FINAL PROJECT REPORT



Forest Cover type Prediction



**“submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM”**

**Submitted by**

|  |  |
| --- | --- |
| **Student Name** | **SIS ID** |
| AMIT KUMAR JHA | 2TINTB4GYY |
| KUNAL SAHU | J6H0PHI60T |
| UJWALITHA PAKKI | Q5K0C9L34P |
| JYOTI DAHIYA | 871C9GWRNS |
| SUDHANSHU SHARMA | 0QE689C5EK |
| FARHAT ALI SIDDIQUI | 74QW5N6OKD |

# **Batch: DSE\_BLR\_SEP-2019**

**Mentor: ROMIL GUPTA**



# **Acknowledgements**

At the outset, we are indebted to our Mentor Romil Gupta for his time, valuable input and guidance and his willingness to go out of the way and ensuring we were adequately prepared to meet the challenges this research project posed, without whom our project would have failed to come to completion. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We also thank the Institute for the resources they provided us and all the faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

AMIT KUMAR JHA

KUNAL SAHU

UJWALITHA PAKKI

JYOTI DAHIYA

SUDHANSHU SHARMA

FARHAT ALI SIDDIQUI

Date:28th Feb 2020

Place: Bangalore

# **Certification of completion**

I hereby certify that the project titled “Forest Cover type Prediction” was undertaken and completed by AMIT KUMAR JHA, KUNAL SAHU, UJWALITHA PAKKI, JYOTI DAHIYA, SUDHANSHU SHARMA, FARHAT ALI SIDDIQUI; students of the DSE-Sep 2019 batch of the Post Graduate Program in Data Science & Engineering, Bangalore under my guidance and supervision.

Mr. Romil Gupta

Date: 28th Feb 2020

Contents

[**Batch: DSE\_BLR\_SEP-2019** 1](#_Toc33735345)

[**Acknowledgements** 2](#_Toc33735346)

[**Certification of completion** 3](#_Toc33735347)

[CHAPTER 1: INDUSTRY REVIEW 6](#_Toc33735348)

[1.1 Literature Survey: 6](#_Toc33735349)

[The Cover type dataset was first used in a machine learning context by Blackard and Dean, as part of Blackard’s doctoral thesis and then as part of an academic article. Both of the original works use the data set as the basis of a supervised learning task. The data set’s instances are drawn from US Forest Service (USFS) Region 2 Resource 6](#_Toc33735350)

[CHAPTER 2: DATASET AND DOMAIN 8](#_Toc33735351)

[2.1 Domain Background: 8](#_Toc33735352)

[2.2 Problem Statement: 9](#_Toc33735353)

[2.3 NEED FOR STUDY: 9](#_Toc33735354)

[2.4 Data Dictionary: 9](#_Toc33735355)

[2.4.1 Variable categorization: 10](#_Toc33735356)

[2.4.2 Wilderness areas details: 11](#_Toc33735357)

[2.4.3 Soil Type Feature Details: 11](#_Toc33735358)

[2.4.4 Forest Cover Type 13](#_Toc33735359)

[(The variable for our prediction): 13](#_Toc33735360)

[*CHAPTER 3:* DATA Exploration 13](#_Toc33735361)

[3.1 The Distribution of Forest Type in: 13](#_Toc33735362)

[3.2 The Distribution of Wilderness area: 15](#_Toc33735363)

[3.3 SOIL TYPE distribution: 17](#_Toc33735364)

[3.4 Elevation: 18](#_Toc33735365)

[3.5 Aspect 19](#_Toc33735366)

[3.6 Slope: 19](#_Toc33735367)

[3.7 Horizontal\_Distance\_To\_Hydrology: 20](#_Toc33735368)

[3.8 Vertical Distance to Hydrology: 20](#_Toc33735369)

[3.9 Horizontal\_Distance\_To\_Roadway 20](#_Toc33735370)

[3.10 Hillshade\_9am 21](#_Toc33735371)

[3.11 Hillshade\_Noon: 21](#_Toc33735372)

[3.12 Hillshade\_3pm: 22](#_Toc33735373)

[3.13 Horizontal\_Distance\_To\_Fire\_Points: 22](#_Toc33735374)

[3.14 Bi-Variate Analysis for correlated features: 23](#_Toc33735375)

[Plotting scatter plots of all features that have correlation greater than 0.5 with each other. 23](#_Toc33735376)

[3.14.1 Aspect Vs Hillshade\_3pm: 23](#_Toc33735377)

[3.14.2 Aspect Vs Hillshade\_9am: 23](#_Toc33735378)

[3.14.3 Slope Vs Hillshade\_Noon: 24](#_Toc33735379)

[3.14.4 Vertical and Horizontal Distance to Hydrology: 24](#_Toc33735380)

[3.14.5 Hilshade\_9am and Hillshade\_3pm : 24](#_Toc33735381)

[3.14.6 Hillshade\_Noon and Hillshade\_3pm : 25](#_Toc33735382)

[3.15 Visualizing the spread and outliers of the data of numerical features: 25](#_Toc33735383)

[3.16 Numeric Feature statistics: 27](#_Toc33735384)

[3.17 Categorical Feature statistics: 28](#_Toc33735385)

[**3.18 Numeric Feature correlation:** 29](#_Toc33735386)

[CHAPTER 4: DATA CLEANING 30](#_Toc33735387)

[4.1 Checking Missing Values: 31](#_Toc33735388)

[4.2 Checking For Duplicate values: 32](#_Toc33735389)

[4.3 Checking if any observations is present in more than one type in same category of Wilderness and Soil Type: 32](#_Toc33735390)

[**4.4 Dimensionality Reduction by removing collinear features:** 32](#_Toc33735391)

[**4.5 Check for Anomalies & Outliers:** 33](#_Toc33735392)

[CHAPTER 5: FEATURE SELECTION 36](#_Toc33735393)

[**5.1 Statistical Test for Categorical Variables:** 36](#_Toc33735394)

[5.2 Statistical Test for Continuous Variables: 38](#_Toc33735395)

[5.3 Top 20 important Features: 38](#_Toc33735396)

[CHAPTER 6: MODEL BUILDING 39](#_Toc33735397)

[6.1 Evaluation Metrics for the model: 40](#_Toc33735398)

[6.2 Classification Results: 41](#_Toc33735399)

[6.2 Final Model: 43](#_Toc33735400)

[CHAPTER 7: CONCLUSION AND SUMMARY 44](#_Toc33735401)

[CHAPTER 8: REFERENCES AND BIBLIOGRAPHY 44](#_Toc33735402)

**CHAPTER 1: INDUSTRY REVIEW**

1.1 **Literature Survey:**

## The Cover type dataset was first used in a machine learning context by Blackard and Dean, as part of Blackard’s doctoral thesis and then as part of an academic article. Both of the original works use the data set as the basis of a supervised learning task. The data set’s instances are drawn from US Forest Service (USFS) Region 2 Resource

Information System data and the classifier must predict the type of forest cover present in a 30 x 30 metre cell given the observed geographic information system (GIS) variables.

Since its introduction, it has become a standard benchmark data set in the literature and has been cited by hundreds of papers. The data set is available as raw data from the UCI Machine Learning Repository and in normalized form from the website of Massive Online Analysis (MOA), an open source framework for data stream

mining.

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.

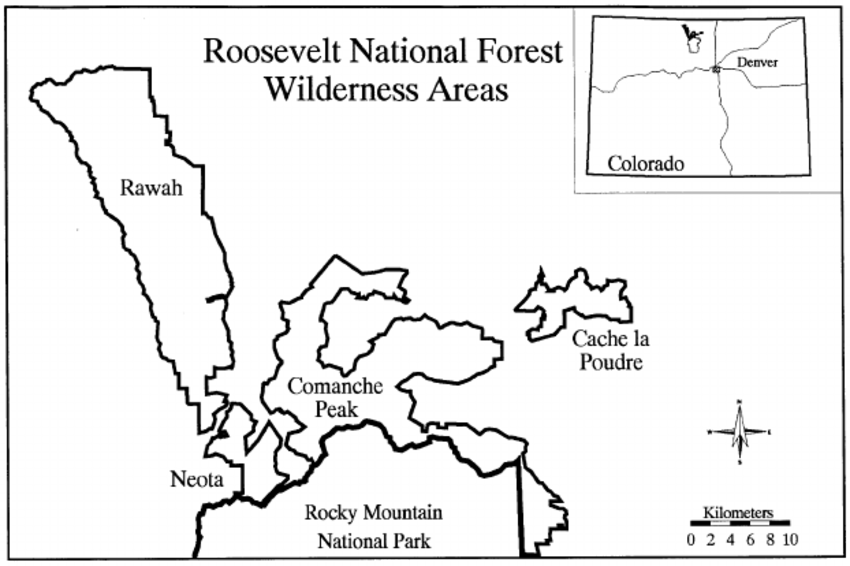
Blackard, Jock A. 1998. "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types." Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado. 165 pages.

**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 |

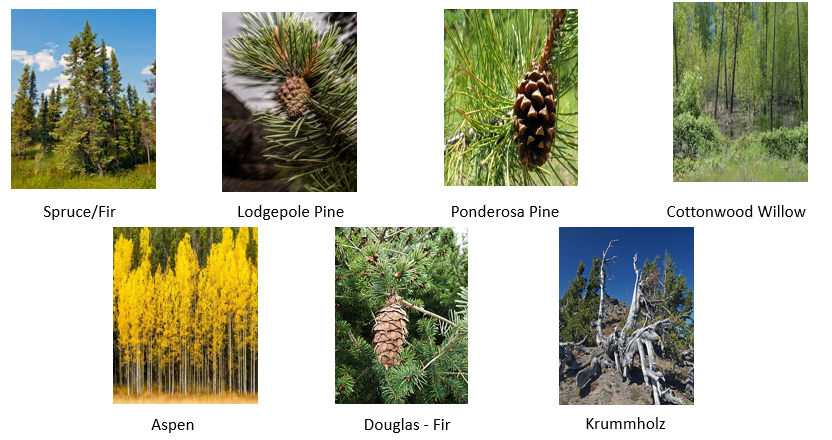
**2.4.4 Forest Cover Type**

**(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.

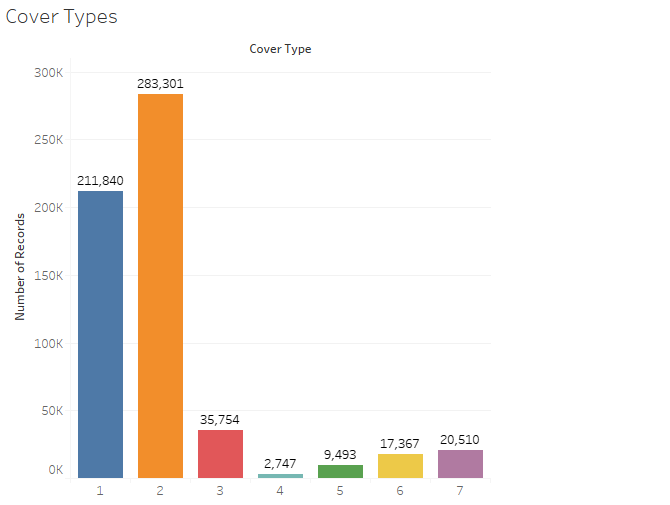
Now let’s look at the names of these types of Forest Cover:

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

**

***CHAPTER 3:* DATA Exploration**

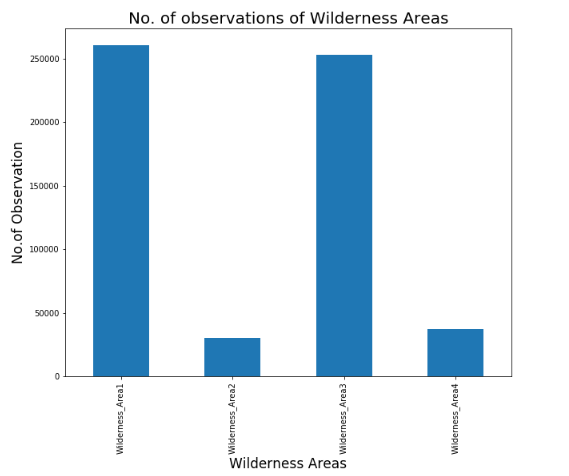
**3.1 The Distribution of Forest Type in:**

**

|  |  |
| --- | --- |
| 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 |

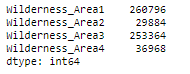
The distribution of these classes is not equal. Spruce and Lodgepole have the most while Cottonwood has the least records. Having such distribution will not give us appropriate results because of unequal amount of distribution. Also, to note here that every observation is assigned to some class of forest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

**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:**



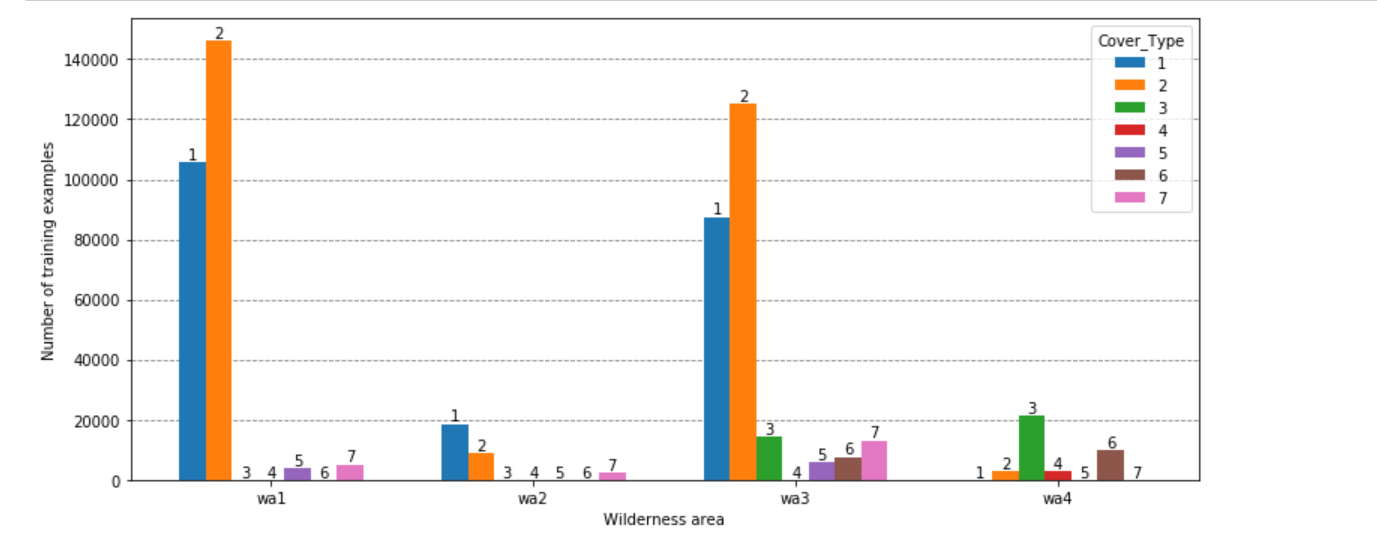
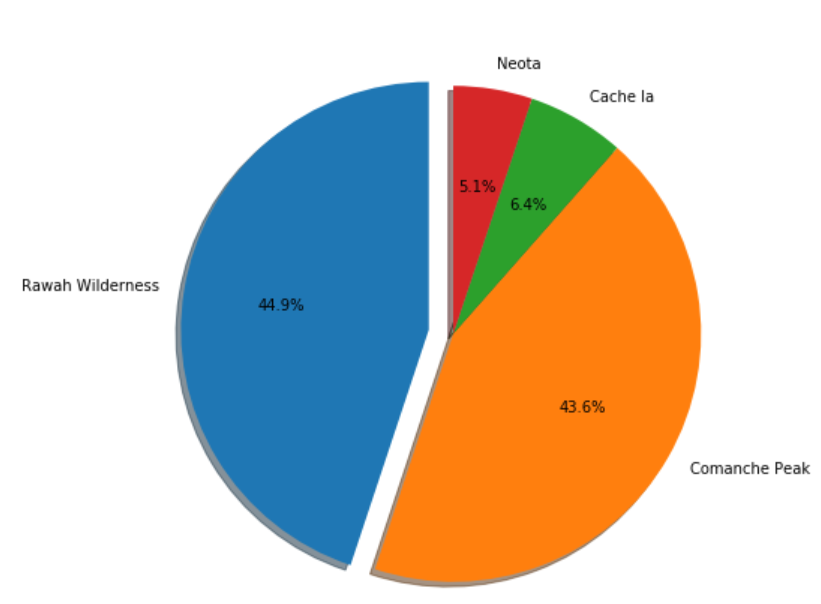
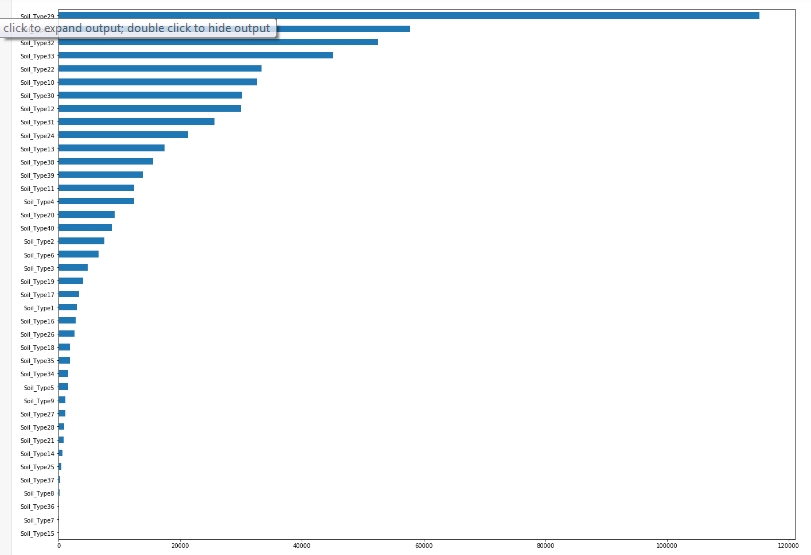


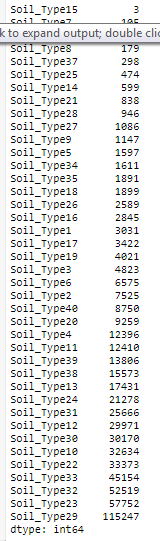
Image above shows the distribution of each cover type in different wilderness area. As you can see cover type 2 is occurring in almost in all the wilderness area and some of them are occurring in only specific wilderness area.



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

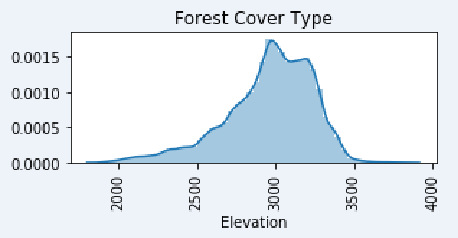
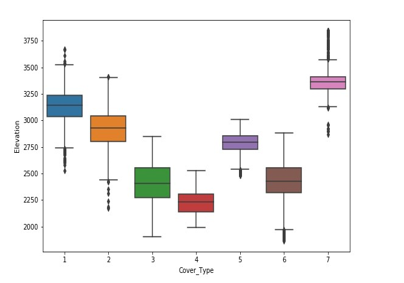
**3.3 SOIL TYPE distribution:**

****



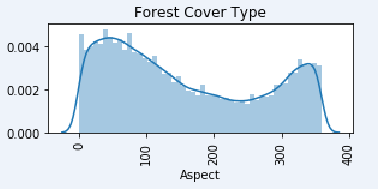
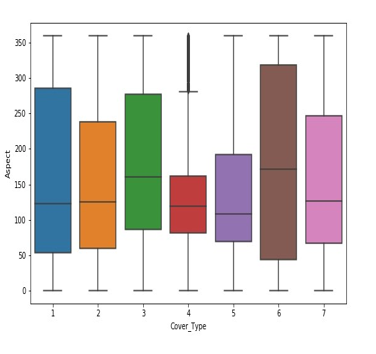
* The least observation are of Soil\_Type15 of 3
* Soil\_Type29 has the highest,

#### **3.4 Elevation:**

**** 

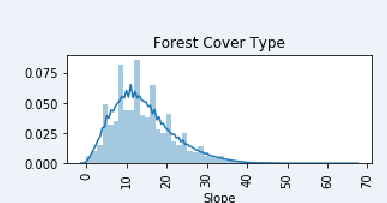
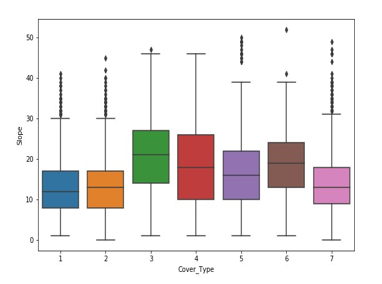
From the diagram we can say that it is a bimodal graph so we can confirm that data is not normally distributed. From the boxplot we can get the inference that median value are different for all the Cover Type may be with this might be one of the important feature to distinguish between the different 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 :***

**** 

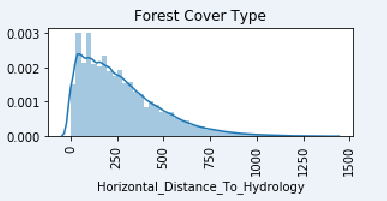
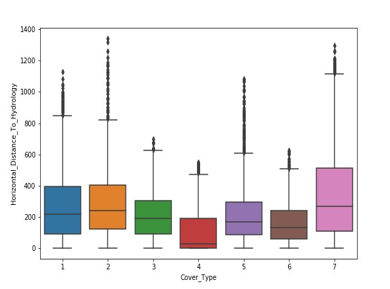
Above show the distribution of Aspect ratio values it is clearly visible that data is not normally distributed we will be doing the statistical test to prove the same. From the box plot we get the idea that the Cover type 1 and Cover type 6 their Aspect values are widely spread as you can see from the 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:**

**** 

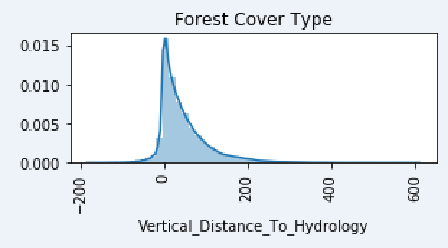
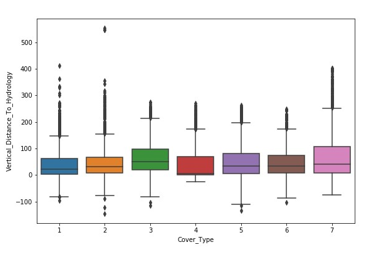
Distribution of Slope values for each forest Cover Type is show in the figure above. From the distribution plot we can infer that it is almost normally distributed but slightly right skewed. Later point of time we will be doing the Statistical test to prove the same. Box plot shows how the Slope values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

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

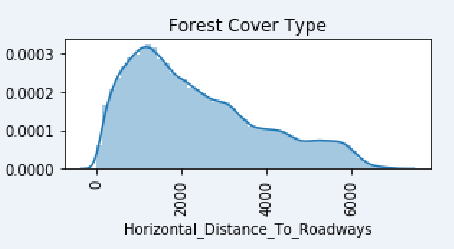
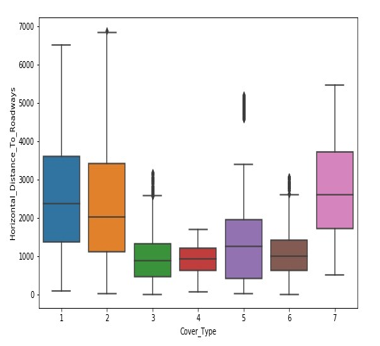
From the distribution plot we can see that it is right skewed graph indicating that data is not normally distributed. The box plot showing the distribution of values for each cover type and as you can see there is little difference in the median value for each cover type and also there are some extreme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

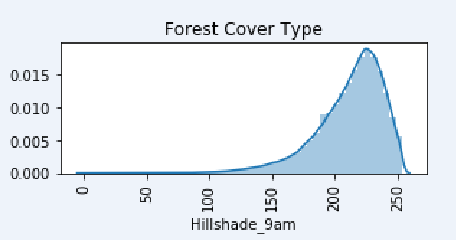
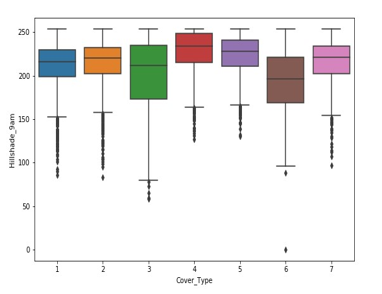
 

As you can see from the boxplot that the value of horizontal distance to roadway is widely spread for Cover type 1

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.

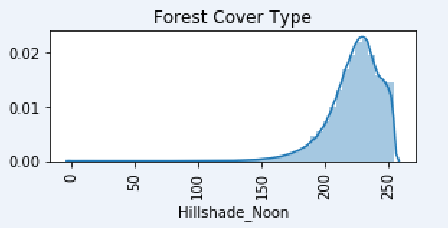
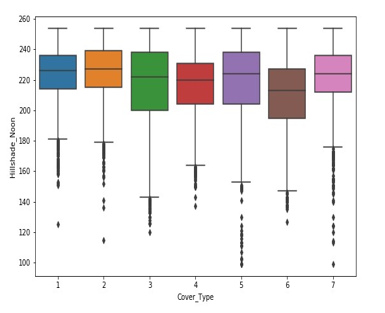
### **3.10 Hillshade\_9am**

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

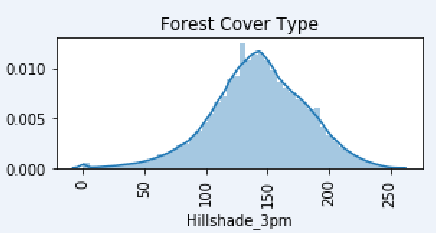
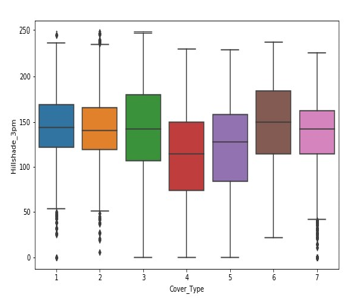
Image above shows how the data is distributed for the Hill shade at 9am variable from the distribution plot we can clearly say that the data is not normal it is highly rightly skewed. Also from the boxplot it is clear that median values lie closer.

**3.11 Hillshade\_Noon:**

** **

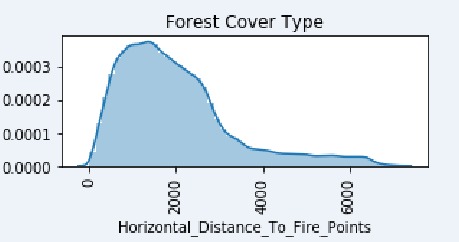
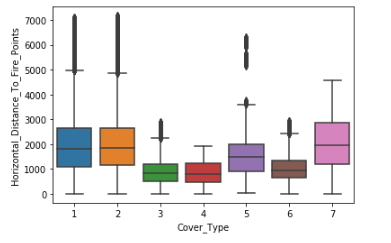
Highly left skewed data not normally distributed. Also from the boxplot it is clear that median values lie closer we have to check statistically whether it is an important feature or not.

**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:**

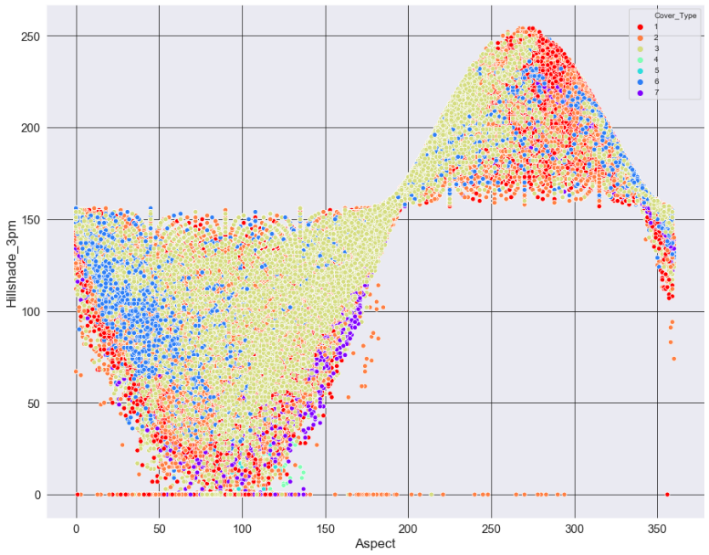
** **

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

**3.14 Bi-Variate Analysis for correlated features:**

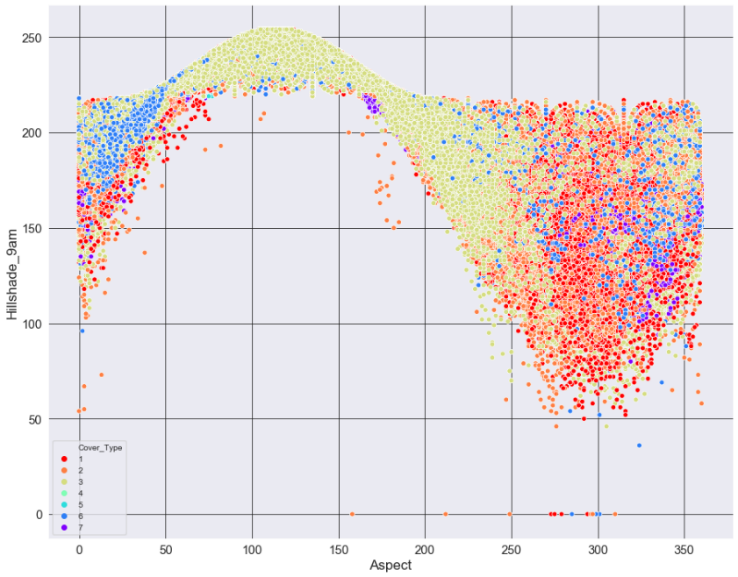
### Plotting scatter plots of all features that have correlation greater than 0.5 with each other.

**3.14.1 Aspect Vs Hillshade\_3pm:**



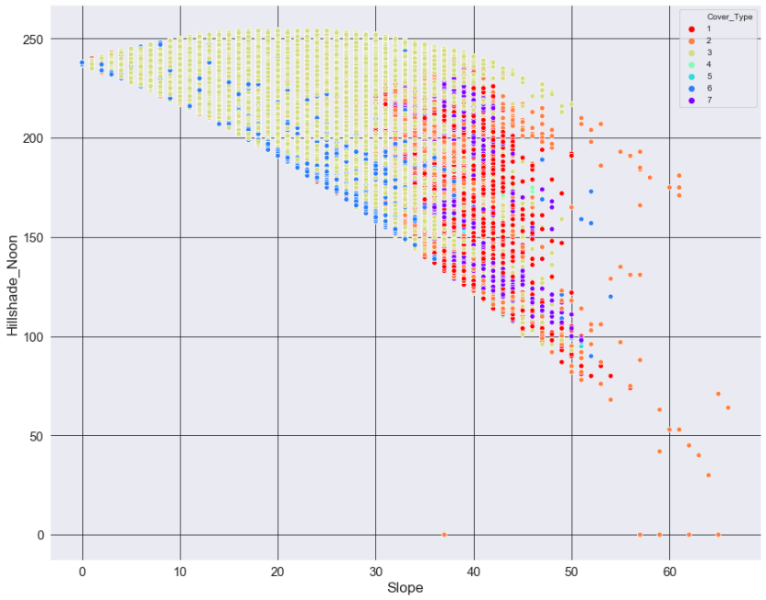
Hillshade\_3pm and Aspect represent relationship of a sigmoid function. The data points at the boundaries of the figure mostly belong to forest cover type class 1 while class 3 takes on most of datapoints in the figure followed by forest cover type class 6. The datapoints when Hillshade\_3pm is 0 belongs to class 1,2,3 or 7 regardless of what Aspect values it has

**3.14.2 Aspect Vs Hillshade\_9am:**



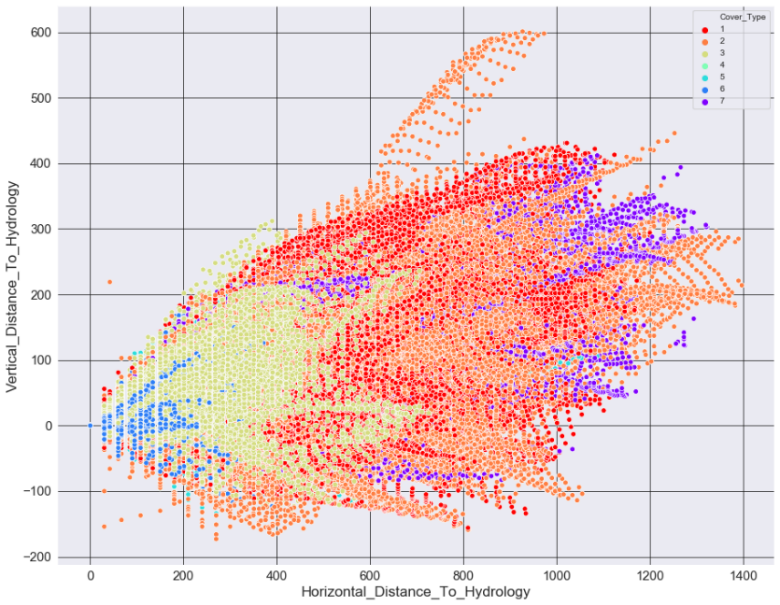
The figure Hillshade\_9am and Aspect also represent relationship of a sigmoid function just its flipped over the y-axis. Class type 3 has the highest observation here followed by the class type 1 and 6.

**3.14.3 Slope Vs Hillshade\_Noon:**



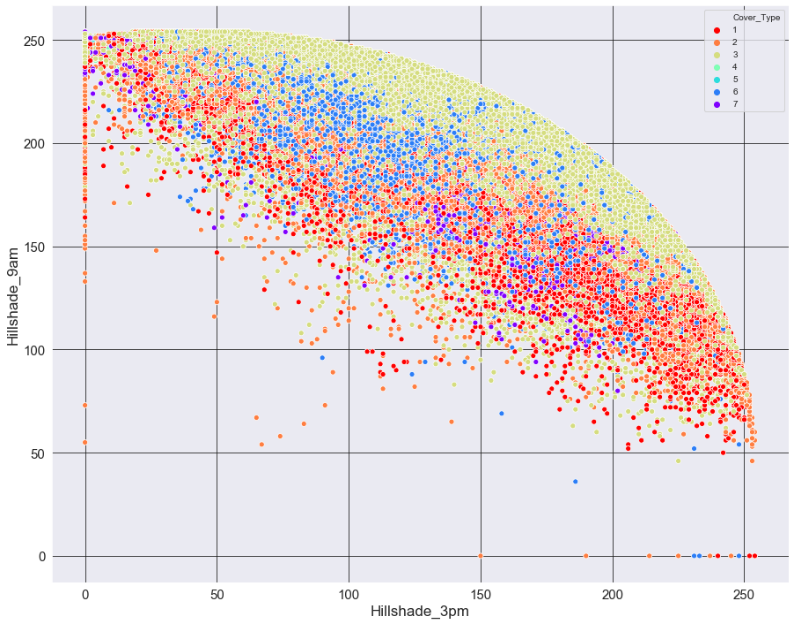
Hillshade\_Noon and Slope have a horizontal 'V' shaped representation. Lower degrees represent class 4 and 6 while high degree values represent class 1, 2 and 7 also we can see decrease in Hillshade\_Noon value as slope increases and it geographically makes sense.

**3.14.4 Vertical and Horizontal Distance to Hydrology:**



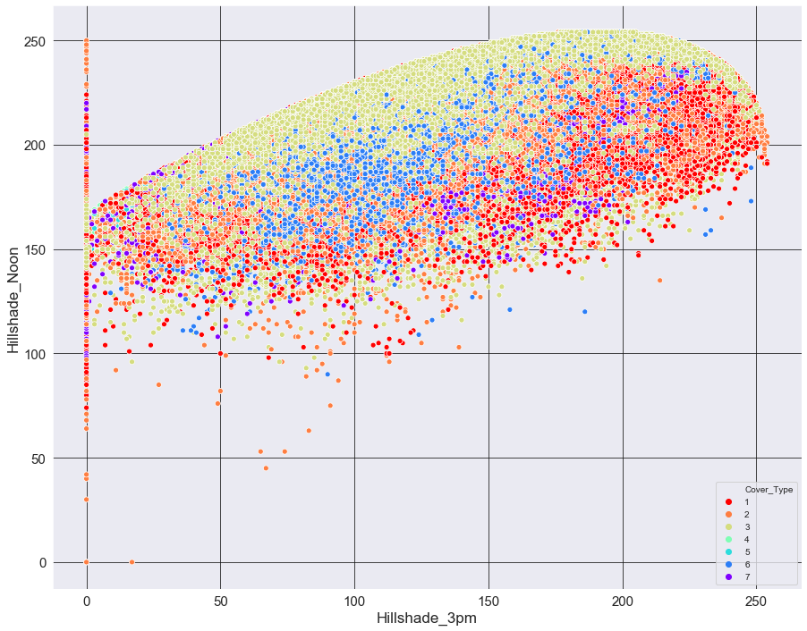
Vertical and Horizontal Distance to Hydrology represent a linear but spreaded out type, not a single line fit to all datapoints. Class type 7 and 2 have more observation here and spreaded out while class type 3 and 6 are densely packed between the range 0-800m of Horizontal Distance to Hydrology

**3.14.5** **Hilshade\_9am and Hillshade\_3pm**  **:**



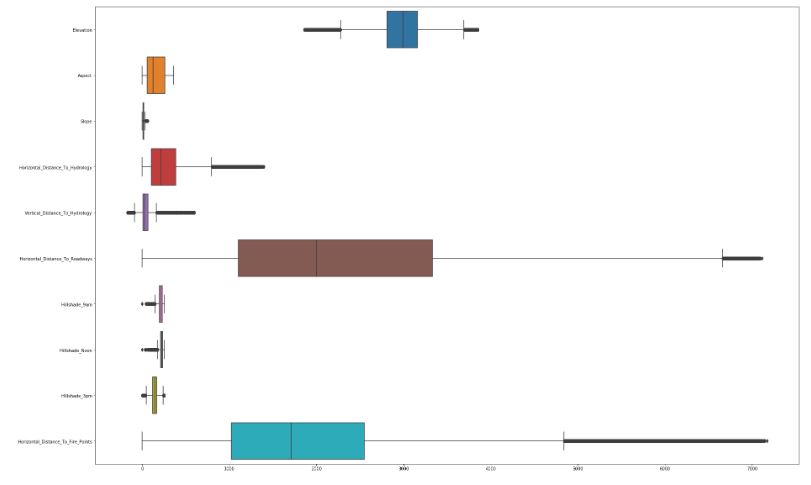
Hilshade\_9am and Hillshade\_3pm figure represents relationship of a sliced-out part of a circle where top most of the datapoints belong to class 3 and middle and bottom area belong to rest of the classes.

**3.14.6** **Hillshade\_Noon and Hillshade\_3pm**  **:**



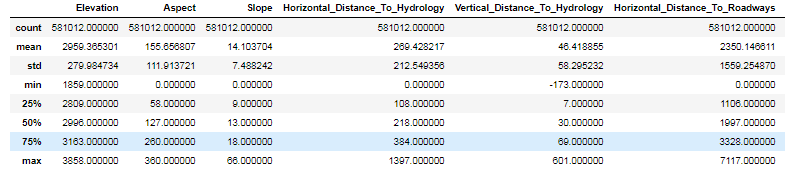
Hillshade\_Noon and Hillshade\_3pm have similar observation as described before just a difference here is that it's flipped over y-axis. We also see similar patterns of datapoints too as before.

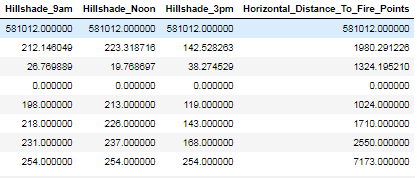
**3.15 Visualizing the spread and outliers of the data of numerical features:**



* 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 dense than Slope. The first 50% of the data, from min to median is more dense than the last 50%, its more spread out.
* Hillshades feature also having similar plot like Slope including many outliers and taking on smaller range. Similar 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.
* Horizontal\_Distance\_To\_Roadways is the most spread data of all features because it has the highest standard deviation score followed by Horizontal\_Distance\_To\_Fire\_Points though this feature has the maximum value. We can see visually only how spread these are and which one is most. Horizontal\_Distance\_To\_Fire\_Points may be having largest number of outliers I guess from this plot. If we compare these two features, the last 50% of the data of Horizontal\_Distance\_To\_Roadways is much more spread and less dense compared to Horizontal\_Distance\_To\_Fire\_Points , hence having high standard deviation score.

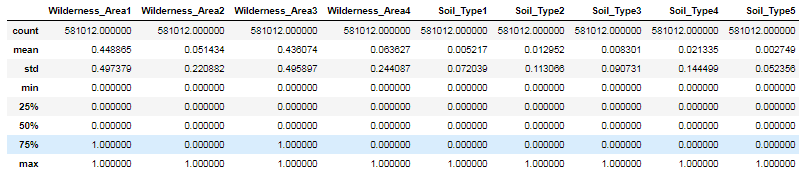
## **3.16 Numeric Feature statistics:**

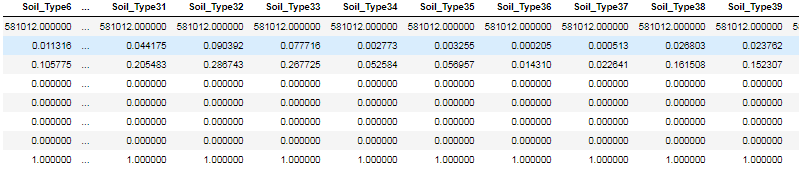




* Mean of the feature vary from as low as 14 to as high as 2959.
* Standard deviation tells us how spread the data is from the mean, here we can see Horizontal\_Distance\_To\_Roadways is the most spread out data followed by Horizontal\_Distance\_To\_Fire\_Points and Elevation.
* All the features have minimum value of 0 except Elevation and Vertical\_Distance\_To\_Hydrology features. Where Elevation has the highest minimum value and Vertical\_Distance\_To\_Hydrology has the lowest, being negative.
* Hillshades features have similar maximum value of 254 while Horizontal\_Distance\_To\_Fire\_Points has the highest followed by Horizontal\_Distance\_To\_Roadways feature and they also have the highest ranges of all features. Slope having lowest maximum value and also being lowest in range followed by Apsect feature.
* The reason some features are so widely spread and having high values and some features don't is because 5 out of 10 variables are measured in meters, includes ('Elevation', 'Horizontal\_Distance\_To\_Hydrology' , Vertical\_Distance\_To\_Hydrology', 'Horizontal\_Distance\_To\_Roadways', Horizontal\_Distance\_To\_Fire\_Points'), so it makes sense that these have high values and ranges. Features like Aspect and Slope are measured in degrees so its maximum value can't go above 360. While Hillshades features can take on max value of 255.

## **3.17 Categorical Feature statistics:**

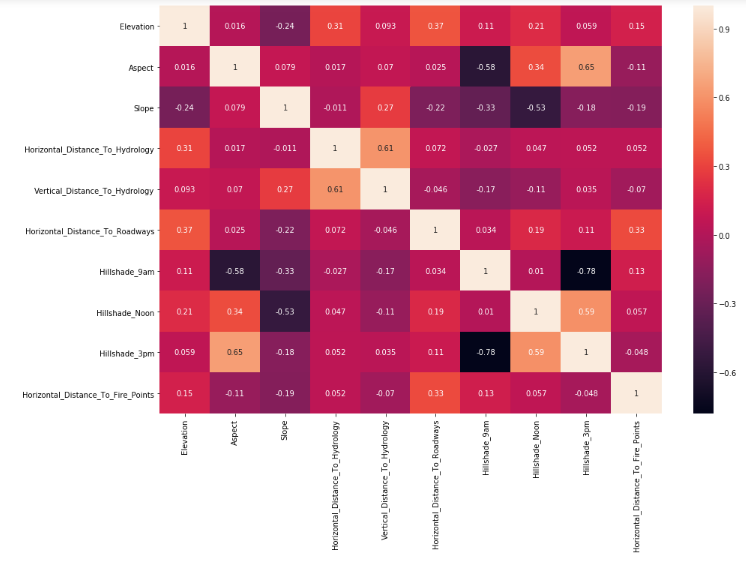




* Generally, we don’t check statistics for categorical features but since here all the values can only be either 0 and 1. The mean can tell us useful information here. Wilderness\_Area1 has the highest mean followed by Wilderness\_Area3, this means that Wilderness\_Area1 has the most presence in the data compared to other Wilderness Areas.
* Most observation have features either Wilderness\_Area1 or Wilderness\_Area3.
* The least amount of observation will be seen from Wilderness\_Area2.
* One more thing to notice here is that when we add all the mean of Wildernesss\_Areas 0.448864 + 0.051434 + 0.436074 + 0.063627 we get result 0.999999 which is approximately 1. This actually makes sense because all the observations can be from any one Wilderness area
* Hence if we look at this in the probability perspective, we can say that, the next observation that we get has 44.8% probability that it’s been taken from Wilderness\_Area1, 43.6% probability that it's taken from Wilderness\_Area3 and so on for others.

By looking at these statistics of two different data types and since the features have different spreads and uneven amount of distribution, we will feature scale these so that all the feature have similar ranges between 0 and 1. Some algorithm are very sensitive to high values hence giving us inappropriate results while some algorithms are not. Do be on safe side we will feature scale it and will do this in Data Engineering Section.

## **3.18 Numeric Feature correlation:**

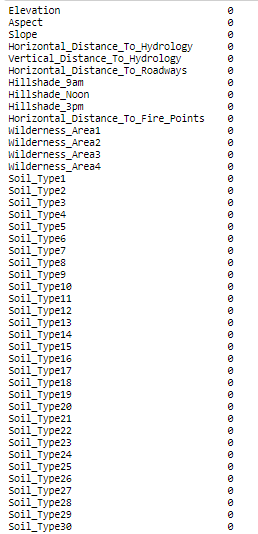


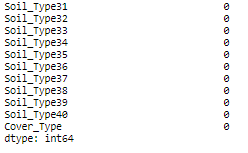
* Hillshade\_3pm and Hillshade\_9am show highly negative correlation while hillshade\_3pm and Aspect show highest positive correlation.
* Hillshade\_3pm and Aspect also had almost normal distribution compared to forest cover types classes. (Plot 4.1)
* Other features which have correlations are Vertical 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:**

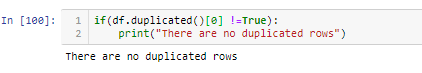
To understand trees dataframe, let’s look at the data types and descriptive statistics. With pandas info method, we can list the non-null values and data types:





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

**4.2 Checking For Duplicate values:**



**4.3 Checking if any observations is present in more than one type in same category of Wilderness and Soil Type:**

We have 0 observations that shows presence in more than 1 Wilderness Area.

We have 0 observations that shows no presence in any Wilderness Area.

We have 0 observations that shows presence in more than 1 Soil Type Area.

We have 0 observations that shows no presence in any Soil Type Area.

**4.4 Dimensionality Reduction by removing collinear features:**

We also see above in visualization section that Wilderness Area and Soil Type Area have no category that has no - observations of it. So every feature has

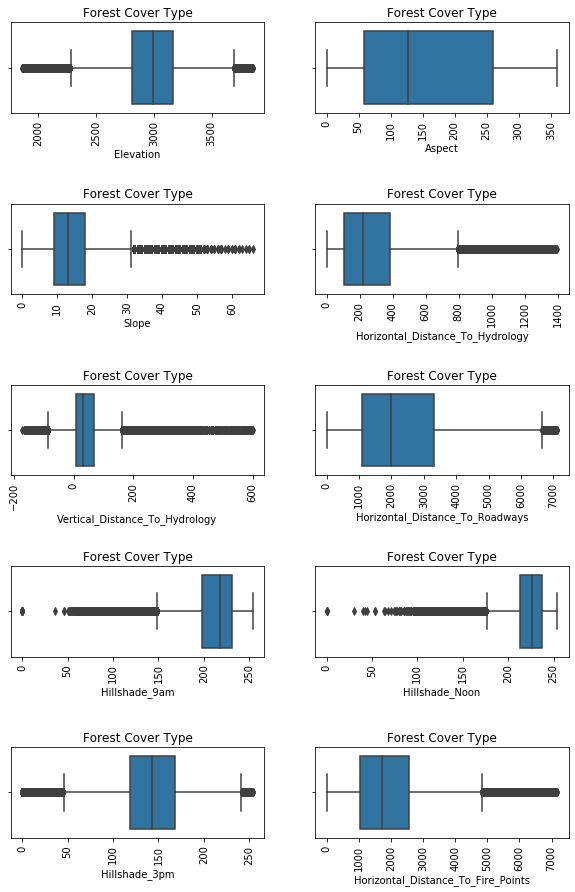
presence or values of an observations so we can't just delete any feature since it may have an important informations for our models in predicting classes.

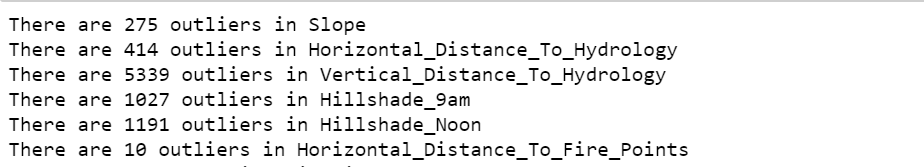
Due to multi-collinearity present we drop the following features:

* Hill\_Shade\_9am

**4.5 Check for Anomalies & Outliers:**

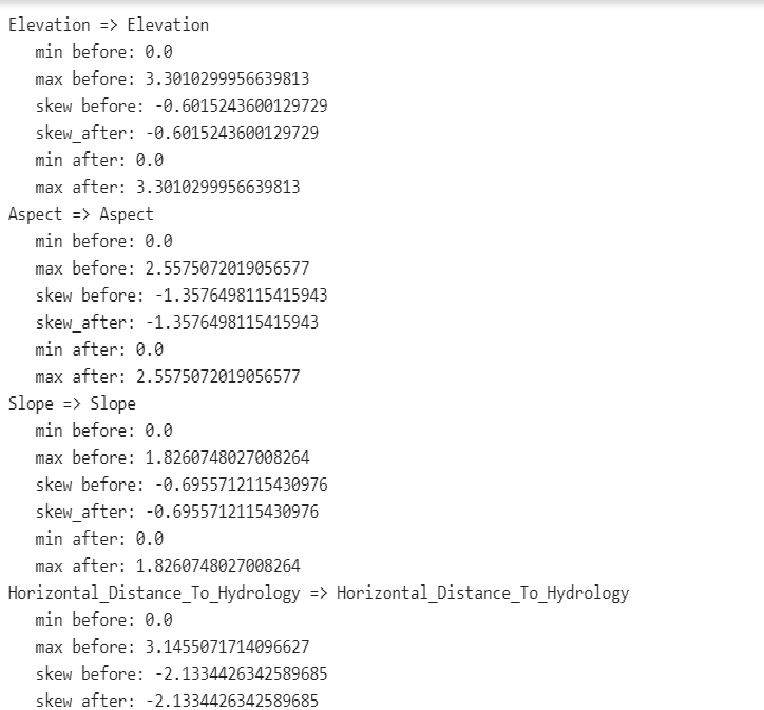
The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory.

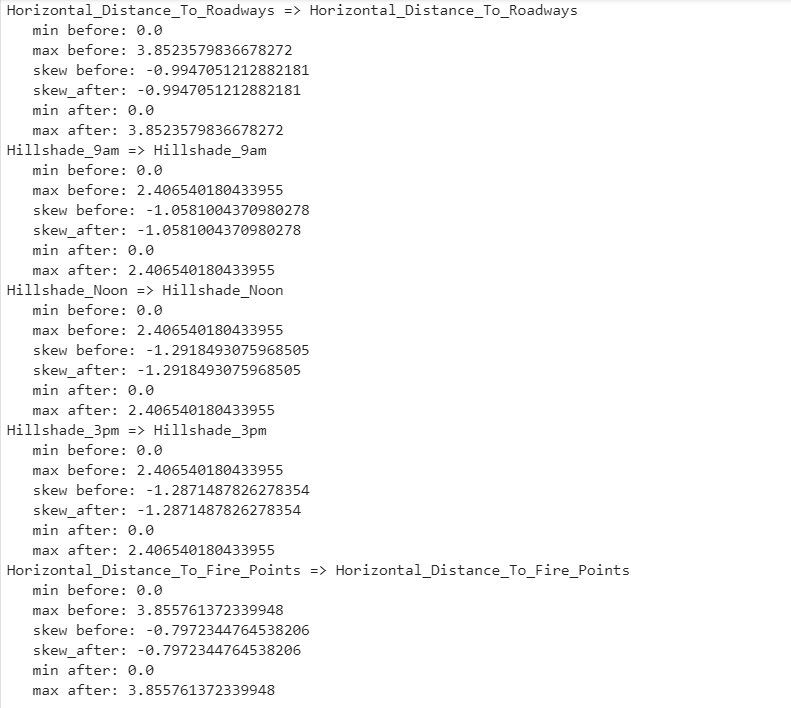




|  |  |
| --- | --- |
| **Outlier Treatment** | **Scaling** |
| Box plot is drawn for Independent features against Target variable and outlier had been detected.  Since the outliers are legitimate, we have decided to retain them in data | Since the base model is Decision Tree which is a non linear model. 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. |
|

**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 eliminates redundant 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.

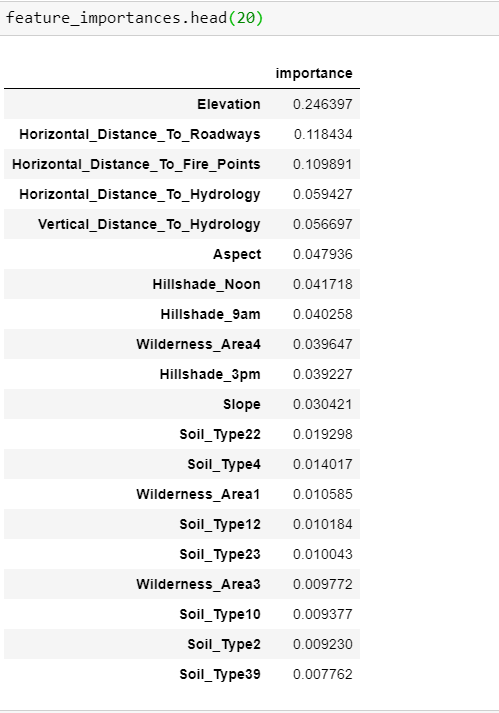
**5.3 Top 20 important Features:**

Since we already have lots of observation now to train the model, we also happen to have lots of features. This will make algorithm run very slowly, have difficulty in learning and also tend to overfit in training set and do worse in testing.

We also see above in visualization section that Wilderness Area and Soil Type Area have no category that has no - observations of it. So, every feature has presence or values of an observations so we can't just delete any feature since it may have an important information for our models in predicting classes.

To approach the problem, we need to see how each feature has an impact on predicting classes, and the best way to do this is by asking the models only.

Classifiers like Random Forest, Gradient Boosting Classifiers and AdaBoost offer an attribute called 'feature\_importance\_' with which we can see that which feature has more importance compared to others and by how much



**CHAPTER 6: MODEL BUILDING**

Models that we choose for this dataset are:

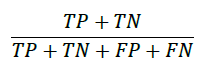
1. Random Forest Classifier (RFC)
2. Logistic Regression
3. Fully grown Decision Tree
4. Decision Tree

**6.1 Evaluation Metrics for the model:**

The evaluation metric we would like to choose is Accuracy.

Accuracy mean how many data points/observations are predicted correctly out of all number of observation.

In a Confusion Matrix point of View Accuracy is calculated by:

****

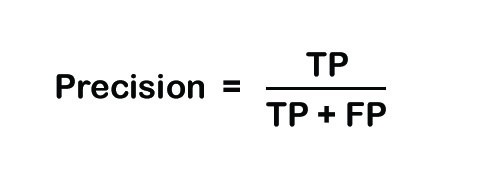
Where:

|  |  |
| --- | --- |
| *TP* | True Positives |
| *TN* | True Negatives |
| *FP* | False Positive |
| *FN* | False Negative |

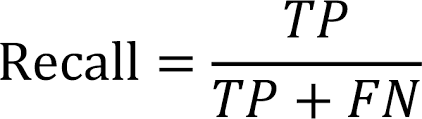
True Positives and True Negatives are the ones which the classifier correctly predicts the class of an observation.

So adding TP and TN gives us all the correct values predicted by the classifier and dividing by the sum all possible states is the same as dividing by total number of observations. This gives us an output value ranging from 0 to 1 where 0 means the classifier has 0% accuracy meaning it has classified all classes wrong and 1 means the classifier has 100% accuracy classifying all classes correctly.

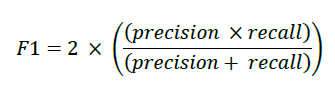
Another metric called is Precision. It is the number of positive prediction values divided by the total number of positive class values predicted. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.



Recall is another metric that evaluates total number of positive predictions divided by the number of positive class values. It’s the intuitively the ability of the classifier to find all the positive samples.



F1 score is the balance between Precision and Recall. It is the weighted average of the precision and recall.



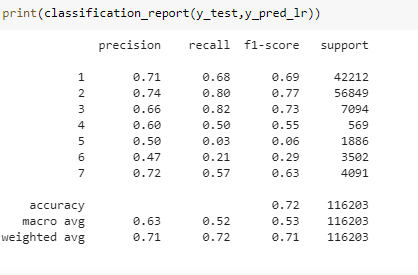
So we will be using Accuracy and F1 score as my evaluation metrics for this project.

**6.2 Classification Results:**

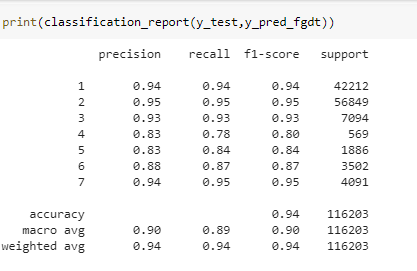
Logistic Regression (LR), Decision tree (DT) and Random Forest (RF) classifiers using five fold cross validations are used to classify. The Accuracy, Precision and F1-Score are presented for each classifier:

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Train accuracy** | **Test accuracy** |
| LR | 71.88% | 71.94% |
| Fully grown DT | 100% | 93.95% |
| DT | 93.91% | 90.23% |
| RF | 96.39% | 93.93% |

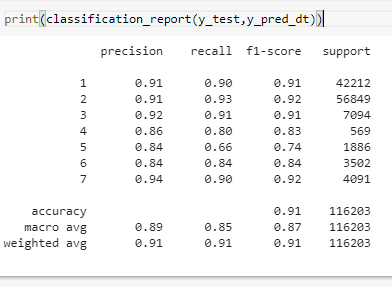
Logistic regression report:



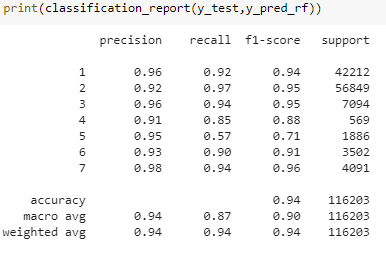
Fully Grown Decision Tree report:



Decision Tree report:



Random Forest report:



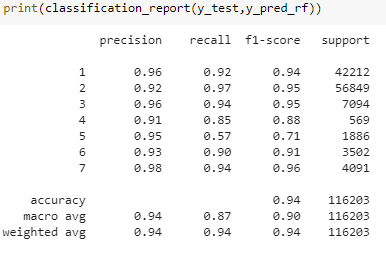
**6.2 Final Model:**

We Compare Accuracy and F1 score results of models and then choose the best model having best evaluation scores.

But it’s also not necessary that a model with highest accuracy is the best model as it might be not doing well predicting positive classes. A model with accuracy 80% might be better than a model with a accuracy score of 90%. This is called the Accuracy Paradox.

Out of these 5 models, RFC will give us better results than any. RFC can perform very well handling high dimensional data.

All models have its pros and cons, and our final solution model will be Random Forest.



**CHAPTER 7: CONCLUSION AND SUMMARY**

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 good training accuracy and test accuracy.

**Tree models are your best friends in multi-class prediction problems.**

Spruce/Fir, Lodgepole Pine and Krummholz are mostly found in Rawah, Neota and Comanche Peak Wilderness Area.

Cache la Poudre Wilderness Area is perfect place for Ponderosa Pine and Cottonwood/Willow.

If you see an Aspen suspect that you might be at the Rawah or Comanche.

Douglas-fir is an easy-going species, that goes along with any wilderness area.

**CHAPTER 8: REFERENCES AND BIBLIOGRAPHY**

[1]“Kaggle.”[Online].Available:https://www.kaggle.com/c/forest-covertype-prediction/forums/t/10693/features-engineering-benchmark

[2] “Uci repository database.” [Online]. Available: <https://archive.ics.uci.edu/ml/machine-learningdatabases/covtype/covtype.info>

[3] J. A. Blackard and D. J. Dean, “Comparative accuracies of artiﬁcial neural networks and discriminant analysis in predicting forest cover types from cartographic variables,” Computers and electronics in agriculture, vol. 24, no. 3, pp. 131–151, 1999.

[4] B. Chandra and V. Pallath Paul, “Prediction of forest cover using decision trees,” J. Ind. Soc. Agril. Statist, vol. 61, no. 2, pp. 192–198, 2007.