

## **Deliverable – 1**

### **Predicting tree types found in the Roosevelt National Forest in Colorado**

**Team members:**

**Sai Anish Chowdary Bezawada (801312402)**

**Pranav Sundaresan Babu (801254287)**

**Sushma Narne (801269072)**

**Sai Rachana Reddy Vanipenta (801266847)**

**Abhishekji Dumala (801307903)**

**Communication plan to include project artifact repository:**

A Word file will be uploaded to Canvas as a way to complete Deliverable 1. Code, screenshots of estimated accuracy and column impact will all be stored in a repository that is established once all other deliverables have been finished. Everyone will have access to the repository as it is made public.

Project repository can be accessed on Github using the following Link given below:

[https://github.com/anish1999-13/Big\\_data\\_Project](https://github.com/anish1999-13/Big_data_Project)

**Data Set Selection :**

**We have selected our dataset from Kaggle. Below are the link to selected data set:**

<https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset>

**Business Problem:**

Understanding of the various forest types and their features is a business opportunity or challenge that must be kept in mind in order to make decisions about land management and conservation initiations. We will be creating a model that would give an overview of the forest's cover type depending on various characteristics such as, elevation type, slope type and the most important is the type of soil. We are interested in determining the features which are the most important for the overview of the type of forest cover.

Keeping all these factors in mind this dataset's business problem is to create a Machine Learning Model which predicts the forest's cover type with the various features/attributes and to determine the most important features that would help us achieve the goal. This model can assist us in making well-informed decisions regarding land management and conservation activities, such as identifying regions that may need various management

techniques depending on their kind of cover or regions that are especially vulnerable to specific disturbances or threats.

## **Input Data Set:**

We have selected our dataset from Kaggle. The Roosevelt National Forest in northern Colorado's Roosevelt National Forest is included in the Forest Cover Type dataset from the UC Irvine Machine Learning Repository. The target variable is a categorical variable that represents the cover\_type of the forest, which has seven potential classes, in the dataset, which also contains continuous and categorical variables.

Here's an explanation of each attribute in the dataset:

Elevation: Elevation in meters

Aspect: Aspect in degrees azimuth

Slope: Slope in degrees

Horizontal\_Distance\_To\_Hydrology: Horizontal distance to nearest surface water features in meters

Vertical\_Distance\_To\_Hydrology: Vertical distance to nearest surface water features in meters

Horizontal\_Distance\_To\_Roadways: Horizontal distance to nearest roadway in meters

Hillshade\_9am: Hillshade index at 9am, summer solstice (0 to 255 index)

Hillshade\_Noon: Hillshade index at noon, summer solstice (0 to 255 index)

Hillshade\_3pm: Hillshade index at 3pm, summer solstice (0 to 255 index)

Horizontal\_Distance\_To\_Fire\_Points: Horizontal distance to nearest wildfire ignition points, in meters

Wilderness\_Area (4 binary columns): Wilderness area designation

Soil\_Type (40 binary columns): Soil Type designation

The "Wilderness\_Area" and "Soil\_Type" attributes are binary columns indicating whether or not the forest observation falls into a particular wilderness area or soil type, respectively.

## **Research Objectives:**

- Evaluating the performance of different machine learning algorithms for the prediction the forest cover type, and identifying the algorithm which performs the best for this model.
- Identify any patterns or trends in the forest cover types over time, and understand how changes in land management and environmental factors may be contributing to these trends.

- Investigate the potential impacts of climate change on the forest cover types and associated vegetation, and understand how the forest cover types may shift over time in response to changing environmental conditions.
- Identify any areas of high conservation value based on the forest cover type and its associated features, and prioritize conservation efforts in these areas.