**Project Summary**

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT Offline Sep-21 |
| Team members | 1. Azhar Madani 2. B Arun Singh Thakur 3. Sri Datta 4. A. Satish |
| Domain of Project | Forest |
| Proposed project title | Forest Cover Type Prediction |
| Group Number | G1 |
| Team Leader | Azhar Madani |
| Mentor Name | Mr. Romil Gupta |

Date: 24th Dec 2021

Signature of the Mentor Signature of the Team Leader

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 4 |
| 4 | Critical assessment of topic survey | 5 |
| 5 | Methodology to be followed | 6 |
| 6 | Timeline chart(Weekly Plan) | 7 |
| 7 | References | 8 |

**Project Details**

## The Cover type dataset was first used in a machine learning context by Blackard and Dean, as part of Blackard’s doctoral thesis and then as part of an academic article. Both of the original works use the data set as the basis of a supervised learning task. The data set’s instances are drawn from US Forest Service (USFS) Region 2 Resource

Information System data and the classifier must predict the type of forest cover present in a 30 x 30 metre cell given the observed geographic information system (GIS) variables.

Since its introduction, it has become a standard benchmark data set in the literature and has been cited by hundreds of papers. The data set is available as raw data from the UCI Machine Learning Repository and in normalized form from the website of Massive Online Analysis (MOA), an open-source framework for data stream mining.

# OVERVIEW

# Business problem statement (GOALS)

1. **Business Problem Understanding:**

We have been given a total of 54 attributes, these attributes contain Binary and Quantitative attributes, and we need to predict which **Forest Cover-Type**is it from the given features.

1. **Business Objective:**

Understanding forest composition is a valuable aspect of managing the health and vitality of our wilderness areas. Classifying cover type can help further research regarding forest fire susceptibility and de/reforestation concerns. Forest cover type data is often collected by hand or computed using remote sensing techniques, e.g. satellite imagery. Such processes are both time and resource intensive. In this project, we aim to predict forest cover type using cartographic data and a variety of classification algorithms.

1. **Approach:**
2. Understand, Clean and Format Data
3. Exploratory Data Analysis
4. Feature Engineering & Selection
5. Compare Several Machine Learning Models
6. Perform Hyperparameter Tuning on the Best Model
7. Interpret Model Results
8. Evaluate the Best Model with Test Data (replying the initiating question)
9. Summary & Conclusions
10. **Conclusions:**

To predict seven different cover types in four different wilderness areas of the [Roosevelt National Forest of Northern Colorado](https://en.wikipedia.org/wiki/Roosevelt_National_Forest) with the best [accuracy](https://miro.medium.com/max/1064/1*5XuZ_86Rfce3qyLt7XMlhw.png).

# TOPIC SURVEY IN BRIEF (200-250 words)

**1.Problem understanding**

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

Natural resource managers responsible for developing ecosystem management strategies require basic descriptive information including inventory data for forested lands to support their decision-making processes. However, managers generally do not have this type of data for inholdings or neighbouring lands that are outside their immediate jurisdiction. One method of obtaining this information is through the use of predictive models.

The study area included four wilderness areas found in the Roosevelt National Forest of northern Colorado. A total of twelve cartographic measures were utilized as independent variables in the predictive models, while seven major forest cover types were used as dependent variables. Several subsets of these variables were examined to determine the best overall predictive model.

**2. Current solution to the problem**

In this project, we aim to predict the forest cover type using many variables that are influencing the outcome

**3. Proposed solution to the problem**

The goal of our classification model is to predict the forest cover type using many variables that are influencing the outcome

# CRITICAL ASSESSMENT OF TOPIC SURVEY

1. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business

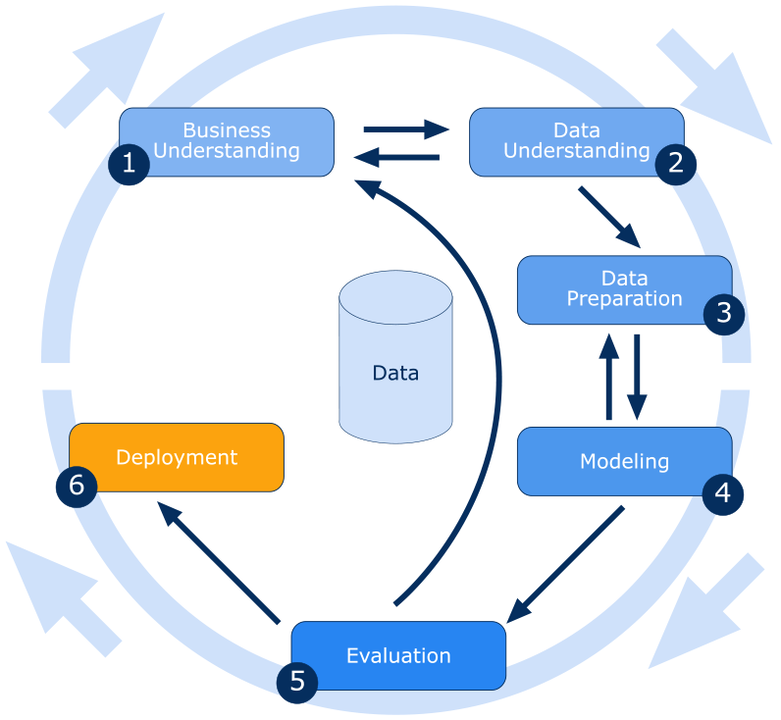
Predicting the different forest cover types

* 1**:** Spruce/Fir
* 2: Lodgepole Pine
* 3: Ponderosa Pine
* 4: Cottonwood/Willow
* 5: Aspen
* 6: Douglas-fir
* 7: Krummholz

1. What key gaps are you trying to solve?

By applying the various Supervised learning models such as classification our model will predict the above scenario. In this project, we aim to predict the forest cover type using many variables that are influencing the outcome. Various feature engineering techniques performed on the datasets to improve over the primary data-set

# METHODOLOGY TO BE FOLLOWED (Explain each step from 1-5)



**Data Understanding**: EDA is the first step in this workflow where the decision-making process is initiated for the feature selection. Some valuable insights can be obtained by looking at the distribution of the target, relationship of the features to the target and link between the features.

**Data Preparation**: The missing values should be handled by using appropriate methods based on the type of attribute. Existence of outliers in the dataset should be identified and handled. Attributes relevant to the problem should be identified.

**Modelling**:

* EDA
* Classification

# Data Evaluation: The results of the modelling should be analyzed and check whether it is giving appropriate results.

**Timeline Chart (Weekly plan):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. No.** | **Task** | **Week** | **Date** | **Status** |
| 1. | Project Synopsis | Week 1 | 24-12-2021 | Completed |
| 2. | Data Handling:   1. Data Sampling 2. Data Cleaning 3. Missing, Null   Value Treatment | Week 2 |  | To Be Started |
| 3. | EDA:   1. Dummy Encoding 2. Feature Engineering 3. Data visualization | Week 3 |  | To Be Started |
| 4. | Model Building | Week 4 |  | To Be Started |
| 5. | Model Tuning | Week 5 |  | To Be Started |
| 6. | Model Validation and Deployment | Week 6 |  | To Be Started |

# REFERENCES:

[Machine learning Workflow](https://towardsdatascience.com/predicting-forest-cover-types-with-the-machine-learning-workflow-1f6f049bf4df)

Blackard, Jock A. and Denis J. Dean. 2000. "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables." Computers and Electronics in Agriculture

[Forest Cover Type Prediction | Kaggle](https://www.kaggle.com/c/forest-cover-type-prediction)

**Notes For Project Team**

*Sample Reference for Datasets (to be filled by team and mentor)*

|  |  |
| --- | --- |
| Original owner of data | UCI Machine Learning Repository |
| Data set information | Multiple files that need to be clubbed together. Pre-Cleaning (Rows & Columns)  Rows: 5,81,012  Columns: 54 |
| Any past relevant articles using the dataset | Yes |
| Reference | Yes |
| Link to web page | [Dataset](https://www.kaggle.com/uciml/forest-cover-type-dataset) |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*