

# Forest Cover Type Classification: A Comprehensive Report

Date: August 24, 2025

## 1. Executive Summary

This report outlines the development and evaluation of machine learning models to predict forest cover types based on cartographic data. The project utilized the well-known Forest Cover Type dataset, comprising 581,012 records and 55 features. After a thorough exploratory data analysis, three ensemble models—Random Forest, XGBoost, and Gradient Boosting—were trained and evaluated. The Random Forest model emerged as the top performer, achieving an impressive accuracy of 95.5%, demonstrating its superior capability for this classification task.

## 2. Data Loading and Preparation

- Data Acquisition: The dataset was provided in a compressed zip file (/content/covertypes.zip). It was first unzipped, and the raw data file (covtype.data.gz) was then loaded into a pandas DataFrame, creating a usable covtype.csv file.
- Initial Inspection: The dataset contains 581,012 entries and 55 columns. All columns were integers, with no missing values. Column 54 was identified as the target variable, Cover\_Type.

## 3. Exploratory Data Analysis (EDA)

The EDA phase was crucial for uncovering patterns, understanding feature distributions, and identifying relationships within the data.

- Target Variable Distribution: Showed class imbalance.
- Numerical Feature Analysis: Histograms, scatter plots, and correlation heatmaps were used to explore features.

## 4. Model Training and Evaluation

The dataset was split into train (80%) and test (20%), with zero-indexed labels. Three models were evaluated.

### 4.1 Random Forest Classifier

Accuracy: 95.5%

Class	Precision	Recall	F1-Score	Support
0	0.97	0.94	0.95	42,557
1	0.95	0.97	0.96	56,500
2	0.94	0.97	0.96	7,121
3	0.92	0.83	0.87	526
4	0.94	0.77	0.85	1,995

5	0.94	0.90	0.92	3,489
6	0.97	0.95	0.96	4,015
Weighted Avg	0.96	0.96	0.95	116,203

**4.2 XGBoost Classifier**

Accuracy: 87.1%

**4.3 Gradient Boosting Classifier**

Accuracy: 77.3%

**5. Conclusion and Recommendation**

Model	Accuracy	Key Observation
Random Forest	95.5%	Best overall performance and robustness.
XGBoost	87.1%	Strong performance, less effective on minority classes.
Gradient Boosting	77.3%	Significantly impacted by class imbalance.

The Random Forest classifier is the recommended model for this task. Its superior accuracy and more balanced performance across all seven cover types make it the most reliable solution.