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
[1]barath.raja2@mail.dcu.ie
[2]hithesh.khandheribalajee2@mail.dcu.ie
[3]rebekah.manimaran2@mail.dcu.ie
[4]priya.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 the machine learning models.**

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

I. 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 machine learning models.

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

II. 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.

In our recent studies on forest cover type prediction, they have used a dataset from the UCI Machine Learning Repository [5] where 15120 samples of 30*30 patches of Roosevelt National Forest [1][6]. 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 accuracy 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 [8].

III. 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 [9] 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 minimum 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 the features and target variables are observed and explored. It is found that there is no strong correlation between any parameter and attributes. There is only very little correlation between numerical features and cover-type.

b) Data Visualization

With, data visualization we are analyzing the most important features that can enhance the prediction model. The dataset consists of both categorical and continuous variables so it is a major task to know which variables contribute more to the cover-type prediction.

From the dataset, we could visualize that (Fig. 1) the cover type 1 (Spruce/Fir) and cover type 2 (Lodgepole Pine) has the highest number of count among the other cover types.

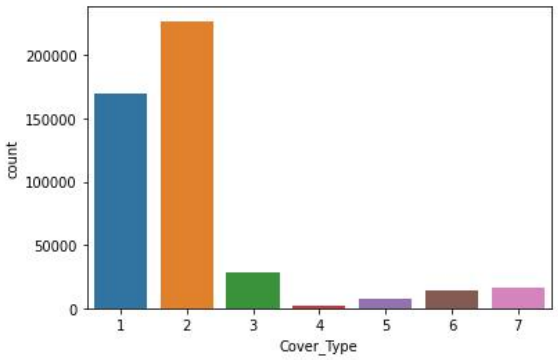

Fig. 1 – Bar plot of different forest cover-types

One of the important features are the soil types, there are 40 soil types in the 4 wilderness areas. To identify only the top 20 important features from 40 soil types we have used ExtraTreesClassifier, an ensemble learning method based on decision trees. This classifier improves the predictive accuracy and controls over-fitting on various sub samples of the dataset. The top 20 soil type (Fig. 2) identified from the feature importance property says that higher the value the more important is the feature.

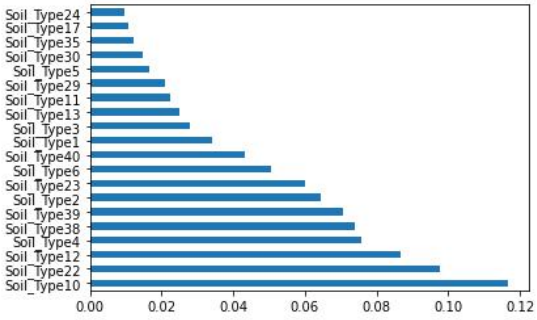

Fig. 2 – Illustrates the most significant soil types

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.
Now, we are visualizing how the cover type plays a role with the top 10 important soil types in the wilderness area.

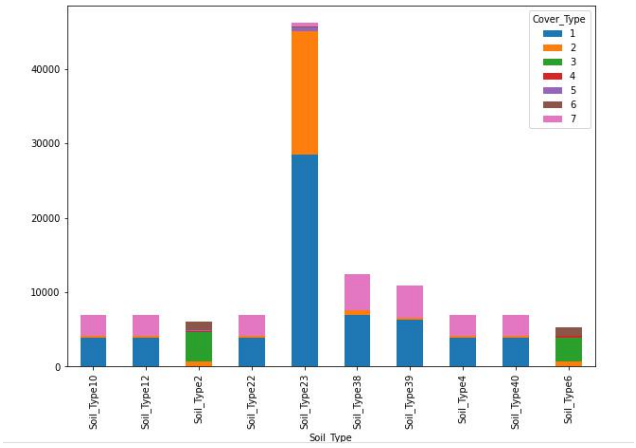

Fig. 3 – Illustrating the presence of cover types in the most significant soil type features

In figure 3, one can observe that the SoilType23 has the maximum cover of the wilderness area. It can be seen that cover type 1 (Spruce/Fir), cover type 2 (Lodgepole pine) and Cover type 7 (Krummholz) are present majorly with the soil types.

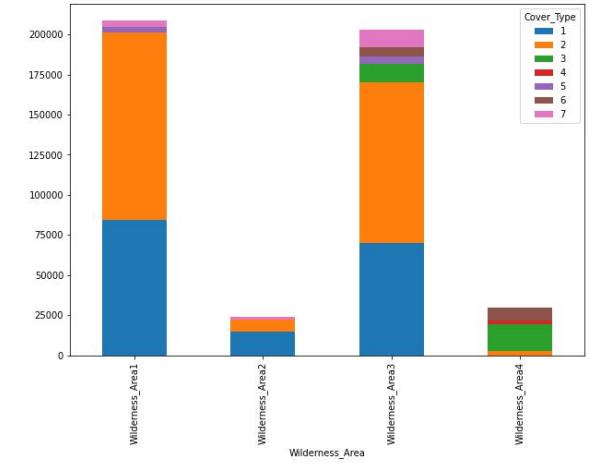

Fig. 4 – comparison of wilderness area with the 7 cover types

In figure 4, there are a total of four wilderness areas. On visualizing we identify, wilderness area 1 and 3 has a higher percentage of cover type than the wilderness area 2 and 4.

From the analysis of visualizing using bar plots, we can say that there are 20 most important soil type features that should not be eliminated because it impacts the predictive accuracy of the model. We can also say that cover type 1 (Spruce/Fir) and Cover Type 2 (Lodgepole pine) contribute more to the soil types and wilderness areas.

IV. FEATURE SELECTION

In order to make my model learn from relevant features, the feature engineering process is done. In this method, the right subset of features or attributes is selected that contributes most to the target variable. We have computed skewness for all the 55 features for our model prediction. 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 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.

By looking at the skewness value of soil types, we can reduce the dimensions of the Soil Types. Since these features improves accuracy, reduces overfitting and misclassification of forest cover-types.

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 our modelling.

V. MODELLING

c) Applying Machine Learning Algorithms

Now, In order to perform our model prediction using the best-fit algorithm we have used the re-sampling method with K fold validation. This cross-validation technique is used to test the effectiveness of the performance of the model in different machine learning algorithms. By using this technique we can generalize our model without any over-fitting or under-fitting issues.

Since our target variable is a multi-class classification, we are using the StratifiedKfold technique for equal distribution of our forest cover-type data. This technique builds a less biased model compared to other methods. In this K fold model, the original dataset will be appearing in training and testing set by splitting the data randomly using k-folds (we are using k = 10).

To come up with the right machine learning model we are using this method to test with logistic regression, decision tree classifier, K nearest neighbor classifier, Linear discriminant analysis, and Gaussian naïve bayes algorithm. This helps to identify the best fit unbiased model using K fold technique and evaluating with cross-validation score of each model. Now each algorithm performs K-Fold validation where the model is trained with K-1 folds (X_train, Y_train) and test using the remaining Kth fold (X_test, Y_test). This process is repeated and appended for 10 Iterations on each algorithm.

The performance metric is calculated using cross-validation score where the mean of all the accuracy and mean of all the standard deviation for each test folds are calculated. This technique produces the best results where the data is equally distributed for all the models in training and testing folds. Five different machine learning algorithms are selected for the following reasons.

d) Logistic Regression

The logistic regression classification model is used with one vs rest scheme for predicting 7 cover-types using the sigmoid function [7]. The model is regularized using the 'liblinear' library which can handle any kinds of input values. The accuracy obtained with the logistic regression algorithm is 70.9% using the 10-Fold cross-validation technique.

e) Decision Tree Classifier

The decision tree classifier is used for predicting model performance because it follows the divide and conquers algorithm. This algorithm splits the training data into subsets on the sunlight, soil type, elevation and aspect features, which are even split into smaller subsets until there is a stop in the process. This results in large information gain where the algorithm classifies extremely fast with unknown records. Thus the accuracy of this algorithm is 93.3%.

f) K-Nearest Neighbor

K-nearest neighbor algorithm is used because of the close proximity between similar input features. This algorithm works by calculating the distances between the query and all the training samples from the data. Since it is classification it returns the mode (most frequent label) of the K labels. This showed the highest accuracy of 96.5%.

g) Linear Discriminant Analysis

The linear discriminant analysis is a dimensionality reduction technique in order to avoid over-fitting, This algorithm computes the mean vector for the 44 features of 7 cover-type, calculates the scatter matrices between cover-types, also finds the eigen vectors and eigenvalues which gets transformed into matrices. The accuracy obtained with 10-Fold cross-validation is 67.8%.

h) Gaussian Naïve Bayes

Lastly, the Gaussian Naïve Bayes algorithm is used which is based on the Bayes theorem describes the probability of the 7 cover-types based on the prior knowledge of the input features. GaussianNB works with mean and standard deviation for the data distribution of 7 forest cover-types. The accuracy came upto 45.6 % using the 10-Fold cross validation method.

Model (10-Fold Val)	Accuracy
Logistic Regression	70.92%
Linear Discriminant Analysis	67.87%
K-Nearest Neighbor	96.52%
Decision Tree Classifier	93.39%
Gaussian Naïve Bayes	45.66%

Fig.5. Table comparing ML algorithms and accuracy

VI. ANALYSIS AND RESULT

After the analysis of various machine learning algorithms, it can be noticed that KNN model outperforms well with 96.52% accuracy and 0.000894 standard deviation (less variance).

Now, the KNN algorithm has been used for the entire dataset with the initialization of K as the number of neighbors. In our method, K initialization has been computed and visualized with the range from 1 to 7 by fitting the model and compiling it with the orginal dataset.

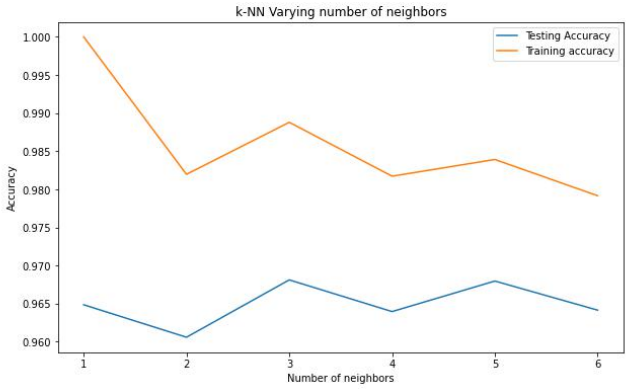

Fig.6 – KNN neigbors vs the training and testing accuracy

From Figure 6, we can conclude that K=5 is the nearest data point neighbor to calculate the euclidean distance between the training and testing dataset. Also, it will assign the most frequent class to the test data.

The Euclidean distance metric is calculated using this formula

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + \ldots + (x_n - x_n')^2}$$

The full analysis of forest cover-type prediction is done with K=5 as the nearest data point and the calculated accuracy score metric as 96.79%.

Later, The performance of the best fit KNN algorithm model is evaluated with accuracy score, classification report and confusion matrix. The precision, recall, f1-score and support are also the metrics produced from the classification report.

The accuracy of 96.79% tells about how perfect will the classifier be classified correctly using the True Positives, True Negatives divided by the total.

The precision score of the macro average is 0.95 which tells how the True positive plays the role with total predictive positive. This shows the percentage of relevant classification of 7 forest cover-type results. The recall score is an important one to choose the best model since False negatives are associated with it. The recall is calculated as True positives by Total actual positives. The model gives an average recall score of 0.92 which tells the percentage of correctly classified cover-types in the forest area.

The F1 score is more useful than accuracy in the case, where you have an uneven amount of cover-type distribution of data. It is calculated by the weighted average of precision and recall. The F1 score of the KNN model is 0.94. The support shows the number of occurrences of 7 forest cover-types in the unknown test dataset.

VII. CONCLUSION

In this experimental study of predicting forest cover type, we have used many trial and error methods with the appropriate machine learning algorithms using the 10-Fold validation technique where the modeling iterations are done on the whole dataset which makes them to learn better features. This came up with a single best fit KNN algorithm that outperformed the previous studies on the same business problem of forest cover-type prediction. We have also used feature engineering to come up with the necessary 44 important features. This result has led us to a model without any bias or over-fitting, under-fitting problems.

The future works of the forest cover type research can be extended with more numbers of the dataset collected which are equally distributed on all the 7 forest cover-types and features can be reduced which minimizes the misclassification of the model. The work can also be extended to a neural network to understand the performance of the model with the respective feature types.

The colab notebooks and papers can be accessed at https://github.com/BarathRaja/CA683_assignment

VIII. 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).

[5] Bache, K. Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2013).

[6] D.J. Newman A. Asuncion. UCI machine learning repository.https://archive.ics.uci.edu/ml/datasets/Covertype, 2007.

[7] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[8] P. Gonzalez-Aranda, E.Menasalvas, S.Millan, F. Segovia " Towards a Methodology for Data Mining Project Development: The Importance of Abstraction"

[9] "The CRISP-DM Model: The New Blueprint for DataMining", Colin Shearer, JOURNAL of Data Warehousing, Volume 5, Number 4, p. 13-22, 2000.