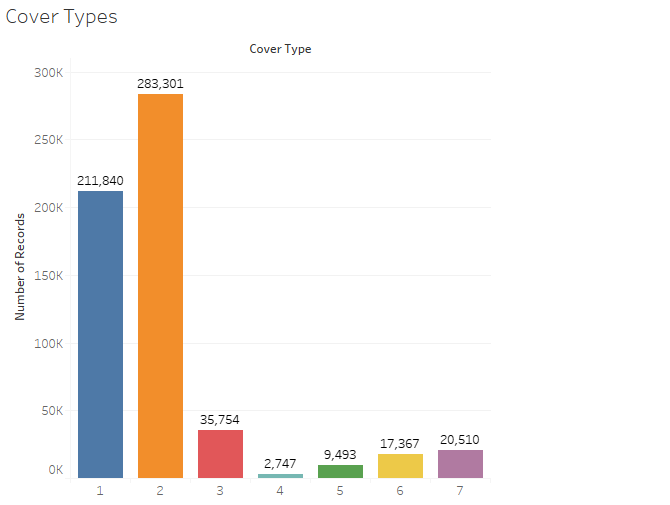
|  |  |
| --- | --- |
| *1* | Spruce / Fir |
| *2* | Lodgepole Pine |
| *3* | Ponderosa Pine |
| *4* | Cottonwood / Willow |
| *5* | Aspen |
| *6* | Douglas-fir |
| *7* | Krummholz |

**

***CHAPTER 3:* DATA Exploration**

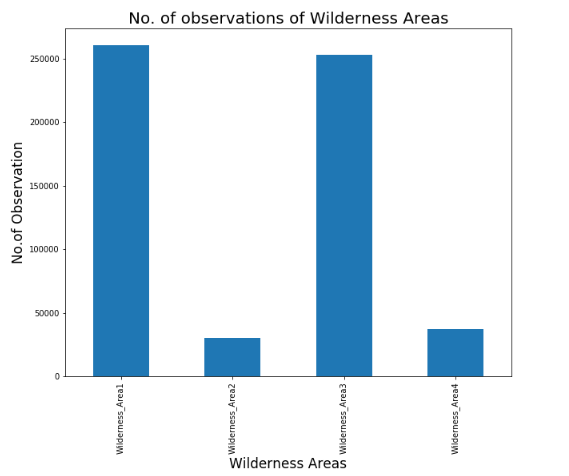
**3.1 The Distribution of Forest Type in:**

* The distribution of these classes is not equal. Spruce and Lodgepole have the most while Cottonwood has the least records.
* Having such distribution will not give us appropriate results because of unequal amount of distribution.
* Also to note here that every observation is assigned to some class of forest type and no observation is empty with that information.

**

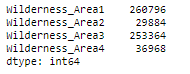
|  |  |
| --- | --- |
| No. of records of Spruce / Fir | 211 840 |
| No. of records of Lodgepole Pine | 283 301 |
| No. of records of Ponderosa Pine | 35 754 |
| No. of records of Cottonwood / Pillow | 2 747 |
| No. of records of Aspen | 9 493 |
| No. of records of Douglas-Fir | 17 367 |
| No. of records of Krummholz | 20 510 |
| Total Records | 581 012 |

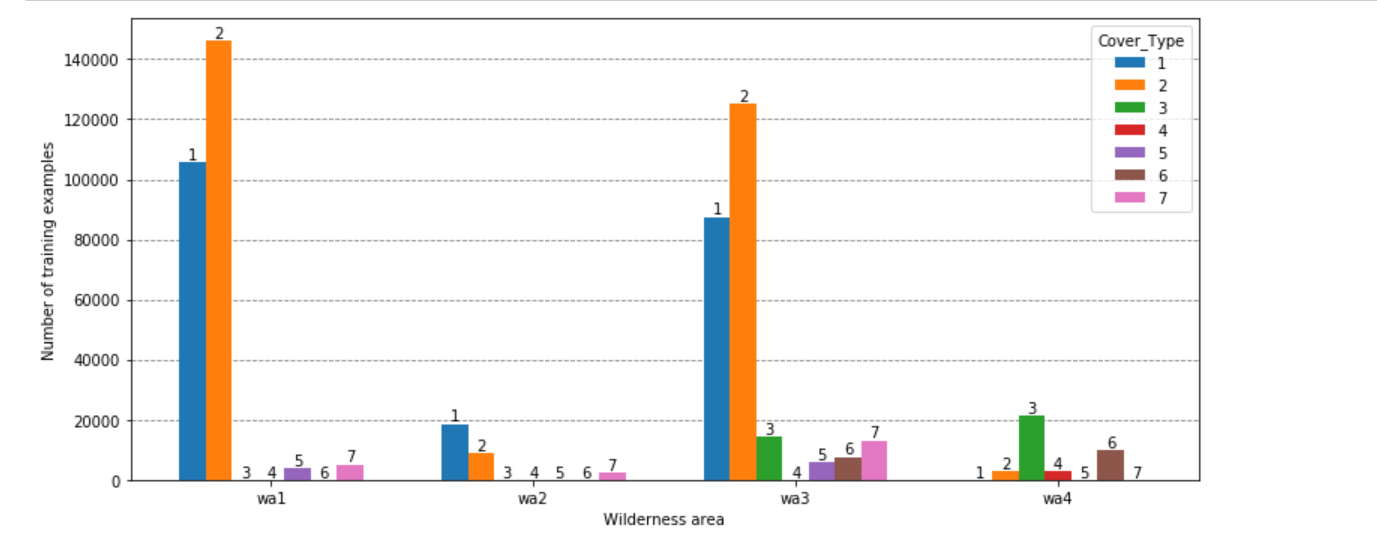
**3.2 The Distribution of Wilderness area:**



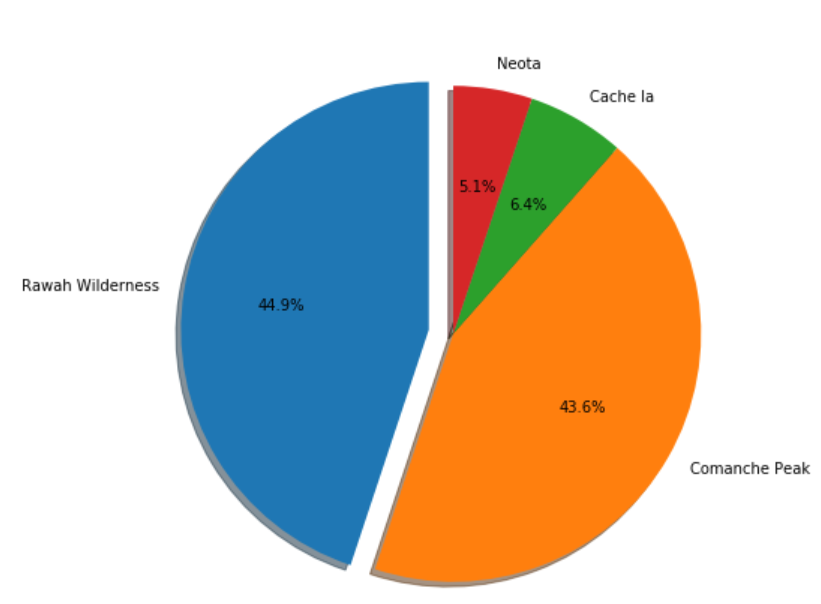
* Wilderness\_Area1 has the most presence followed by Wilderness\_Area3, both have quite close observations and so were their mean value.
* Wilderness\_Area2 having the least observation.

**Count of wilderness area:**



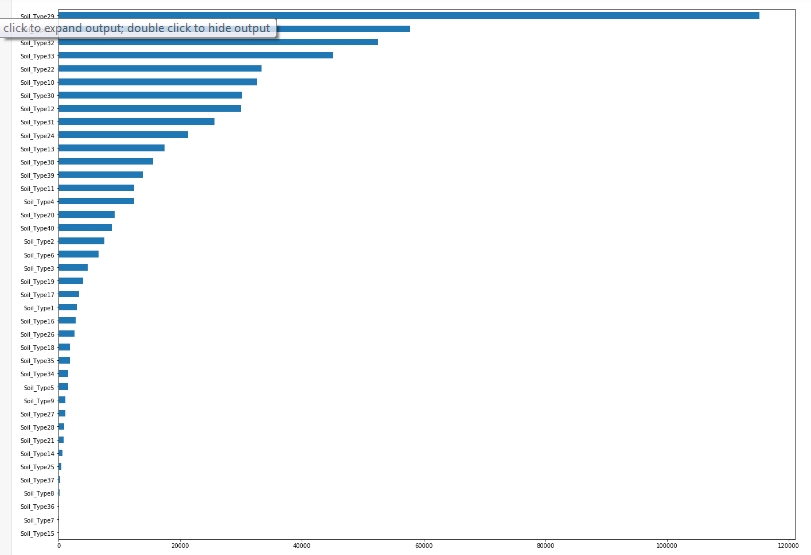


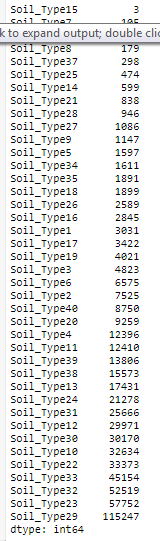
* Image above shows the distribution of each cover type in different wilderness area. As you can see cover type 2 is occurring in almost in all the wilderness area and some of them are occurring in only specific wilderness area.



* Shows the distribution of the wilderness area in the data the Rawah Wilderness and Comanche Peak wilderness area is occurring 87% of the data.

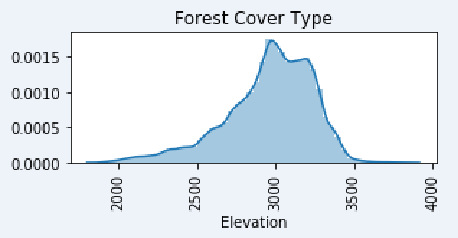
**3.3 SOIL TYPE distribution:**

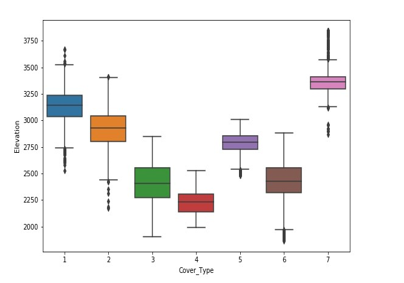
****



* The least observation are of Soil\_Type15 of 3
* Soil\_Type29 has the highest

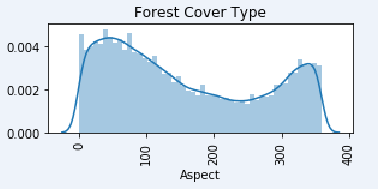
#### **3.4 Elevation:**

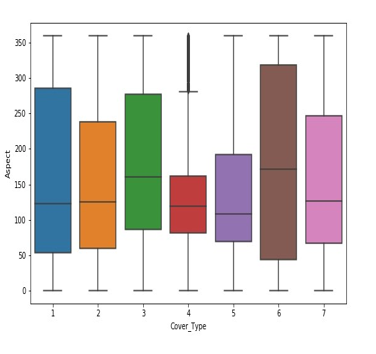
****



* From the diagram we can say that it is a bimodal graph so we can confirm that data is not normally distributed. From the boxplot we can get the inference that median value are different for all the Cover Type may be with this might be one of the important feature to distinguish between the different Cover type. For Some of the Cover Type as you can see from the figure above there are some extreme values or some outliers so we need identify whether they are extreme values or if outliers then have to do proper computation methods.

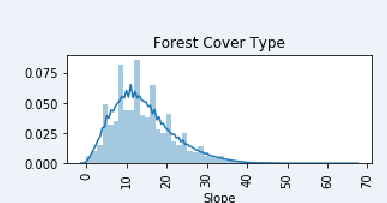
#### **3.5 Aspect :**

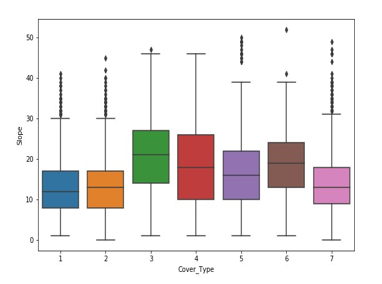
****



* Above show the distribution of Aspect ratio values it is clearly visible that data is not normally distributed we will be doing the statistical test to prove the same. From the box plot we get the idea that the Cover type 1 and Cover type 6 their Aspect values are widely spread as you can see from the size of the box. Some Cover type median values are overlapping so it might not be a good variable for making the right prediction.

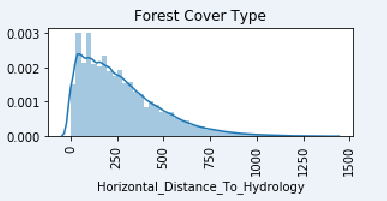
### **3.6 Slope:**

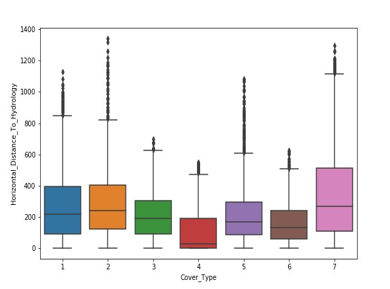
****



* Distribution of Slope values for each forest Cover Type is show in the figure above. From the distribution plot we can infer that it is almost normally distributed but slightly right skewed. Later point of time we will be doing the Statistical test to prove the same. Box plot shows how the Slope values have been distributed in each Cover Type and the centre line depicting the median value.

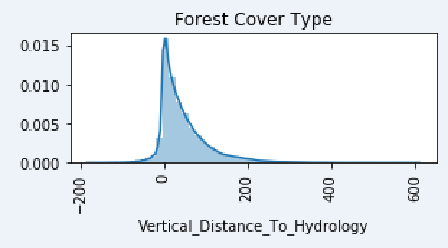
### **3.7 Horizontal\_Distance\_To\_Hydrology:**

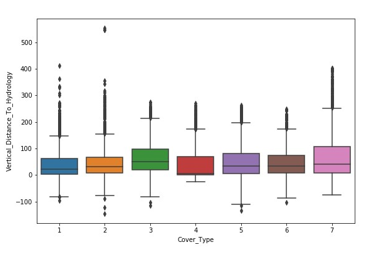




* From the distribution plot we can see that it is right skewed graph indicating that data is not normally distributed. The box plot showing the distribution of values for each cover type and as you can see there is little difference in the median value for each cover type and also there are some extreme values.

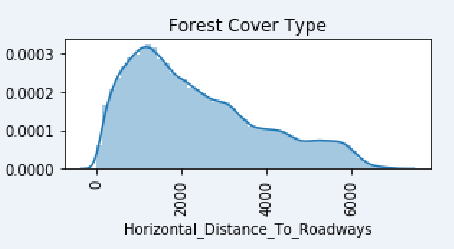
**3.8 Vertical Distance to Hydrology:**

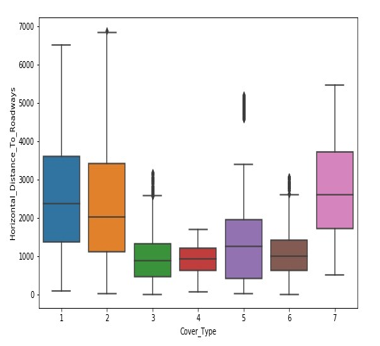
****



* From the boxplot it is visible that the median value of vertical distance to hydrology is almost equal so this is not a good feature for building model but will see whether it a important feature with the Statistical test.

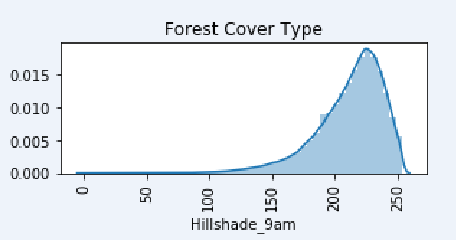
**3.9 Horizontal\_Dista nce\_To\_Roadway:**



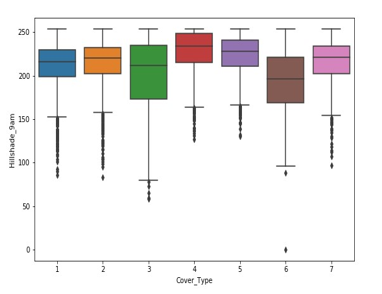


* As you can see from the boxplot that the value of horizontal distance to roadway is widely spread for Cover type 1 and Cover type 2 in which Cover type 2 is having the highest horizontal distance to roadway value.
* The median line value is also different for some of the cover type.

### **3.10 Hillshade\_9am**

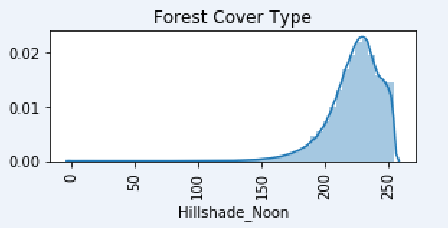


* Highly left skewed data not normally distributed. Also from the boxplot it is clear that mean values lie closer.

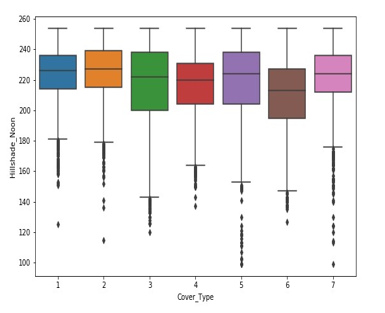


* Image above shows how the data is distributed for the Hill shade at 9am variable from the distribution plot we can clearly say that the data is not normal it is highly rightly skewed. Also from the boxplot it is clear that median values lie closer.

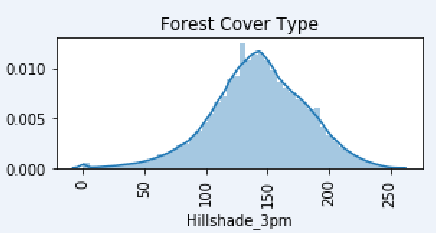
**3.11 Hillshade\_Noon:**

****

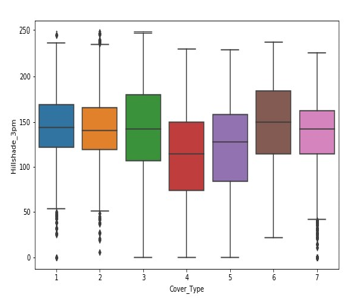
* Highly left skewed data not normally distributed. Also from the boxplot it is clear that median values lie closer we have to check statistically whether it is an important feature or not.

****

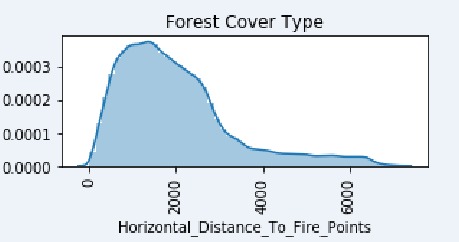
**3.12 Hillshade\_3pm:**



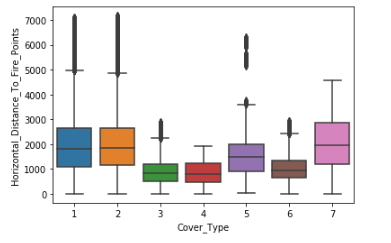
* Data seems to be normally distributed but can only confirm with the normality test. For some of the features the median values lies nearer.



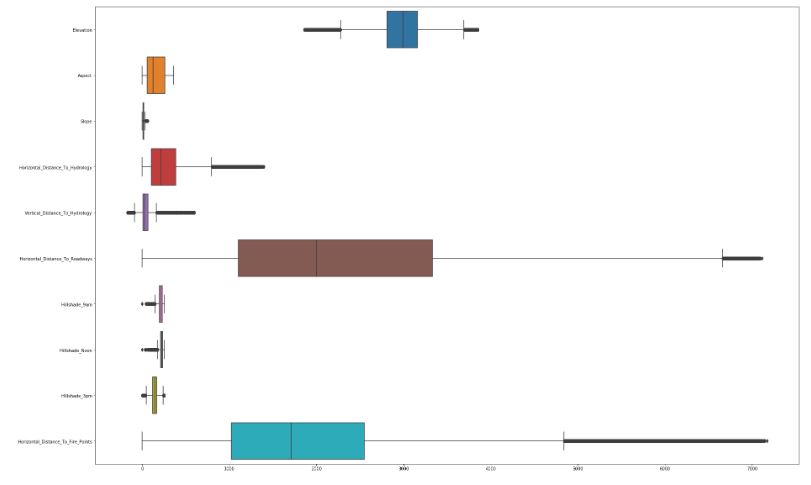
**3.13 Horizontal\_Distance\_To\_Fire\_Points:**

****

* Right Skewed distribution. From boxplot we can see that the median values are apart.

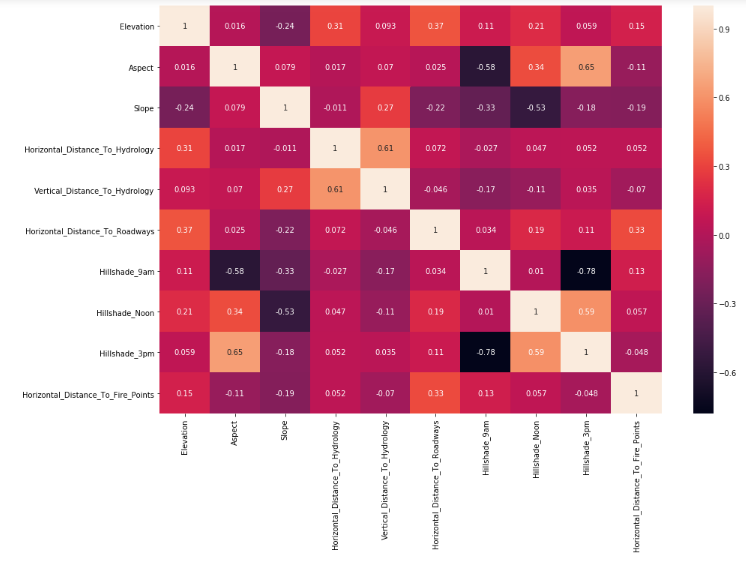
****

**3.14 Visualizing the spread and outliers of the data of numerical features:**



* Slope is the most squeezed box plot feature! It's densely packed taking on least range compared to all features. Having little range means mean and median will be quite close and we saw that before in the table, it has a difference of approx 1. It does have a few outliers though.
* Aspect feature is the only one which do not have any outliers having a range of 360. Since both Aspect and Slope are measured in degrees, Aspect takes on much bigger range than Slope because it has lowest max score, hence Aspect is much less densed than Slope. The first 50% of the data, from min to median is more densed than the last 50%, its more spread out.
* Hillshades feature also having similar plot like Slope including many outliers and taking on smaller range. Similiar plot is for Vertical\_Distance\_To\_Hydrology except here the minimum value is negative as we had seen in the table.
* Elevation and Horizontal\_Distance\_To\_Hydrology are the only features that doesn't have minimum value of 0. Elevation instead is plotted in middle having many outliers too.

## **3.15 Numeric Feature correlation:**

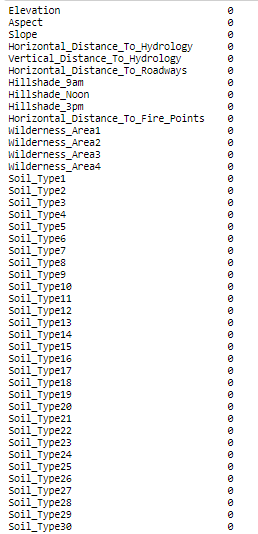


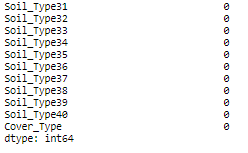
* Hillshade\_3pm and Hillshade\_9am show highly negative correlation while hillshade\_3pm and Aspect show highest positive correlation.
* Hillshade\_3pm and Aspect also had almost normal distribution compared to forest cover types classes. (Plot 4.1)
* Other features which have correlations are Vertical and Horizonal Distance to Hydrology, Hillshade\_3m and Hillshade\_Noon,Hillshade\_9am and Aspect and Hillshade\_Noon and Slope. So in total we have 6 pairs of correlation.

**CHAPTER 4: DATA CLEANING**

**4.1 Checking Missing Values:**

* To understand trees dataframe, let’s look at the data types and descriptive statistics. With pandas info method, we can list the non-null values and data types:

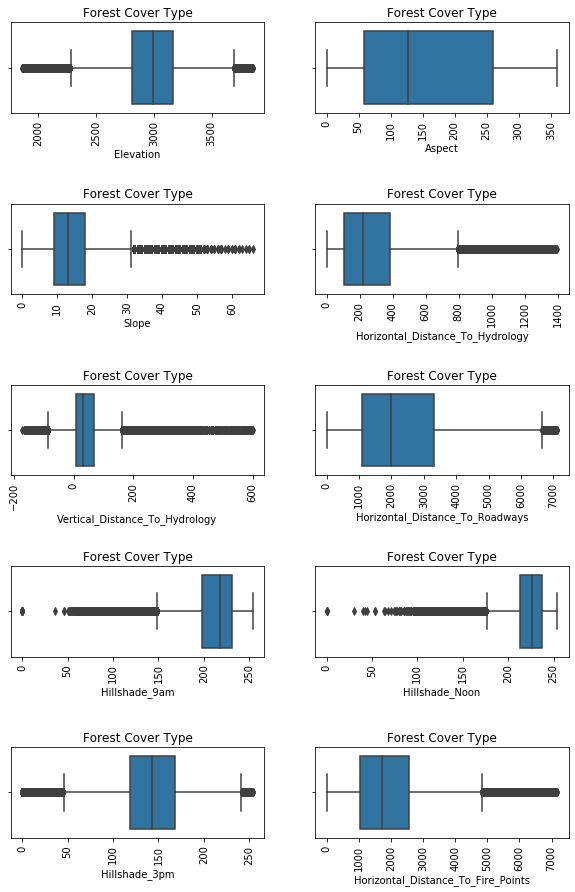


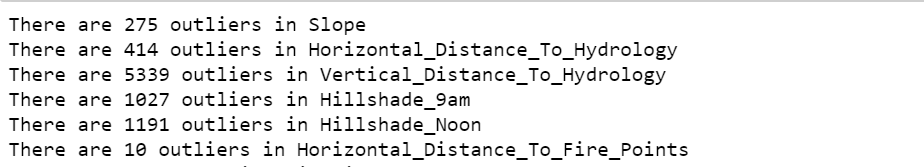


#### **Missing Attribute Values: None**

**4.2 Check for Anomalies & Outliers:**

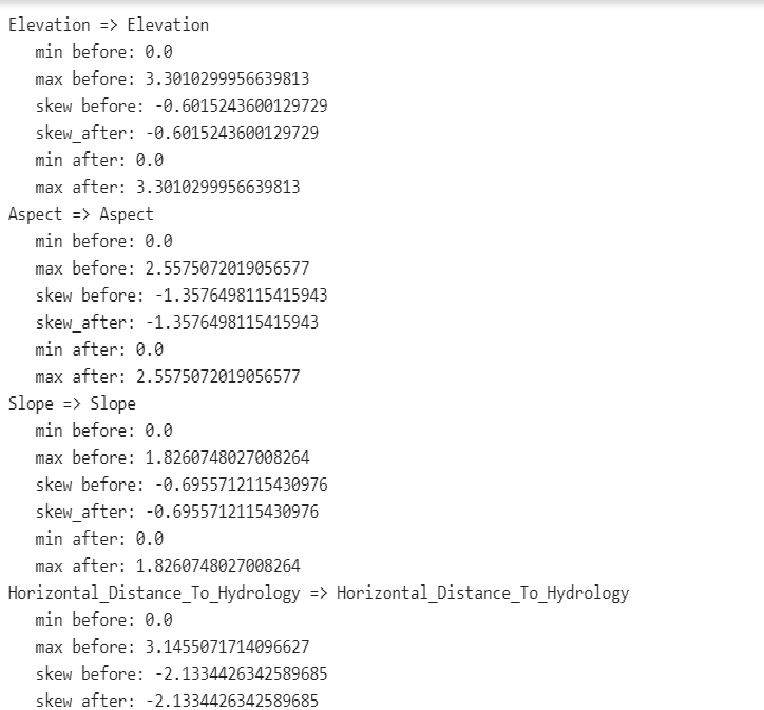
* The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory.

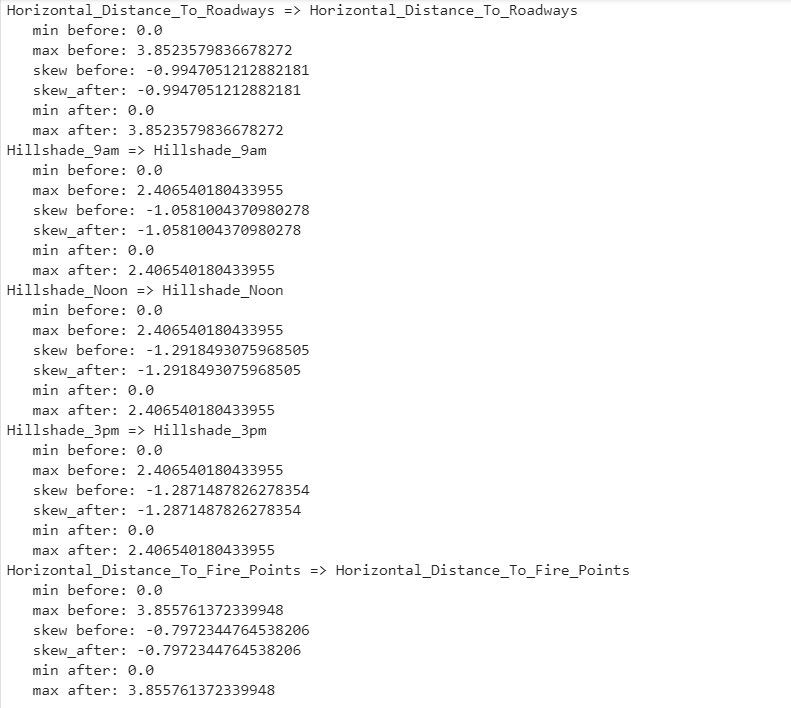




|  |  |  |
| --- | --- | --- |
| **Outlier Treatment** | **Oversampling** | **Scaling** |
| Box plot is drawn for Independent features against Target variable and outlier had been detected.  Since the outliers are legitimate, we have decided to retain them in data | Since our Dataset is highly imbalanced, we used SMOTE  oversampling  technique in order to tackle class | Since the base model is Decision Tree which is a non linear model. So we don’t require scaling here. But in future, if we are considering Logistic or KNN, we will be doing scaling and then building the model. |
| imbalance. |

**Applying Log Transformation for the Features:**





**CHAPTER 5: FEATURE SELECTION**

* Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminatesredundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.
* In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features.

**5.1 Statistical Test for Categorical Variables:**

* **Chi-Square test**

pvalue for Wilderness\_Area1 is: 0.0

pvalue for Wilderness\_Area2 is: 0.0

pvalue for Wilderness\_Area3 is: 0.0

pvalue for Wilderness\_Area4 is: 0.0

pvalue for Soil\_Type1 is: 0.0

pvalue for Soil\_Type2 is: 0.0

pvalue for Soil\_Type3 is: 0.0

pvalue for Soil\_Type4 is: 0.0

pvalue for Soil\_Type5 is: 0.0

pvalue for Soil\_Type6 is: 0.0

pvalue for Soil\_Type7 is: 1.7129654908603296e-21

pvalue for Soil\_Type8 is: 2.4898563581535384e-11

pvalue for Soil\_Type9 is: 3.892084170771216e-138

pvalue for Soil\_Type10 is: 0.0

pvalue for Soil\_Type11 is: 0.0

pvalue for Soil\_Type12 is: 0.0

pvalue for Soil\_Type13 is: 0.0

pvalue for Soil\_Type14 is: 0.0

pvalue for Soil\_Type15 is: 8.890831242782292e-19

pvalue for Soil\_Type16 is: 2.397288685254257e-168

pvalue for Soil\_Type17 is: 0.0

pvalue for Soil\_Type18 is: 0.0

pvalue for Soil\_Type19 is: 5.151252311605873e-282

pvalue for Soil\_Type20 is: 2.9220109658927094e-247

pvalue for Soil\_Type21 is: 9.989045946160907e-275

pvalue for Soil\_Type22 is: 0.0

pvalue for Soil\_Type23 is: 0.0

pvalue for Soil\_Type24 is: 0.0

pvalue for Soil\_Type25 is: 1.8052965766843747e-28

pvalue for Soil\_Type26 is: 0.0

pvalue for Soil\_Type27 is: 4.842165593085665e-50

pvalue for Soil\_Type28 is: 1.3244879394223233e-167

pvalue for Soil\_Type29 is: 0.0

pvalue for Soil\_Type30 is: 0.0

pvalue for Soil\_Type31 is: 0.0

pvalue for Soil\_Type32 is: 0.0

pvalue for Soil\_Type33 is: 0.0

pvalue for Soil\_Type34 is: 5.62552190567215e-230

pvalue for Soil\_Type35 is: 0.0

pvalue for Soil\_Type36 is: 1.0506714447191135e-182

pvalue for Soil\_Type37 is: 0.0

pvalue for Soil\_Type38 is: 0.0

pvalue for Soil\_Type39 is: 0.0

pvalue for Soil\_Type40 is: 0.0

* Statistical tests carried on categorical variables and it is inferred that all the variables are passing the statistical test there by selecting all the features for model building.

**5.2 Statistical Test for Continuous Variables:**

* **One way Anove test:**

Statistical test for Numerical features(ANOVA):

pvalue for Elevation is: 0.0

pvalue for Aspect is: 0.0

pvalue for Slope is: 0.0

pvalue for Horizontal\_Distance\_To\_Hydrology is: 0.0

pvalue for Vertical\_Distance\_To\_Hydrology is: 0.0

pvalue for Horizontal\_Distance\_To\_Roadways is: 0.0

pvalue for Hillshade\_9am is: 0.0

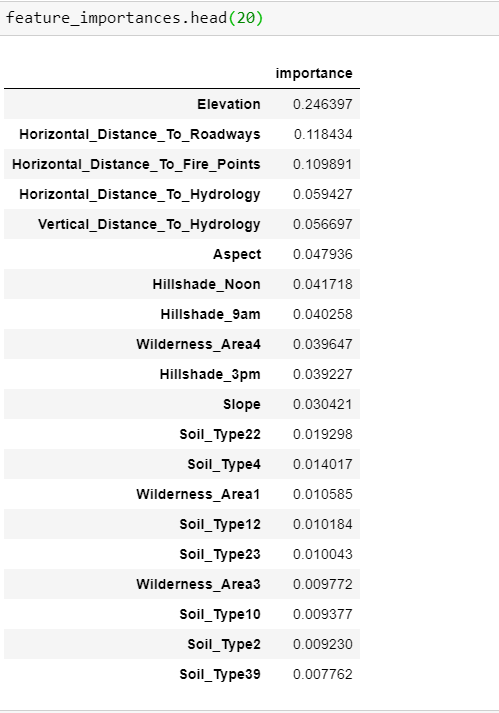
pvalue for Hillshade\_Noon is: 0.0

pvalue for Hillshade\_3pm is: 0.0

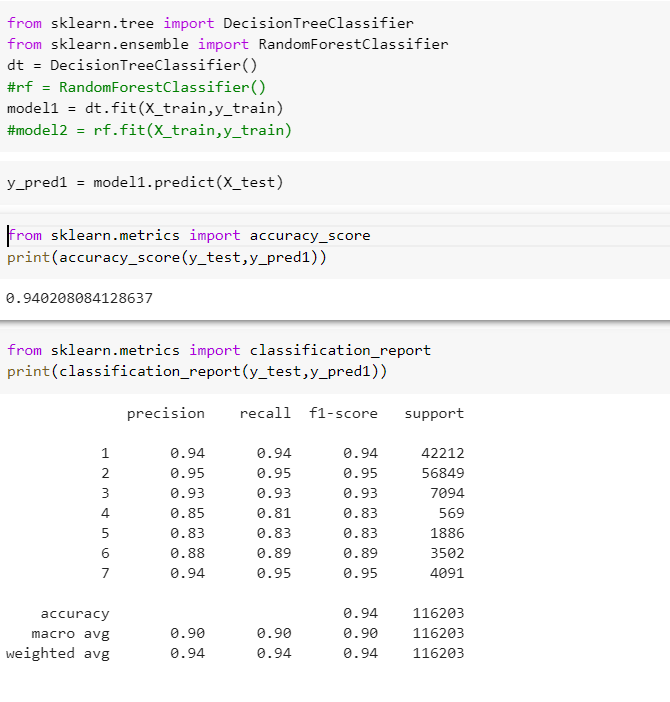
pvalue for Horizontal\_Distance\_To\_Fire\_Points is: 0.0

* Since Pvalue is less than alpha(0.05) for all continuous variables, all numerical features need to be considered.

**5.3 Top 20 important Features:**



Chapter 6 – Base Model



* The main reason we selected base model as Decision Tree is that decision tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand. The key idea is to use a decision tree to partition the data space into cluster (or dense) regions and empty (or sparse) regions.

### **Chapter 7 – Conclusions**

* In this project, we aim to predict the forest cover type using many variables that are influencing the outcome. Various feature engineering techniques performed on the datasets to improve over the primary data-set. Random forest with oversampling the data performed so well giving a training accuracy of 90% and testing accuracy of 94%.The problem we faced was that there was multicollinearity in the data and the variables which has that is the most contributing factor for the final outcome so we cannot drop those variables as it is an important feature, the only way is to treat it so first we went tried combining both the feature but at that time our accuracy was going down. So other way to reduce multicollinearity is by Principal component analysis which gave a similar training and testing accuracy. One of the defect of the Principal component analysis is that the data interpretability is lost.

### **Chapter 8 – Future Work**

* Data that existed was imbalanced and was not a normal dataset so in future if a better dataset comes up where there is less class imbalance we will be able to get a better result out of that.
* Since, there is Multicollinearity in the data so we use VIF technique to removed the features which are highly correlated but the drawback of this technique is that we lose the data interpretability.