GROUP NO. 10: SEMESTER PROJECT DRAFT REPORT

Group Members: Rohit Arora, Abhishek Kumar Agrawal, Nisarg Gandhi, Tyler Stocksdale.

Title: Forest Cover Type Prediction

1. Problem Statement

To predict the forest cover type, the 7 predominant kind of tree cover namely - Spruce/Fir (1)*, Lodgepole Pine (2)*, Ponderosa Pine (3)*, Cottonwood/Willow (4)*, Aspen (5)*, Douglas-fir (6)*, Krummholz (7)*, from strictly cartographic variables by evaluating four Wilderness areas in the Roosevelt National Forest located in the Front Range of northern Colorado, USA.

**In subsequent document we have used number in brackets to represent the corresponding cover type.*

1.1. What is Given?

For our project we obtained the data from Kaggle.com (also hosted in UCI repository), under the data set tilted "Forest Cover Type Prediction". This data was primarily collected and derived from the US Geological Survey and United States Forest Service.

Our dataset consists of 54 attributes (excluding Cover Type and ID) which consist of 40 different Soil Types, 4 different Areas of Wilderness, Elevation, Aspect, Slope, Horizontal and Vertical distance to Hydrology, Horizontal Distance to Roadways & Fire Points, and Hillshade at 9am, Noon, and 3pm. The *training set* consists of 15120 observations. These data records include all of the cartographic variables plus an indicator of the correct cover type. From this training data, we are to predict the cover type for the *test data set* of 565892 observations. The test data set has the same cartographic variables, but does not include the cover type.

1.2. Constraints

We identified following constraints at the onset of the project:

1. Difference in size of Training and Test Data Set: The ratio of training set to test set is 1:37, as a result we had less data to train our model to completely represent the variability in test data.
2. Different Attribute Types: Our data consist of both Qualitative and Quantitative values.
3. Missing Remote Sensed Data: Remote Sensed data provide useful information for such kind of classification problems which can be classified using Semi Supervised Learning or KNN.

2. Background

The original study (Blackard, Jock A. and Denis J. Dean, 1999[1]) compared two alternative techniques for predicting forest cover types. Their results of comparison indicated that Feed-Forward Artificial Neural Network performed better than Linear Discriminant Analysis.

In the Artificial Neural Network (ANN) based approach, the training data was randomly divided into 3:1 ration (1620 observations for training per cover type, and 540 observations for validation per cover type), and the Training, Validation and Test data were together linearly scaled in the range 0 to 1. The study used 54-120-7 based ANN architecture with Learning Rate=0.05, Momentum Rate = 0.5 and utilized Backpropagation learning algorithm for training Neural Network. The prediction Accuracy for ANN base model was found to be 70.58% on test data set.

In Linear Discriminant Analysis (LDA), the same data bifurcation was performed as in ANN, however there were two major differences:

i. Validation Set was not utilized, only training set was used for training model.
ii. Response variable (Cover Type) was coded as single variable that assumed value between 1 to 7, as opposed to ANN where response variable were coded as a series of seven binary variables.

The overall prediction accuracy over test dataset for LDA was found to be 58.38% which is less than ANN's prediction accuracy of 70.58%.

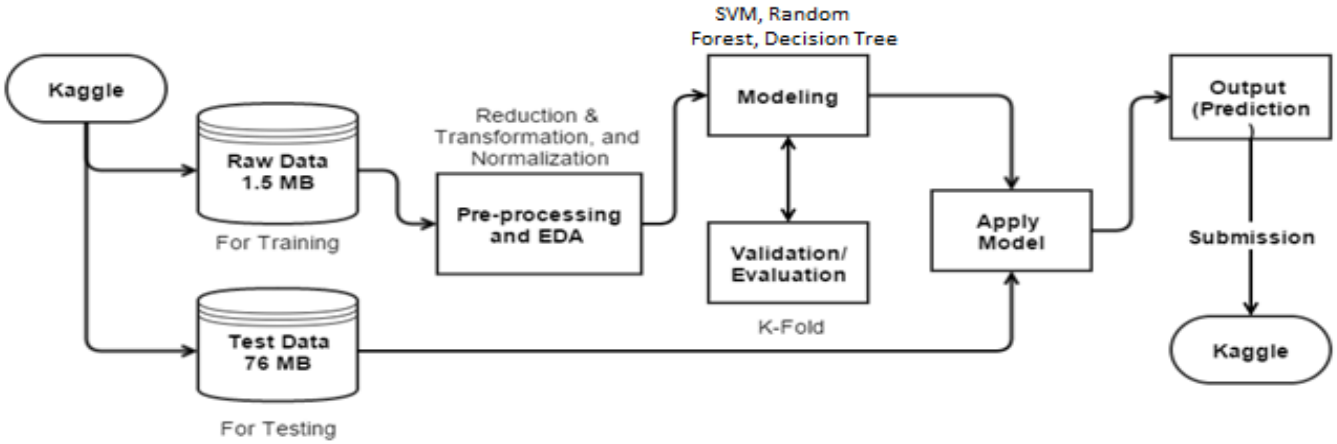
2.1 Our Learnings

The study (Blackard, Jock A. and Denis J. Dean, 1999[1]) provided us some better understanding of the data which helped us during initial Exploratory Data Analysis stage. We learnt about Artificial Neural Network, which is one of the popular classification techniques and has not yet been covered in the class.

For the purpose of our project we experimented with several State of the Art Classification techniques some of which we have studied as a part of our curriculum (Random Forest, Decision Tree, KNN, Naïve Bayes, and Rule Induction) and some popular ones which showed positive indications (such as SVM, Generalized Boosted Regression). And, our results using SVM and Random Forest were found to be better than ANN.

3. Our Contribution, Experiments & Claims

For detailed understanding of our contribution and project execution steps can be mapped to the below workflow diagram. The subsequent sub-sections further explain each component of the diagram.



3.1 Data

In the section 1.1 of this report we have mentioned about the source, nature and size of our training and testing data set in detail.

3.2 Pre-processing

In the pre-processing step we have adopted various Reduction, Transformation and Normalization techniques.

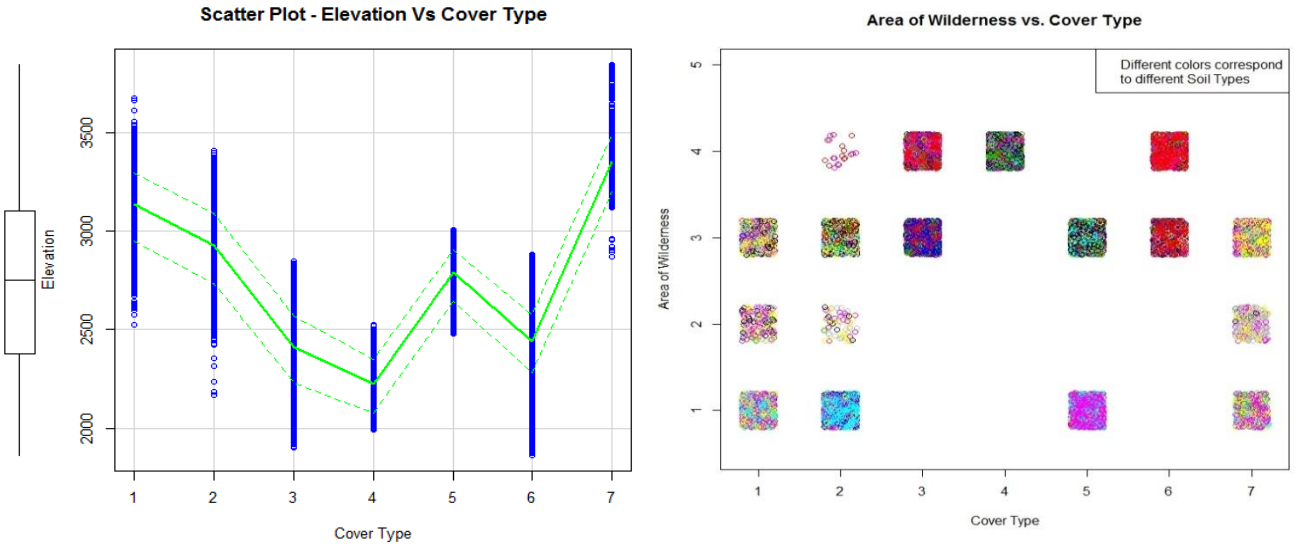
i. **Reduction & Transformation:** In this step we have converted the Nominal sparse data, i.e. Wilderness Area, and Soil Type to an equivalent decimal nominal value. Then, we transformed

these attributes and Cover Type to alphanumeric equivalent like (S1, S2, S3, etc.), (A1, A2, A3, etc.), and (C1, C2, C3, etc.) for Soil Type, Wilderness Area, and Cover Type respectively.

ii. **Normalization:** In the next step we normalized the quantitative variables using the transformation x'=(x-μ)/σ to create a new variable that has means 0 and standard deviation 1. Because the mean and standard deviation of the training data and test data differ, we chose to use a different transformation to achieve better accuracy in our classification. We used the transformation x'=(x-min)/(max-min) where max and min were selected from the combined training and test data. This ensured that classification algorithms created using the training data were consistent when used on the test data.

3.3 Exploratory Data Analysis[#]

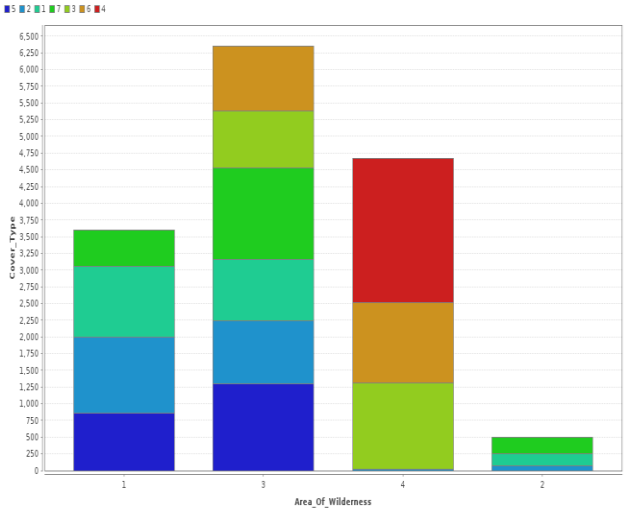
The purpose of the exploratory data analysis was to find comparative significance of each attribute in estimating Cover Type and grouping of some of the attributes which might display similar behavior such as Soil Type. We did this by determining the relationship between attributes themselves and between attributes and Cover Type (in our train data). This was done by finding correlation matrix, and making plots such as Scatter plots, Jitter plots, Stacked Bar charts, Scatter Matrix, and Box plots.

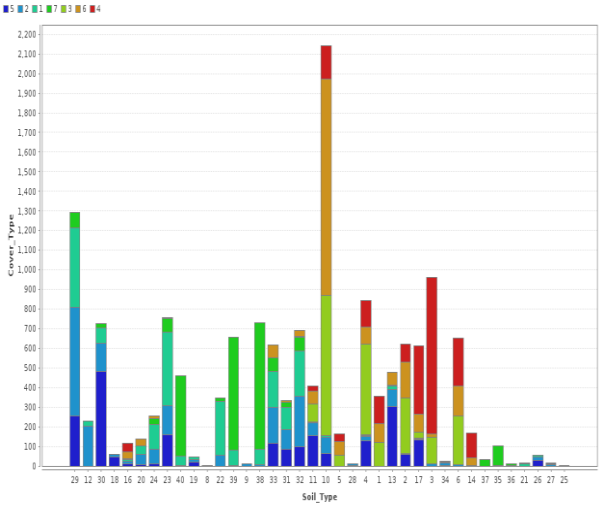


In the Scatter plot above the x-axis represents seven Cover Types. The plot shows the Elevation corresponding to each cover type, such as Cover Type 7 has elevation greater than 3000 only. On the Left of label 'Elevation' is the Box Plot for Elevation that shows the distribution of elevation, Range, Median, Maximum and Minimum Values. The strong and dotted green lines represent median and interquartile range of elevation respectively for each cover type.

The Jitter plot represents the only three nominal variables in our training data set. The plot compares the Cover Type (x-axis) with Wilderness Areas (y-axis). These are both nominal data types represented by integers. On a traditional scatter plot these data records would be plotted on top of each other, but in this plot, a "jitter" was added to each data record. In this way, the density and number of records in each category can be visualized. In addition, each data record is colored according to its Soil Type. Due to this, it is visually apparent when one category has a multitude of soil types, or a majority soil type. This one plot provides us with very useful insight on the three nominal variables that exist in our data set.

[#] *The plots for EDA can be found in the folder 'plots/eda' in the zip file.*





The stacked bar plots are colored based on cover type. Each color corresponds to a different cover type. Each bar corresponds to an area of wilderness or soil type. These plots provide us with a very useful visualization of distribution of cover types with respect to areas of wilderness and soil types. For example, it is apparent that cover type 4 is only located in area of wilderness 4.

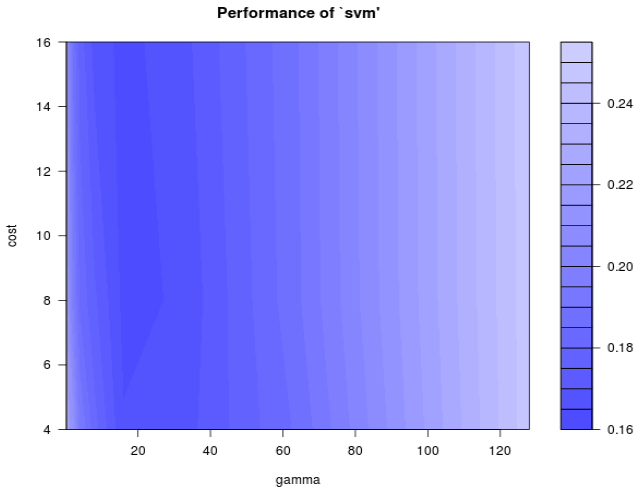
3.4 Modeling, Validation and Evaluation[$]

After data pre-processing and EDA we proceeded with model building and validation process. We experimented with several State of the Art methods to find suitable classification technique for our data. Some of these models were tuned and then used for the prediction of actual test data. These predictions validated against the Kaggle submissions (ground truth) are shown below.

Classifier	Accuracy	Top Recall	Top Precision	Test Data Accuracy(Kaggle)
KNN	83.31%	4 > 7 > 5 > 6	7 > 3 > 4 > 5	69.67%
SVM	84.80%	4 > 7 > 5 >6	7 > 4 > 5 >3	**76.69%**
Decision Tree	78.28%	4 > 7 > 5 > 6	7 > 4 > 5 >3	58.89%
Naive Bayesian	66.20%	7 > 4 > 5 > 1	7 > 4 > 5 >3	Not Submitted
Random Forest	88.23%	4 > 7 > 5 > 3	7 > 4 > 5 > 6	**76.6%**
Gradient Boost Model	87.12%	4 > 7 > 6 > 5	7 > 3 > 4 > 5	69.82%
Rule Induction	76.41%	4 > 7 > 5 > 6	4 > 7 > 5 > 6	58.32%

Out of the above classifiers, we chose below classifiers for further parameter tuning.

i. **Support Vector Machine(SVM):** The motivation behind selecting SVM classifiers is to obtain linearly separable hyperplanes in order to predict the cover type accurately. We chose Radial Basis kernel Function (RBF) as the default kernel function for the process. After performing grid search[2] over SVM RBF kernel we obtained the range of best parameters which is shown in the plot above (on right). The vertical bar indicates the error rate w.r.t. to the gradient.



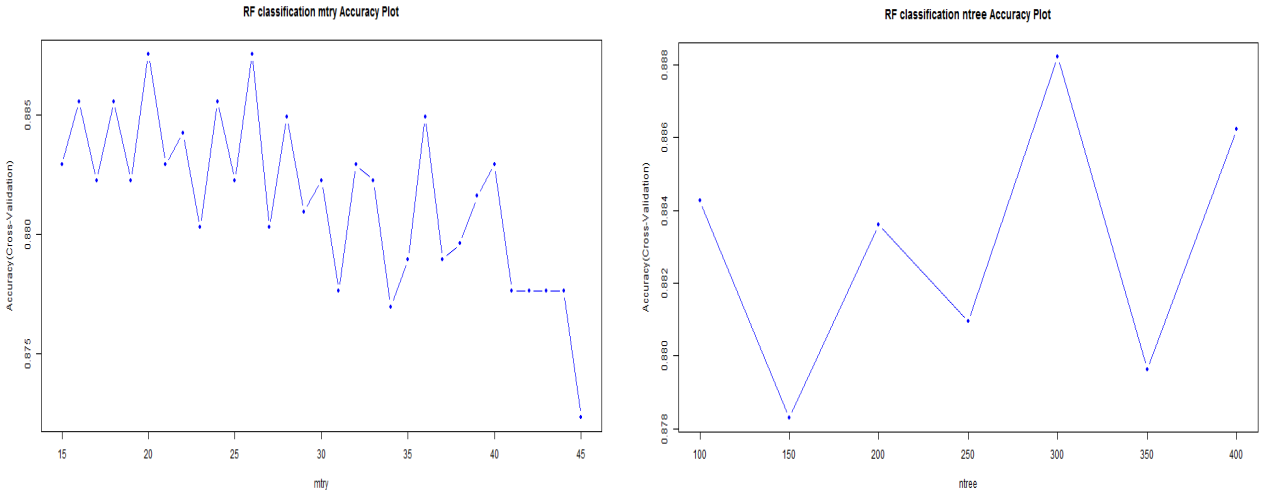
[$] *The plots for Modeling, Validation and Evaluation can be found in the folder 'plots/classifiers' in the zip file.*

ii. **Random Forest:** The motivation behind Random Forest is that it grows many classification trees based on the number of random attributes (mtry and ntree). The classifier chooses the class label[3] having the most votes (amongst all the trees in the forest). This ensemble method guarantees not to overfit and outputs the class with the majority votes, thus reducing the misclassification error. Also random forests can be used to rank the importance of variables in a classification problem in a natural way[4].

Two important parameters, on which the accuracy of the Random Forest depends, are:

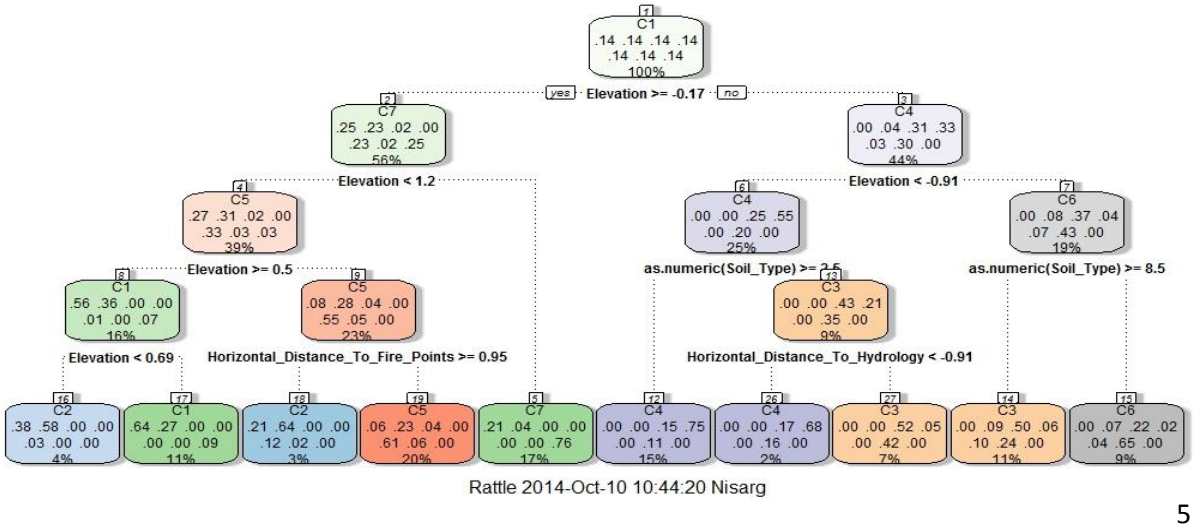
a. **ntree** = number of trees in the forest
b. **mtry** = number of random attributes chosen for each classification tree

Following is the graph of the grid search made on both the parameters to get the best accuracy classifier:



Using these tuned parameters, ntree = 300 and mtry = 20, we achieved the cross validation accuracy of 88.88%. The Kaggle test set submission achieved 75.6% accuracy.

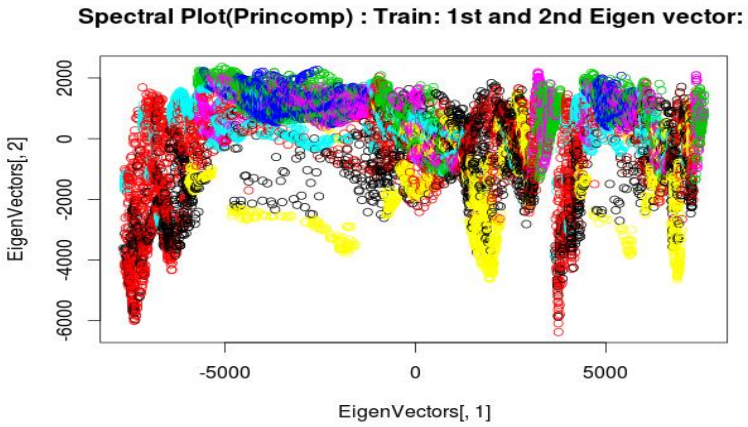
iii. **Decision Tree Classifier:** The motivation behind selecting Decision Tree is to identify the principal attributes in the dataset which are identified across the various levels in decision tree. Also, decision tree algorithms have proved useful in the past for classification of numeric as well as categorical attributes. This method uses the criteria of 'Information Gain' to split the attributes. Below is the decision tree generated for the forest cover type data.



Our 10-fold cross validation results indicated that the Decision Tree classifier has an accuracy of 81% on the trained data and 58.89% in Kaggle submission. We also obtained useful attribute weights in the decision tree tuning process. Elevation(**41**), Soil Type(**33**), Horizontal Distance To Roadways(**10**), Horizontal Distance To Fire Points(**5**), Area Of Wilderness(**4**), Horizontal Distance To Hydrology(**3**), Hillshade 3pm (**1**), Hillshade 9am (**1**), Vertical Distance To Hydrology (**1**), Aspect (**1**) etc. These weights and features were used for feature selection and weight parameters for other classifiers.

3.5 Other Work

Apart from modeling we also performed *Principal Component Analysis* over the training data, and the figure on right represents spectral plot for the first two eigenvectors of the data with colors indicating different cover-type. More detail on the spree plots of the variability of eigen-vectors can be found inside the folder 'plots'.



4. Conclusion & Future Work

We explored and implemented several State of the art methods which enhanced our data analysis skill set that we utilized to improve prediction accuracy of our test data set. Support Vector Machine and Random Forest classifiers were examined in depth by applying various parameter tuning techniques in order to find best fitting parameter values. We obtained 76.69% of test accuracy using SVM classifier which is an improvement over the results of (Blackard, Jock A. and Denis J. Dean, 1999[1]) which as 70.58%.

However, based on Classification accuracy obtained by other teams on Kaggle we understand that there is scope of further improvement by gain more indepth understanding of other advance classifiers as Deep learners, Adaboost, Kernelized SVM etc. Another scope of improvement can be different attribute selection techniques that can be used for tackling the problem of skewed distribution in the output class. Our work in Principal Component Analysis for dimensionality reduction and decision tree attribute weights outputs can be used for feature generation and important attribute selection.

Given the problem statement our end results are very promising. The methods and techniques applied during the process of experimentation can be further generalized for other multi-class classification problems.

5. References

[1] Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables (2000) by J. A. Blackard and D. J. Dean. In: Computers and Electronics in Agriculture 24(3), pp. 131-151.
[2] "A Practical Guide to Support Vector Classification"
[3] "Random Forests Leo Breiman and Adele Cutler" Random Forests. Web. 16 Nov. 2014.
[4] "Using GBM for Classification in R"
[5] "Random Forest" Wikipedia. Wikimedia Foundation, 14 Nov. 2014. Web. 16 Nov. 2014.