Forest Cover-Type Prediction

Barath Raja (19211063), Hithesh Sai K B (19210864),

Priya Dharshini B (19210538), Rebekah Jennifer (19210518)

Department of Computing

Dublin City University

Dublin, Ireland

1barath.raja2@mail.dcu.ie

2hithesh.khandheribalajee2@mail.dcu.ie

3rebekah.manimaran2@mail.dcu.ie

4priya.balaji2@mail.dcu.ie

Abstract— Predicting forest cover type in forests and natural reserves provide an advantage in the conservation and management of nature. The process of measuring and recording the cover types is time-consuming and costly in some situations. In these situations, predictive models provide an alternative method for obtaining data. In this study, we aim to predict forest cover types from cartographic variables using various machine learning models.

Keywords— cover-type, data mining, forest types, ExtraTreesClassifier, KNN, RandomForest, Logistic Regression

1. INTRODUCTION

Forests play a significant role in the economic, environmental and social roles in the development of a nation. In major circumstances, the contribution of the forest-based sector has been minimal compared to its potential. But, these forest lands serve as an escape valve to a majority of the population without access to agricultural land. There has always been a concern about the effects of global warming, the loss of genetic material and forest fires in these forests and natural reserves. This contributes to various problems like soil erosion, diminution of water sources and the elimination of wildlife habitats. [1] Such situations affect the livelihood of the people who depend on the forest. To avoid these types of disasters, the natural resources information is stored in the federal land management agencies for inventory management. Forest cover type is one of the basic information that is recorded in these inventories. With this information, there is so much potential for research in the fields of environmental conservation, flora and fauna research, and geological studies. In this study, we aim to predict forest cover types from cartographic variables using various machine learning models.

One of the major challenges in machine learning identifying the right algorithm to get the expected results for the dataset, here we are experimenting with various algorithms like Random Forest Classifier, K-nearest neighbor and Logistic regression.

1. RELATED WORK

There have been many research and various methods to predict and classify forest cover types that can help in further research of forest fire susceptibility, the spread of the infestation [2], and other deforestation problems.

A. Literature Survey

In our recent studies on forest cover type prediction, they have used a dataset from the UCI Machine Learning Repository where 15120 samples of 30\*30 patches of Roosevelt National Forest [1]. In these implementations of predicting forest cover type, they have used 54 cartographic features [1] and also by removing the 44 Boolean features and making them into dimensions of 10 features of the data. The features and labels include the elevation, hydrologic, soil, and sunlight and the 7 cover types. In the study, they have implemented a variety of classification algorithms such as Multi-class support vector machine and K-Means Clustering using Principal Component Analysis. The principal component analysis is a method to reduce the dimensions of the data by making the mean to zero and variance to one. This has been visualized in three dimensions for 8000 samples. When applying the data with reduced dimensions, the runtime of multi-class SVM has also been reduced but the loss of the variance will decrease the performance.

In Multi-Class SVM [1], the data is trained using the Boolean and without Boolean information where the 7 forest cover types are classified into 21 separate binary classifiers to predict the cover types of trees in the wooden area. After training the model it has been tuned with two hyper-parameters to produce better accuracy using grid search and 10 cross-validations [1]. The results of the model obtained are 81.35% training and 78.24% testing accuracy. Also by removing the Boolean features accuracies dropped to 75.21% and 72.75%.

Likewise, K Means Clustering [1] has been used for the same dataset to classify the cover types, here the data is grouped into clusters where the model is developed without the labels of the data. Each of the clusters is observed and named based on the most common cover type. This has been run for 10 times for better accuracy. The results showed that when k=7 for each of the 7 cover types performance was very poor and once the number of clusters has increased the test error reduced with 0.38 for the complete dataset and 0.55 for dataset without Boolean parameters.

The study evaluated that reducing the dimensions using PCA and transforming the data from 54 features to 10 features with Multi-Class SVM and K-Means Clustering performed worse in training and testing than using the entire dataset. Although the positive aspect would be this work reduced the overfitting and demonstrated lower generalization error.

In one of the previous studies and experimentation of using artificial neural networks and discriminant analysis [3] says that the results of the feed-forward artificial neural network model predict more accurately about the forest cover type than the traditional statistical model based on Gaussian discriminant analysis [3]. In the approach of ANN one hidden layer and backpropagation learning algorithm is used with mean squared error (MSE) function. The 54 input variables are analyzed for the reduction process to identify the variables that did not contribute to the overall predictive capability of the system. The experiment showed that 150 hidden nodes were used to minimize the MSE with the best learning rate and momentum rate of 0.05 and 0.5. Also, the classification accuracy of the prediction model was 70.58%.

The second approach in the discriminant analysis [3] is implemented based on two main assumptions. One being the data distributions of all dependent and independent variables are normal and second is the covariance matrix for different groups are equal. The classification accuracy for the discriminant analysis model was 58.38%.

The results of this study and experiments conclude that the ANN model outperforms the DA model in the prediction of forest cover type. The negative aspect implies that both models misclassify ponderosa pine, Douglas-fir, and cottonwood/willow cover types with each other. This is because of the geographic proximity of the different cover types. Also, another factor that impacts the approach of both the classification models is the amount of computational time that is required to develop the prediction.

The methodology used in this project management process is CRISP -DM [4] (Cross Industry Standard Process for Data Mining). The application of using this in our research is discussed below with exploratory data analysis.

1. PROCESS FLOW

In Phase 1, the Business perspective on the application of classifying forest cover types from cartographic variables is studied and analysed. The advantages of predicting forest cover type trees in the wooden area help in conservation and proper management of forest trees without any fire, infestation [2] and disasters. The business goal on this application is to understand which trees species grow predominantly in what kind of wilderness area with data collected from hill shade, slope and soil aspects for the challenges of US Forest management services. This is implemented by understanding the seven forest cover types from four different wilderness areas in Roosevelt National Forest of Northern Colorado.

The next step is the data understanding phase where all the features are analysed with verification on the quality of the data and finding the outliers. There are 54 features with one target variable with 581012 instances. This is a Multi-class Classification for seven discrete categories of forest cover types. From the descriptive statistics of the dataset, we could say that the features of the data are complete and numeric without any missing/null values. It includes 10 continuous variables and 44 Boolean variables with one-hot encoded columns such as soil type and wilderness area. Along with this an anomaly is identified and checked in cover type’s data frame on Vertical Distance to Hydrology column. This feature explains the vertical distance from the nearest surface water where negative values are displayed. These values show that the nearest surface water is below the sea level so the values shall remain the same.

During the Data Preparation stage, the goal is to focus on data transformation and feature selection. We analysed the different features and their correlation with each other using python libraries but there were some challenges on the measurements used for different features. To resolve the issues we reversed our process flow to the data understanding stage to find that meters and degrees are used for measuring the input features which leads to an unrelated wide spread of data. All the feature inference are explained in EDA.

*a) Exploratory Data Analysis (EDA)*

On performing the univariate analysis on the features, it has a min. value of 0 excluding ‘Elevation’ and ‘Vertical Distance to Hydrology features’. The latter, has the lowest value, being negative, these values show that the nearest surface water is below that data point or it is below the sea level. The mean value of these features varies from 14 to 2959. In the features, 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). Features like Aspect and Slope are measured in degrees so its maximum value can't go above 360. And Hill shades features can take on a max value of 255.

The correlation between and among the features are explored to better understand underlying relationships in the data using a correlation matrix: The features which are strongly correlated with a threshold above 0.5 can be found below: **a.** ‘Elevation’, ‘Horizontal Distance To Fire Points’, ‘Wilderness\_Area1’ and ‘Wilderness\_Area4’ have stronger correlations with the target cover type. **b. ‘**Aspect’, ‘Slope’ and ‘Distances to Hydrology’ have weak correlations with the target cover type. **c.** ‘Wilderness\_Area3’ has a higher correlation with the ‘Aspen’ cover type. **d.** Soil types are distinct and unrelated to each other **e.**Certain soil types have higher correlations with certain cover types.

*b) Data Visualization*

In the next steps, we visualize some important features of the dataset, to get a better understanding, as inferred, the features consist of a mix of both categorical and continuous variables, one of the important features are the soil types, there are 40 types, in the four wilderness areas. To identify which of the top 20 soil types that affect the accuracy of the model. Implementing, ExtraTreesClassifier, an ensemble learning method based on decision trees. That randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. There are 40 soil types in the given data, by using feature importance we have plotted the top 20 soil types which have high importance for predicting the cover type. [2]

Analyzing the most significant soil types by representing them in the form of graphs is what the bar plot analysis does and can be observed by:

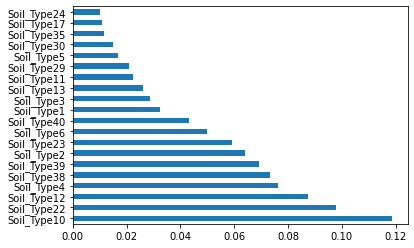


Fig. 1 bar plot illustrating the most significant soil types

The ‘Soil Type10’ is said to have the most significant effect on the model, it has the maximum feature importance of 0.12 approx., which is followed by the ‘Soil Type22’ with the values of 0.10.

On analysing, soil types are related to cover types and wilderness areas. On analysing, soil types are related to cover types and wilderness area.

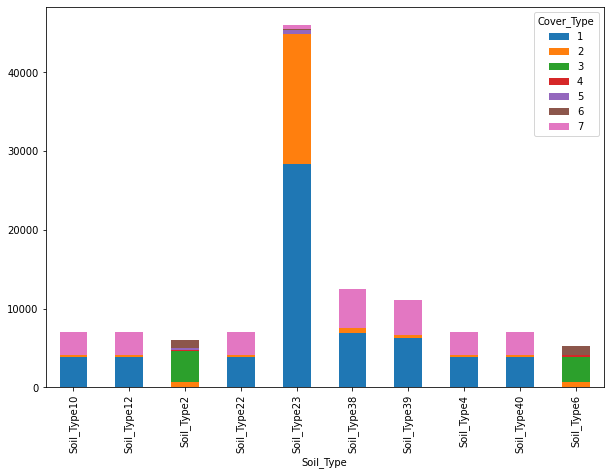


Fig.2. bar plot illustrating the pesence of various cover types in the most significant soil types

In figure 2, one can observe that the SoilType23 has the maximum cover of the wilderness area. It can also be visualized that the cover type Spruce/Fir, Lodgepole Pine and Krummholz are present in most of the top 10 soil types.

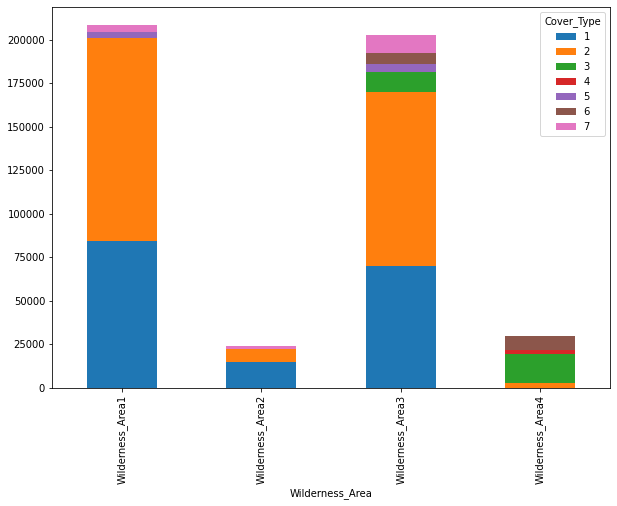


Fig.3. bar plot illustrating the pesence of various cover types in the wilderness areas

In figure 3. There are a total of four wilderness area. On visualizing we identify, wilderness area 1 and 3 has higher percentage of cover type than the wilderness area 2 and 4. And cover type 1 and 2 contribute highest number to wilderness area 1 and 3.

1. FEATURE SELECTION

To reduce the complexity of the dataset and to remove the features which do not impact on the prediction of the model, we perform feature selection. On identifying the skewness of the features, Soil\_Type15 has the highest positive skewness and is followed by Soil\_Type7, 36, 38. This is called right-skewed distribution, where the mode of the feature is to the left-most followed by median and mean. This means that most of the observations have will have 0 value for this feature. The other features like ‘Elevation’ and ‘Hillshade’ are having negatively skewed distribution, it's the opposite of the positively skewed distribution, where the mode is to the rightmost followed by median and mean.

In general, looking at skew scores of Soils it seems like we can reduce our dimensions by removing some Soil Types only if they don’t have any different information to give our models and improving its performance.

We remove the following soil features with varied skewness and occurrences less than 1000 to improve the performance,

('Soil\_Type7','Soil\_Type8','Soil\_Type14','Soil\_Type15','Soil\_Type21','Soil\_Type25','Soil\_Type28','Soil\_Type36','Soil\_Type37)

Thus, we have 44 features, with which we can perform out modelling.

1. MODELLING

*c) Applying Machine Learning Algorithms*

• Logistic regression:

Logistic regression is a classification model. It uses a logistic function to frame the binary output model. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

• K-nearest neighbors:

K-nearest neighbor is a non-parametric method used for classification and regression. In KNN classification, a majority voting is applied over the k nearest. The value of k is always odd. Computations happen the only runtime.

• Random Forest:

This algorithm for machine learning that performs both regression and classification tasks. Random Forest is a collection of decision trees and the average/majority vote of the forest is selected as the predicted output. This is less prone to overfitting than a Decision tree and gives a more generalized solution.

1. analysis and result

The performance of the models are evaluated using 2 different evaluation metrics accuracy and f1 score.

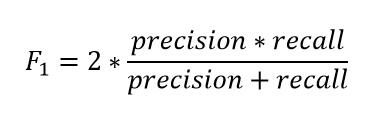
• Accuracy is the measure of the number of right prediction data divided by total number of observations hence giving a value ranging between 0 and 1, while 0 is no correctly predicted class whereas 1 is all correctly predicted class. We can multiply the result by 100 to get the accuracy score in terms of percent.

• F1 score is more useful than accuracy especially in the case where you have uneven amount of class distribution as in our case. It's the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

|  |  |  |
| --- | --- | --- |
| Vector Space | | |
| Model | Accuracy | F-1 Score |
| Random Forest  Classifier | 95.2% | 91.98% |
| K-Nearest  Neighbor | 96.52% | 93.11% |
| Logistic  Regression | 61.92% | 23.11% |

Fig.4. Table comparing various the accuracy and F-1 Score of different models

One can observe from figure 4 that the accuracy of K-NN model is high when compared with the other models, getting accuracy 95.2 % and 93.11% of F1 score. It takes the highest training and evaluation time. KNN has performed well predicting classes, n\_neighbors, which is the parameter of KNN, set to 5 by default. The F1 scores are more important as they give us an weighted average score of both precision and recall, where precision is intuitively the ability of the classifier not to label as positive a sample that is negative and recall it's the number of positive prediction divided by the number of positive class values and F1 score is the balance of both of these. F1 score can be calculated using,



Bar plot:

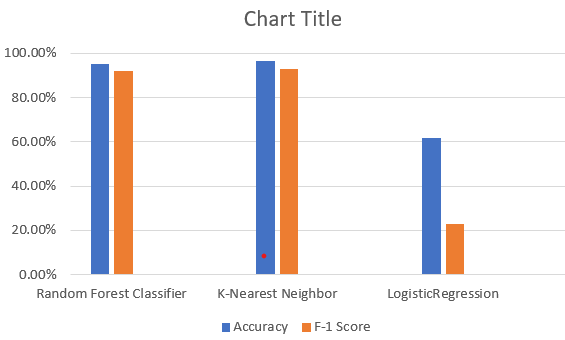


Fig.4.1 bar plot comparing accuracy and F-1 Score of different models

1. CONCLUSION

In this experiment, we observed that using feature engineering to select the necessary features for the model improves the accuracy of the model which also helped to reduce the training time considerably. Among the models used in the experiment KNN out performed the other classical models such as Random Forest classifer and Logistic Regression.

`References

1. Crain, K., and G. Davis. "Classifying forest cover type using cartographic features." Published report (2014).
2. D.A. Leatherman, Colorado State Forest Service entomologist (retired); 2/99. http://www.ext.colostate.edu/pubs/insect/05528.html Revised 9/11.
3. Blackard, Jock A., and Denis J. Dean. "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables." Computers and electronics in agriculture 24, no. 3 (1999): 131-151.
4. Chapman, Pete. "Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler),"." CRISP-DM 1.0. Step-by-step data mining guide (1999).