Forest Cover Type Prediction
(Multiclass Classification)

Saatvik Sinha **MT2025722** & Affan Shaikh **MT2025016**

December 10, 2025

Abstract

The goal of this project is to predict the type of forest cover present in different regions of wilderness using machine learning techniques. The dataset used in this study contains measurements of environmental variables such as elevation, slope, distances to roadways and water bodies, hillshade values, and soil type indicators. These features allow machine learning models to learn how geographical and ecological conditions influence the type of trees that grow in a particular location. The prediction task is important because accurate forest cover classification can help in environmental conservation, wildfire risk assessment, land usage planning, and biodiversity monitoring.

Several machine learning models were trained and compared to determine which approach works best for this dataset. The project includes linear models, distance-based models, tree-based models, boosting models, support vector machines, and deep learning using neural networks. Each model was evaluated using metrics such as accuracy, precision, recall, and F1-score to understand its strengths and weaknesses. To improve fairness in comparison, the same preprocessing steps, the same train–test split, and the same evaluation framework were applied to all models.

The results show that the dataset is highly non-linear and contains complex interactions among environmental variables and soil types. Because of this, simple linear models like Logistic Regression and Linear SVM were not able to capture underlying patterns and performed poorly. Neural Networks showed strong performance after tuning, which confirms the presence of non-linear patterns in the data. However, the best performance was achieved by the Random Forest model, which reached an accuracy of 95.52%. This model works well when features interact in non-linear ways, and it benefits from combining many decision trees to reduce overfitting and improve generalization.

Overall, this study demonstrates that machine learning can be used effectively to automate forest cover prediction and that tree-based ensemble models are the most suitable for structured tabular environmental data.

Table of Contents

1 Dataset Description

The dataset used in this project is the UCI Forest CoverType dataset, a widely used benchmark dataset for forest cover classification. It contains information about geographic and soil characteristics of different plots of land from four wilderness areas located in the Roosevelt National Forest in northern Colorado, USA. The main objective of the dataset is to predict the **Cover_Type**, meaning the dominant kind of tree species expected to grow in that area.

The dataset consists of **581,012 individual land samples**, each represented by 54 attributes that describe environmental and ecological conditions. These attributes can be grouped into two main categories:

1.1 Feature Groups

1. Continuous numerical features (10 columns): These include geographical measurements that vary over a continuous scale.

- **Elevation (in meters)**: height above sea level.
- **Aspect**: direction the slope faces (0°–360°).
- **Slope**: steepness of the terrain.
- **Horizontal & Vertical Distance to Hydrology**: distance to water sources.
- **Horizontal Distance to Roadways / Fire Points**: accessibility and fire risk.
- **Hillshade at 9am, Noon, 3pm**: sunlight exposure on the terrain.

2. Categorical one-hot encoded features (44 columns): Instead of being continuous measurements, these features represent the presence or absence of certain characteristics.

- **Wilderness Area (4 columns)**: Which protected area the sample belongs to.
- **Soil Type (40 columns)**: Soil composition and geology information.

1.2 Target Variable

The target variable in this dataset is the **Cover_Type**. There are 7 cover type classes:

Class ID	Tree Species (Cover Type)
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz

Table 1: Target Variable Classes

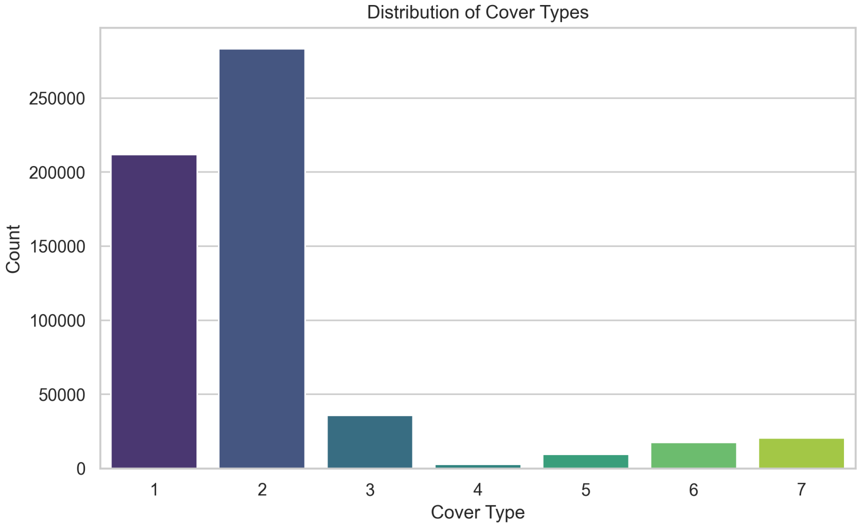

Figure 1: Distribution of Forest Cover Types

2 Methodology

The goal of the methodology is to describe the exact steps followed to develop, train, evaluate, and compare different machine learning models. The process is divided into five major stages.

2.1 Data Preprocessing

The dataset was first loaded and inspected. The following steps were performed:

1. Checked dataset structure and for missing values (none were found).

2. Created an engineered feature — **Hydro_Euclidean Distance**, which combines horizontal and vertical distance to hydrology.
3. Performed train–test split using **80%** of data for training and **20%** for testing, using stratification.
4. Scaled features using `StandardScaler` for distance-based models (KNN, SVM, Neural Networks). Tree-based models used non-scaled data.

2.2 Exploratory Data Analysis (EDA)

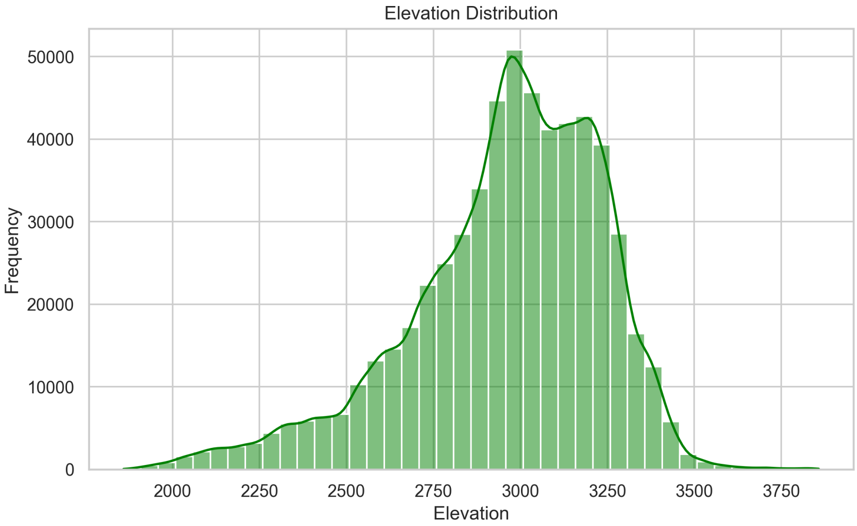

Figure 2: Distribution of Elevation Across Samples

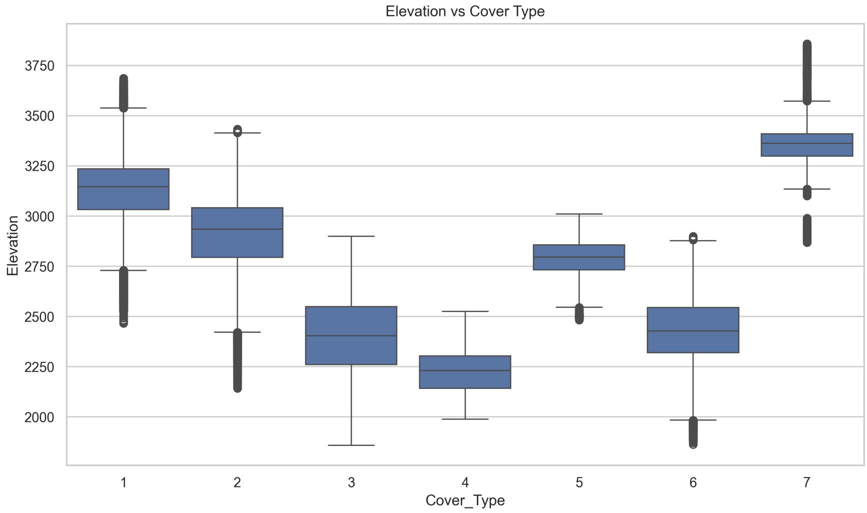

Figure 3: Elevation vs Cover Type

EDA was performed to understand feature distributions and relationships.

- **Histograms** visualized the distribution of numerical features such as elevation.
- **Boxplots** (Elevation vs Cover Type) demonstrated how land characteristics influence specific tree species.
- **Correlation Heatmaps** identified relationships among numerical features.

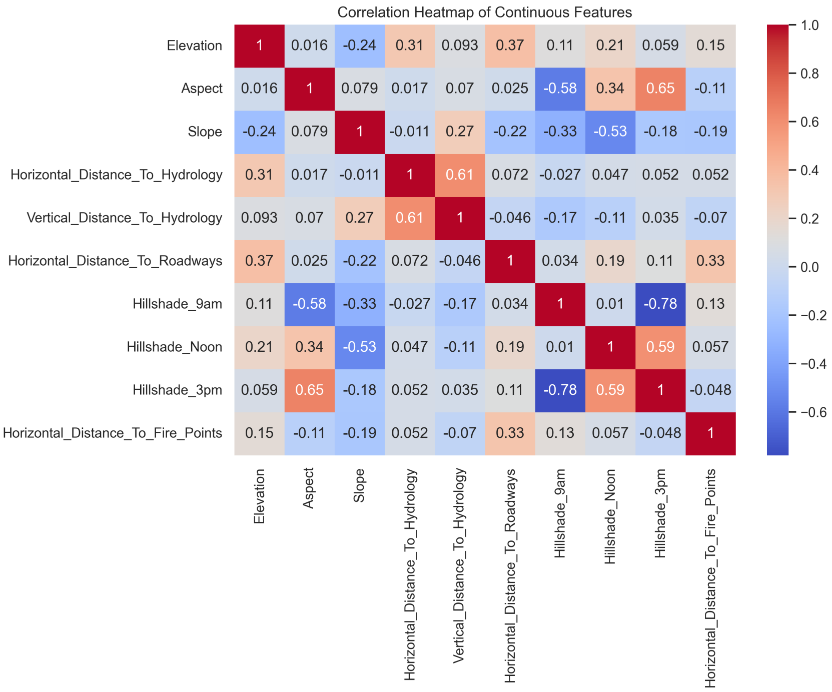

Figure 4: Correlation heatmap of continuous features

2.3 Model Building

A wide range of supervised learning models were developed:

Model Category	Algorithms Used
Linear models	Logistic Regression
Distance-based	K-Nearest Neighbors (KNN)
Tree-based	Decision Tree
Ensemble (Bagging)	Random Forest
Ensemble (Boosting)	HistGradientBoosting
Margin-based	Linear SVM
Deep Learning	Neural Network (MLP)

Table 2: Machine Learning Models Implemented

3 Results

This section presents the performance of all machine learning models trained on the Forest Cover Type dataset.

3.1 Baseline Model Performance

Model	Accuracy
Random Forest (150 trees)	95.47%
Decision Tree	93.84%
Neural Network (MLP)	87.16%
KNN (k=15)	86.48%
HistGradientBoosting	83.08%
Logistic Regression	71.92%
Linear SVM	71.14%

Table 3: Baseline Accuracy Results

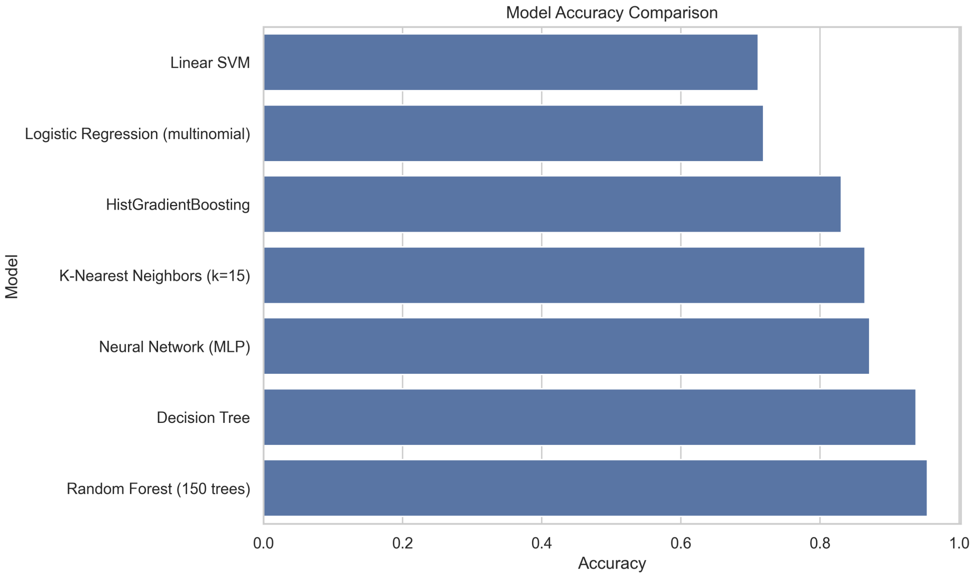

Figure 5: Accuracy Comparison Across All Models

3.2 Performance After Hyperparameter Tuning

Model	Accuracy After Tuning
Random Forest (tuned)	95.52%
Neural Network (tuned MLP)	93.75%
HistGradientBoosting (tuned)	87.13%

Table 4: Tuned Accuracy Results

3.3 Final Ranking of Models

The Random Forest model (tuned) emerged as the winner.

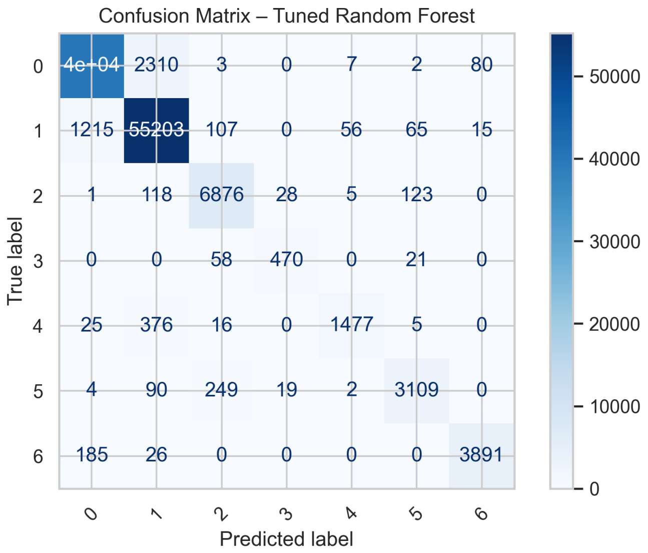

Figure 6: Confusion Matrix of the Tuned Random Forest Model

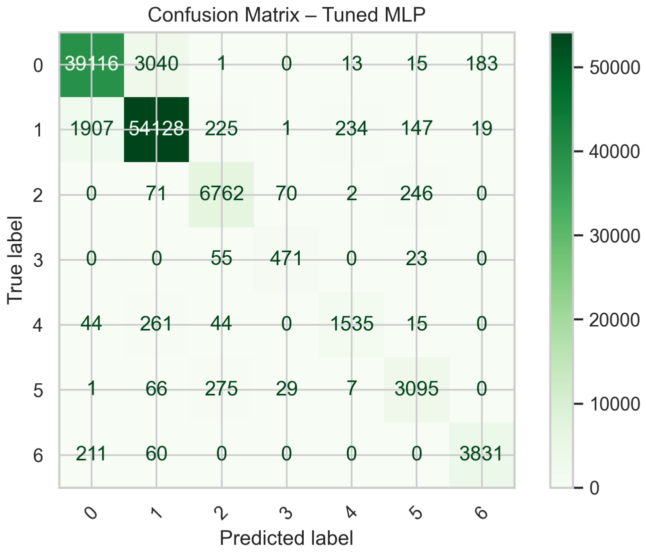

Figure 7: Confusion Matrix of the Tuned Neural Network (MLP)

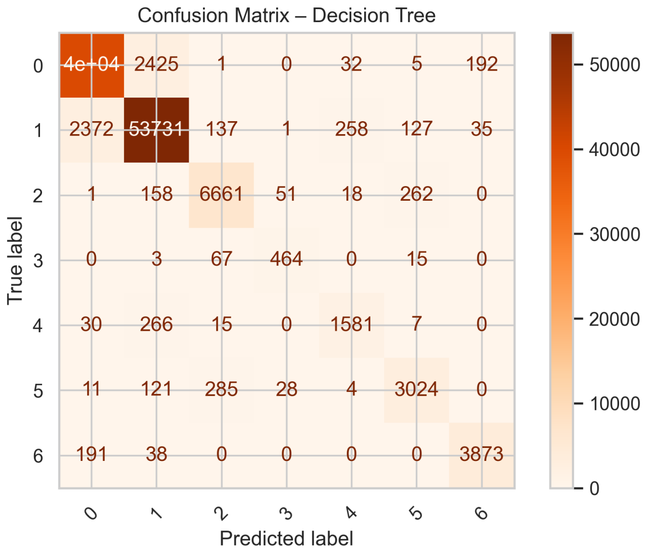

Figure 8: Confusion Matrix of the Decision Tree Classifier

4 Comparative Analysis

4.1 Overall Model Comparison

The results clearly show that **tree-based learning methods** performed the best on the Forest Cover dataset. The Random Forest (tuned) achieved the highest accuracy of 95.52%. Linear models such as Logistic Regression and Linear SVM performed significantly poorly due to the highly non-linear nature of the dataset.

4.2 Why Top-Performing Models Worked Better

Random Forest: Able to capture complex non-linear interactions and feature dependencies.

Decision Tree: Learns interpretable decision boundaries but prone to overfitting.

Neural Network (MLP): Learns deep non-linear relationships but more sensitive to hyperparameters.

4.3 Why Low-Performing Models Failed

Logistic Regression & Linear SVM: Cannot handle non-linear boundaries in multi-dimensional environmental data.

KNN: Struggles with high-dimensional one-hot encoded soil features and is computationally heavy.

5 Conclusion

The project successfully demonstrated that model selection has a major impact on prediction accuracy for forest cover classification. Tree-based ensemble methods, especially the **tuned Random Forest**, achieved the highest accuracy of 95.52%. This proves they are best suited for structured, high-dimensional ecological datasets.

6 Future Work

- Use advanced ensemble methods such as XGBoost, LightGBM, or CatBoost.
- Try stacked or hybrid models for improved generalization.
- Add spatial and climatic information to improve ecological accuracy.

- Handle class imbalance using SMOTE or class weighting.
- Deploy the model via a web or mobile application.