greatlearning
Learning for Life

Project Summary

Batch details	
Team members	1.) Batchu Heam chandu 2.) Yash Vardhan Khetawat 3.) Mamidi Reshma 4.) V.Sriram 5.) Vishnu Vamsi 6.) S Esha Yadav
Domain of Project	Machine Learning
Proposed project title	Forest Type Prediction
Group Number	5
Team Leader	Batchu Heam chandu
Mentor Name	ANIMESH TIWARI

Date: 21-02-2024

Batchu Heam chandu

ANIMESH TIWARI

Signature of the Mentor Leader

Signature of the Team

Table of Contents

Project Details

OVERVIEW

The development of ecosystem management strategies is the responsibility of natural resource managers, who need basic descriptive data, such as forest land inventory data, to guide their decision-making. For holdings or neighboring lands that are outside of their immediate jurisdiction, managers typically do not have access to this kind of information. The application of predictive models is one way to get this data.Based on the feature values inserted in the trained model, these models can accurately predict the type of cover of any specific area.

The Roosevelt National Forest situated in Colorado's northern region, in the predictive model, there were four wilderness areas included in the study research. The dependent variables were the seven main types of forest cover, and the independent variables were the twelve cartographic measures. Several subsets of these variables were looked at to decide which predictive model was the best overall.

This work aims at potent prediction of forest cover using various classification algorithms, such as random forests, K-Nearest Neighbor and other machine learning models have been tested to develop robust and accurate predictive models for classifying the type of forest cover. Using common ecological and environmental data, the forest department could easily point out the related forest area cover and its type to take any action. This was made possible with the improved accuracy of prediction using Random Forest model and a user-friendly interface to reach out the model.

TOPIC SURVEY:

In the paper, machine learning techniques which are used for the forest cover prediction of specific or various types using remote sensing and cartographic variables. The machine learning techniques used are regression, decision tree, GBM and random forest classifier. The objective is to compare these techniques and identify the one that provides the best prediction accuracy. They

want to develop an automated system for tree species classification that is applicable in real-life scenarios. Different machine learning algorithms were tested, and the highest classification performance was achieved using the random forest algorithm, with an overall accuracy of 86%. These papers also explored the use of remote sensing tools to map forest type and conditions. The data used in this paper are Aerial photographic data as well as Landsat Enhanced Thematic Mapper (ETM+) data. The ECSU campus would then be mapped in terms of its land cover using remote sensing datasets that had been verified in the field.

In their study, researchers evaluated the efficacy of discriminant analysis and artificial neural networks for predicting various types of forest cover. They found that neural networks were generally more accurate than discriminant analysis for predicting forest cover. While there are other papers comparing how fast the algorithm predicts the forest cover with better accuracy. One of the best methods for prediction involves using an immune and genetic approach. This biological approach gave by far the most accurate predictions, although it is not always possible to get hands on biological data. There are infinite possibilities for prediction algorithms. By using cartographic variables, the model can keep learning and predicting with the ever-changing terrain and other geographical features.

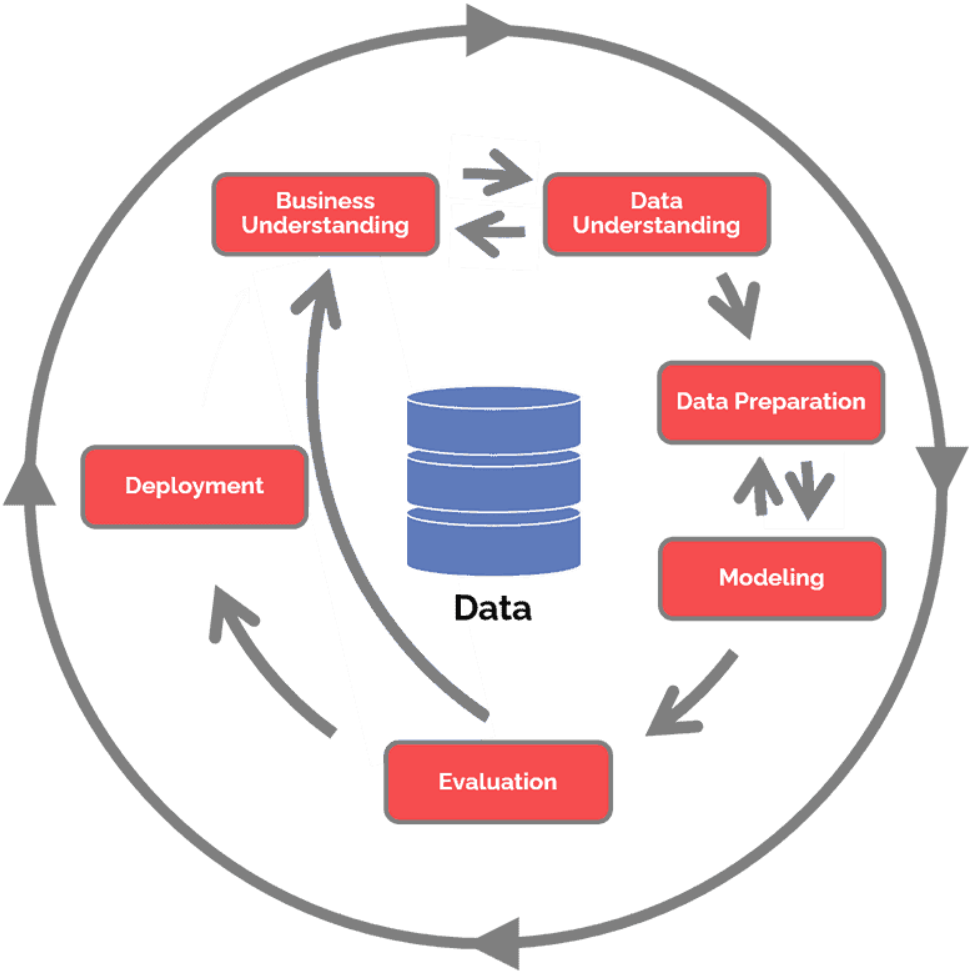
Business problem goals:
Today, one of lot global problems is forest deforestation and monitoring. The article describes the system creation for forests control and monitoring. This work aims are to develop a model for the forest cover type determination based on environmental characteristics and machine learning as the currently developing project part "Monitoring the trees condition using drones". The project aims are to simplify and partially automate the control and monitoring of trees using drones and machine learning to improve the forest situation. The task is to create a model for predicting what tree types grow in the area based on environmental characteristics. Therefore, the main system will be able to compare the existing values/characteristics with the predicted ones (which tree should normally there) to find discrepancies. Eventually, the main system will be able to use this information to report and inform relevant staff and authorities. This work is based on a data set for learning system that includes observations of trees from four areas of Roosevelt National Forest in Colorado. All observations are cartographic variables (without remote sensing) from \(30\times 30\)-m forest areas. In total, there are more than half a million measurements. The work aim is the development of a forest cover types classification model depending on the environment and its characteristics.

CRITICAL ASSESSMENT OF TOPIC SURVEY:

For the increase in efficiency of the model for predicting the outputs accurately, we successfully decreased the time taken, that is Time Complexity of the model, by applying various dimensionality reduction methods. The accuracies and performances of the employed algorithms

were compared, and it was discovered that Random Forest provides better prediction with an accuracy rate of 78% and better performance as compared to other algorithms. By making use of confusion matrix, it became easier to understand and evaluate the model by recognizing the false and correct predictions. On further analysis, it was observed that the major misclassified predictions were due to the similar attributes, and some of the attributes which had differences were incredibly small. Using all these modifications and features, the predictions of the forest cover for Forest department and such could be used for determining the type of forest on the affected area or the area of interest. The prediction data can be fetched using a user interface by providing the required attributes which are easily available.

METHODOLOGY:



A standard method exists for building a model, but various modifications may be necessary to meet the specific requirements of the desired model. The model-building process involves multiple steps, including collecting resources, data cleaning, pre-processing, training, testing, and

deployment. During the resource collection stage, relevant data was collected and collated. Unnecessary data gets cleaned. Pre-processing involves transforming data into a suitable format. The model is then developed using the training data, evaluating the model's accuracy while testing data. The detailed process has been explained below.

A. Dataset Exploration:
   The dataset chosen for the model building was Forest Cover Type. This dataset is a collection of cartographic information about forest cover types for the purpose of predicting forest cover types based on various geographical and environmental factors. It includes information about 7different cover types in Northern Colorado's Roosevelt National Forest. The dataset consists of 15,210 observations with 56 different attributes. The dataset is often used for classification and data mining tasks and has been widely studied in machine learning research.
B. Data Pre-Processing:
   Data preprocessing, which is a component of data preparation, refers to any type of processing carried out on raw data to prepare it for another data processing procedure. It has traditionally been a significant opening phase in the process of data mining. For example, by dropping the unnecessary feature or by combining and making a new single feature helps in achieving a better time complexity One of the pre-processing methods, Dimensionality Reduction has been used in the model, as there were too many individual binary types for the same data and features that can be reduced.
C. Model Selection and Evaluation:
   Seven algorithms were compared using both raw and processed data to evaluate their accuracy and efficiency. The results showed that Random Forest algorithm performed the best with both raw and processed data and was selected for further evaluation and prediction. The technical details of the Random Forest algorithm have been explained in the following section:
   1) Random Forest:
      Random Forest is a widely used machine learning algorithm developed by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. In this article, we will understand how random forest algorithm works, how it differs from other algorithms and how to use it.
      A Random Forest is like a group decision-making team in machine learning. It combines the opinions of many "trees" (individual models) to make better predictions, creating a more robust and accurate overall model.
      Random Forest Algorithm widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems

effectively. The algorithm's strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.

2) Working of Random Forest Algorithm:

Before understanding the working of the random forest algorithm in machine learning, we must investigate the ensemble learning technique. Ensemble simplymeans combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

Bagging

It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

Boosting

It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

random forest | methods of ensemble

As mentioned earlier, Random Forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

Bagging

Bagging, also known as Bootstrap Aggregation, serves as the ensemble technique in the Random Forest algorithm. Here are the steps involved in Bagging:

Selection of Subset: Bagging starts by choosing a random sample, or subset, from the entire dataset.

Bootstrap Sampling: Each model is then created from these samples, called Bootstrap Samples, which are taken from the original data with replacement. This process is known as row sampling.

Bootstrapping: The step of row sampling with replacement is referred to as bootstrapping.

Independent Model Training: Each model is trained independently on its corresponding Bootstrap Sample. This training process generates results for each model.

Majority Voting: The final output is determined by combining the results of all models through majority voting. The most commonly predicted outcome among the models is selected.

Aggregation: This step, which involves combining all the results and generating the final output based on majority voting, is known as aggregation.

Steps Involved in Random Forest Algorithm:

Step 1: In the Random Forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

important Features of Random Forest:

Diversity: Not all attributes/variables/features are considered while making an individual tree; each tree is different.

Immune to the curse of dimensionality: Since each tree does not consider all the features, the feature space is reduced.

Parallelization: Each tree is created independently out of different data and attributes. This means we can fully use the CPU to build random forests.

Train-Test split: In a random forest, we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

Stability: Stability arises because the result is based on majority voting/ averaging.

The Random Forest model is trained on the environmental features to predict forest cover types. The training process involves constructing multiple decision trees and combining their predictions to enhance accuracy and generalization. The evaluation metrics, including precision, recall, F1-

score, confusion matrix, and Cohen's Kappa Score, are employed to comprehensively assess the model's performance on the test set. Random Forests are chosen for their ability to handle complex relationships and provide robust predictions, making them well-suited for tasks like predicting forest cover types based on diverse environmental attributes.

D. Model Evaluation:

	precision	recall	f1-score	support
1	0.76	0.73	0.74	506
2	0.76	0.67	0.71	523
3	0.81	0.77	0.79	514
4	0.91	0.96	0.94	524
5	0.89	0.93	0.91	512
6	0.80	0.84	0.82	528
7	0.92	0.97	0.94	522
accuracy			0.84	3629
macro avg	0.83	0.84	0.84	3629
weighted avg	0.83	0.84	0.84	3629

The provided report describes the performance of a Random Forest classification model on a dataset containing seven forest cover types. The target variable consists of categories labeled 1 to 7. The evaluation metrics include precision, recall, and F1-score for each class, as well as macro and weighted averages. The precision values range from 0.76 to 0.92, indicating the accuracy of positive predictions. Recall values vary from 0.67 to 0.97, reflecting the model's ability to capture relevant instances. F1-scores, balancing precision and recall, range from 0.71 to 0.94. The support values detail the actual occurrences of each class in the dataset. The overall accuracy of the model is reported at 84%, implying correct classification for 84% of instances. Macro average metrics, considering equal weight for all classes, are 0.83 for precision, 0.84 for recall, and 0.84 for F1-score. Weighted averages, accounting for class proportions, yield the same values. The Random Forest model demonstrates robust performance across the seven forest cover types, with balanced precision, recall, and F1-score. The 84% accuracy suggests reliable classification on the given dataset.

REFERENCES

1. Blackard, J. A., & Dean, D. J. (1999). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Computers and Electronics in Agriculture,24(3).
2. Badulescu, L. (2017). Data mining classification experiments with decision trees over the forest
covertype database. In 21st International Conference on System Theory Control and Computing (ICSTCC),236.
3. Xue, J.-H., & Hall, P. (2015). Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(5), 1109–1112.
4. Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. The Journal of Machine Learning Research,17(1), 841–881.
5. Blackard, J. A. (2000). 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

Notes For Project Team

Original owner of data	US Forest Service (USFS)
Data set information	The data set contain 15,120 rows and 56 columns.
Any past relevant articles using the dataset	Forest Cover Type Prediction by Hyped Splatoon
Reference	https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset/data?select=covtype.csv
Link to web page	https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset/data?select=covtype.csv

****