Report Of Forest Cover Type Prediction

Rupali
Department of Computer Science
IIITD
Delhi,India
rupali19095@iiitd.ac.in

Abstract: The process of prediction in data mining involves different steps. This document is about these steps involved such as visualization, preprocessing and prediction. There are different classifiers that can be used in that can be used in prediction algorithms i.e Naive Bayes Classifier, Random Forest Classifier, K-Nearest Classifier, Decision Tree Classifier and comapring there accuracies using different parameters.

I. Introduction

Forest Cover Type Prediction is a intresting area of classifiaction in US. This data is derived by USFS(US Forest Services) and US Geological Survey. Here goal is to predict the Cover Type of different Forests.Complete data set is in numerical format so any type of label encoding or One hot encoding is not used. Problem statement includes predicting the Cover Type of forest.

II. Data

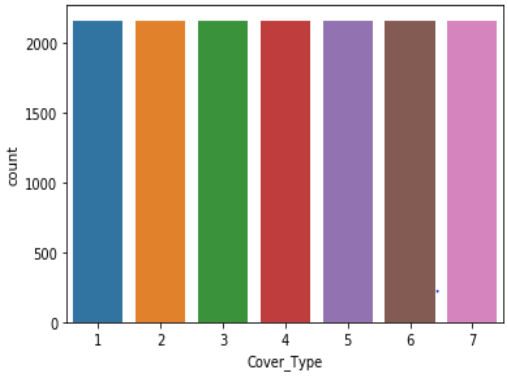
Data have been taken from kaggle Competition Forest Cover Type Prediction.The area on which this study was done is Roosevolt National Forest.Data consist of binary columns for some the attirbutes such as 'Wilderness' and 'Soli Type'.Attributes given are Elevation(m), Aspect(), Slope(), Horizontal Distance to Hydrology, Vertical Distance to Hydrology, Horizontal Distance to Roadways, Hillshade 9pm, Hillshade Noon, Hillshade 3pm, Horizontal Distance to Fire Points, Wilderness Area(4 column Binary data), Soil Type(40 Columns Binary Data), Cover Type.The training data consist of 15120 observations and 56 attributes in total. Testing Data Consist of 565892 Observations 55 attributes i.e expect Cover Type.

III Data Visualization: Exploratory Analysis

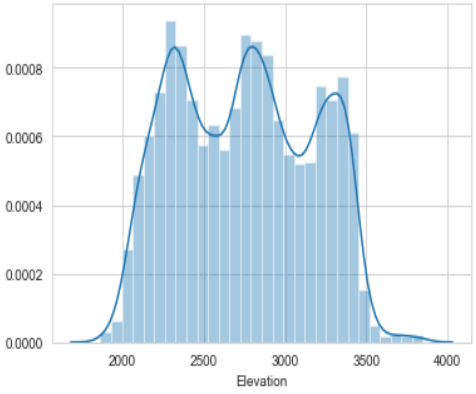
Null Values:

There are no null values in the data.As we can see using nullvalues() command for the dataframe.

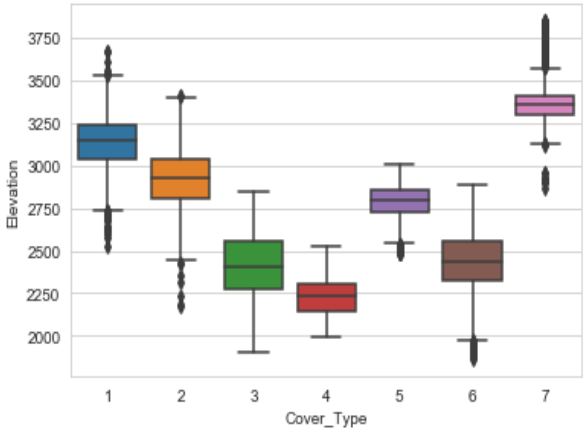
Records Of Each Cover Type: This is the graph for each Cover Type and its number of records. Intially data was unbalanced. I have used technique



Univariant Plots: These are the plots which are used for analysis of single variable.
Bivariant Plots: These are the plots which are used to analyse exactly two variables.
Univariant plot of Elevation:
This shows that how the density of records between the range of record values.As here maximum records have values 2000-2500.

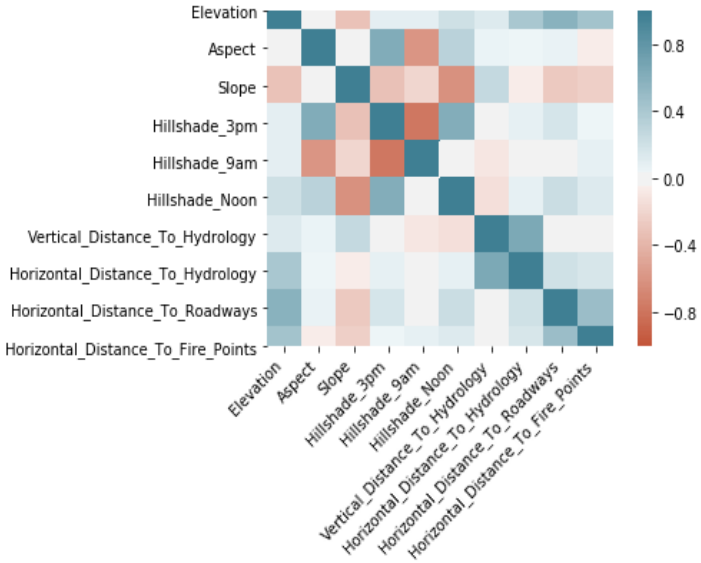


Bivariant Plot between Elevation and different Cover Type:



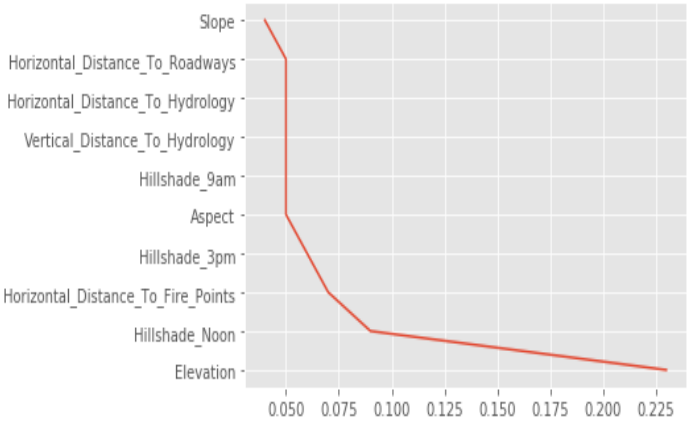
Here we can see that cover type 4 has minimum elevation whereas cover type 1 and 7 has maximum elevation.Similarly we have seen all the attributes w.r.t final Cover Type.

Correlation Matrix: It is taken as the measure to find how two attributes are related to each other.If this measure is 0 then the items are not correlated if it goes to 1 then they are highly correlated.
In the given data none of the given attributes is highly correelated so we can merge them to make any derived attribute.
Here no two attributes are highly correlated.



Standard Deviation: Here standard deviation of two attributes is exactly equals to 0 which means their values remains almost constantly equal to 0. So we can drop the attribute Soil Type15 and Soil Type 7

Feature Selection: When using Regression Classifier for feautre selection. I have come to know that Elevation is the most important feature on which cover type is going to depend.



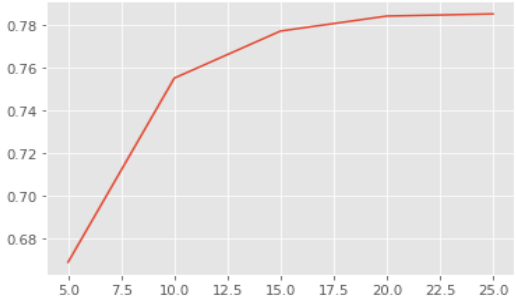
As feature importance for Elevation is 0.225 which is maximum.And so we can see top 10 features important to the data provided.

Score	Features
0.23	Elevation
0.09	Hillshade_Noon
0.6	Horizontal_distance_to_fire_points
0.07	Hillshade_3pm
0.05	Aspect
0.05	HillShade_9am
0.05	Vertical_Distance_To_Hydrology
0.05	Horizontal_Distance_To_Hydrology
0.05	Horizontal_Distance_To_Roadways
0.04	Slope

Table for Feature Score

IV Analysis of Different Classifiers

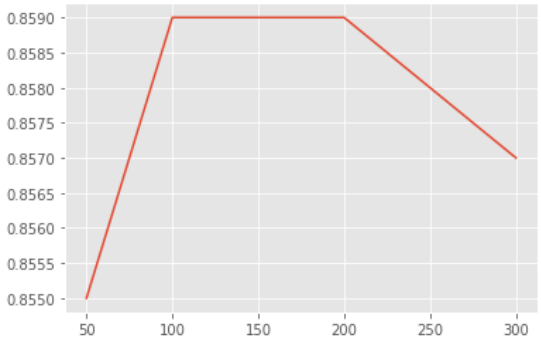
Decision Tree Classifier: First Decision Tree Classifier is used as number of attributes that are independent to each other are very less. Now the accuracy plot for this graph is as follows:



Height	Accuracy
5	0.669
10	0.755
15	0.777
20	0.784
25	0.785

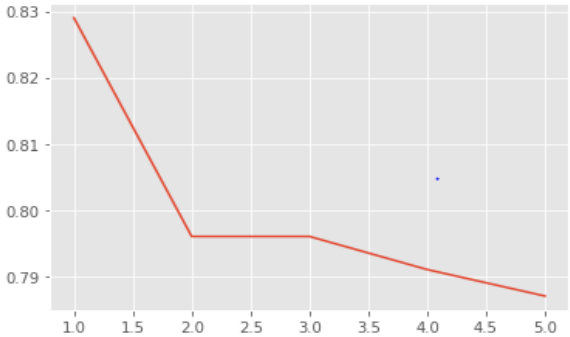
Accuracy will increases with Height of decision tree and becomes constant after a specific height.

Random Forest Classifier: Even after increasing height of decision tree accuracy does not increases so I used Random forest classifier which to give better accuracy.Graph of accuracy vs number of estimators is shown below:



N_estimator	Accuracy
50	0.855
100	0.859
200	0.859
250	0.858
300	0.857

KNN Classifier: This is used as it gives output based on simillarities of data points. Accuracy graph vs k nearest neighbors is shownbelow:

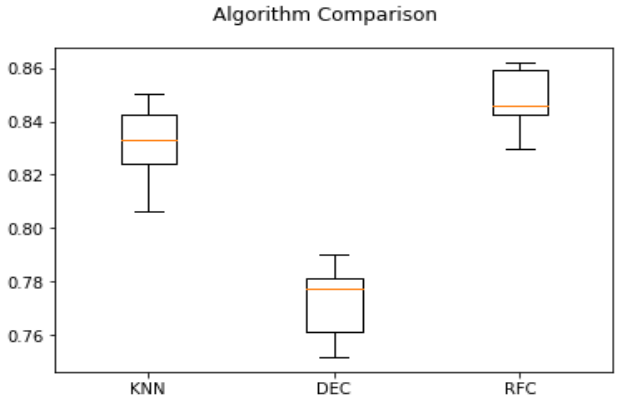


K-value	Accuracy
1	0.829
2	0.796
3	0.796
4	0.791
5	0.787

V Comaparing Diffe

Classifiers are Comapred using Kfolds method.As we kno compare using cross validation score.

Accuracies of different classifiers at best parameters using



References:

- https://towardsdatascience.com/dealing-with-multicla
- https://www.researchgate.net/publication/28265559
  iables
- https://seaborn.pydata.org/examples/many_pairwise
- https://www.kaggle.com/briantc/forest-cover-training
- https://www.kaggle.com/sociopath00/forest-type-pre
- **https://towardsdatascience.com/better-heatmap**
  41445d0f2bec
- **https://towardsdatascience.com/random-forest-**
- **https://machinelearningmastery.com/compare-**
- **https://www.geeksforgeeks.org/decision-tree-im**