Predicting Forest Cover Types

Rohitangshu Bose
Roll No: MT2025106
Rohitangshu.Bose@iiitb.ac.in
Course: Machine Learning
International Institute of Information Technology, Bangalore

December 2025

Abstract

This project develops machine learning models to classify forest cover types using environmental, topographical, and terrain-based indicators from the Forest Cover Type dataset. The data contain 581,012 samples and 54 features describing elevation, slope, hydrology distances, hillshade values, soil type, and wilderness area. Comprehensive exploratory data analysis was conducted to study feature distributions, understand ecological structure, examine correlations, and assess significant class imbalance.

Preprocessing steps included label correction, standardization using `StandardScaler`, and SMOTE-based oversampling on the training split to alleviate severe imbalance, particularly for minority cover types. Three models were trained and compared: an Artificial Neural Network (ANN), Logistic Regression, and Linear Support Vector Machine (SGD). The ANN achieved the strongest performance with a test accuracy of 0.9037, demonstrating the effectiveness of nonlinear modeling for complex ecological classification tasks.

Visual analyses—such as histograms, boxplots, scatter plots, correlation matrices, confusion matrices, and model comparison charts—provide insight into both the dataset structure and model behavior. The project highlights the importance of nonlinear modeling, class balancing, and architectural regularization for robust performance. The full implementation is available at:
`https://github.com/Rohitangshu2026/AIT511-Machine-Learning-Course-Project-2`.

Contents

1 Introduction

Accurate prediction of forest cover types is central to ecological modeling, forest management, biodiversity assessment, and wildfire risk estimation. The Forest Cover Type dataset contains detailed terrain and soil information, enabling machine learning models to capture relationships among environmental variables and vegetation patterns. Traditional classification relies largely on expert heuristics, whereas modern learning methods reveal nonlinear interactions and structural patterns in multidimensional ecological data.

This project follows a structured workflow inspired by a previous health-related classification study, extending the same methodological rigor to a large-scale environmental task. The objective is to: (1) conduct detailed exploratory data analysis, (2) implement appropriate preprocessing, (3) train multiple classification models, (4) compare their performance, and (5) support findings with visual reasoning.

2 Methodology

The methodological pipeline includes:

1. **Data loading and inspection** (shape, types, statistical summary).
2. **Exploratory Data Analysis (EDA)**, including class distribution, feature distributions, boxplots, correlation analysis, and scatter visualization.
3. **Preprocessing** involving missing-label removal, feature scaling, and SMOTE balancing.
4. **Model development** using ANN, Logistic Regression, and Linear SVM.
5. **Evaluation** through accuracy scores, F1-metrics, confusion matrices, and comparative visualizations.

3 Dataset Description

The dataset contains 581,012 observations, 54 features, and 7 target classes. Features include:

- **Continuous**: Elevation, Aspect, Slope, Hillshade values, Hydrology distances, Roadway/Fuel distances.
- **Binary**: 4 wilderness areas, 40 soil types.

The target variable `Cover_Type` ranges from 1 to 7. For modeling, labels were shifted to 0–6.

4 Exploratory Data Analysis

4.1 Class Distribution

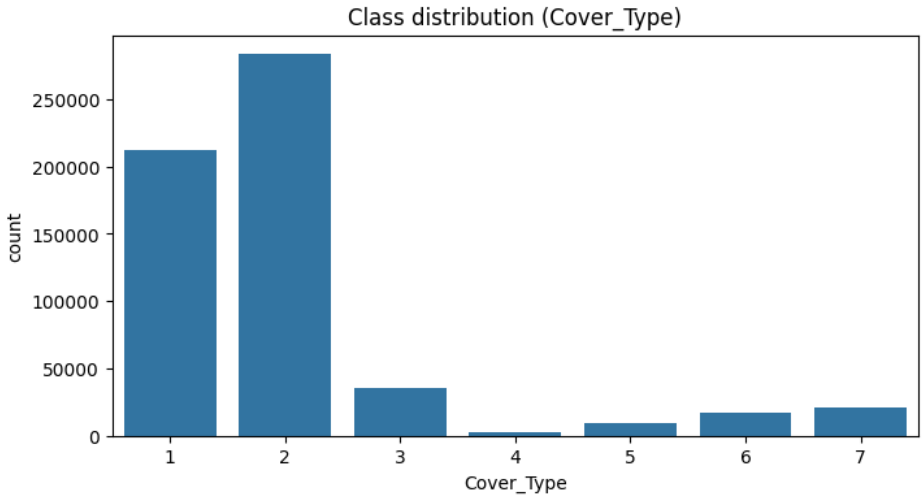

Figure 1: Distribution of the seven forest cover types.

The dataset exhibits extreme imbalance. Cover Types 1 and 2 constitute the majority, while Types 3–7 represent only a small proportion. This imbalance significantly affects linear models and necessitates the use of oversampling for stable training.

4.2 Feature Distributions

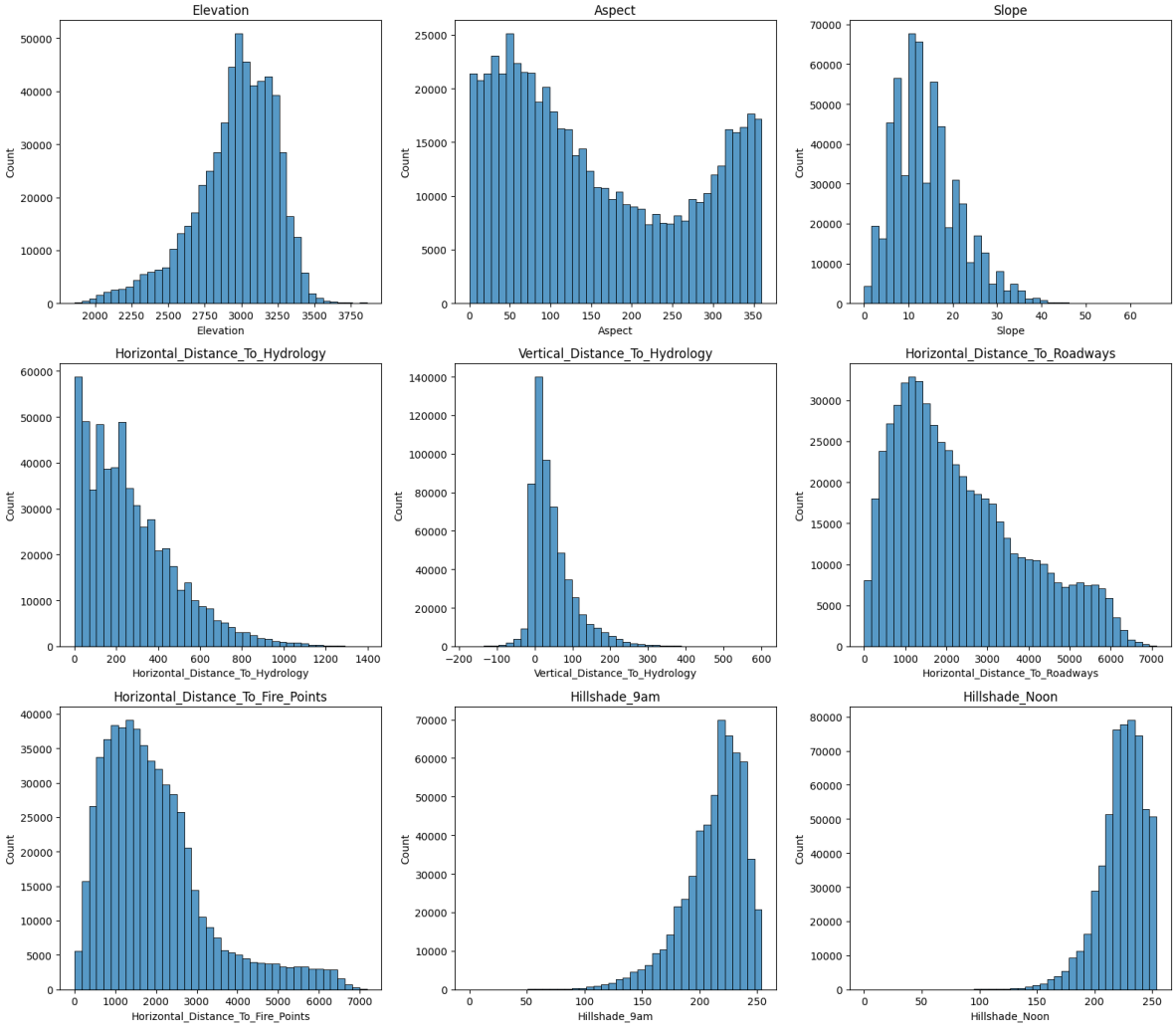

Figure 2: Distributions of major continuous terrain features.

Elevation shows clear multimodal structure, reflecting ecological zones. Hydrology and road distances exhibit long right tails, characteristic of terrain datasets. Aspect displays a U-shaped pattern typical of directional data, while hillshade features are approximately Gaussian. These varied distribution shapes underscore the need for robust scaling and nonlinear modeling.

4.3 Boxplots Across Cover Types

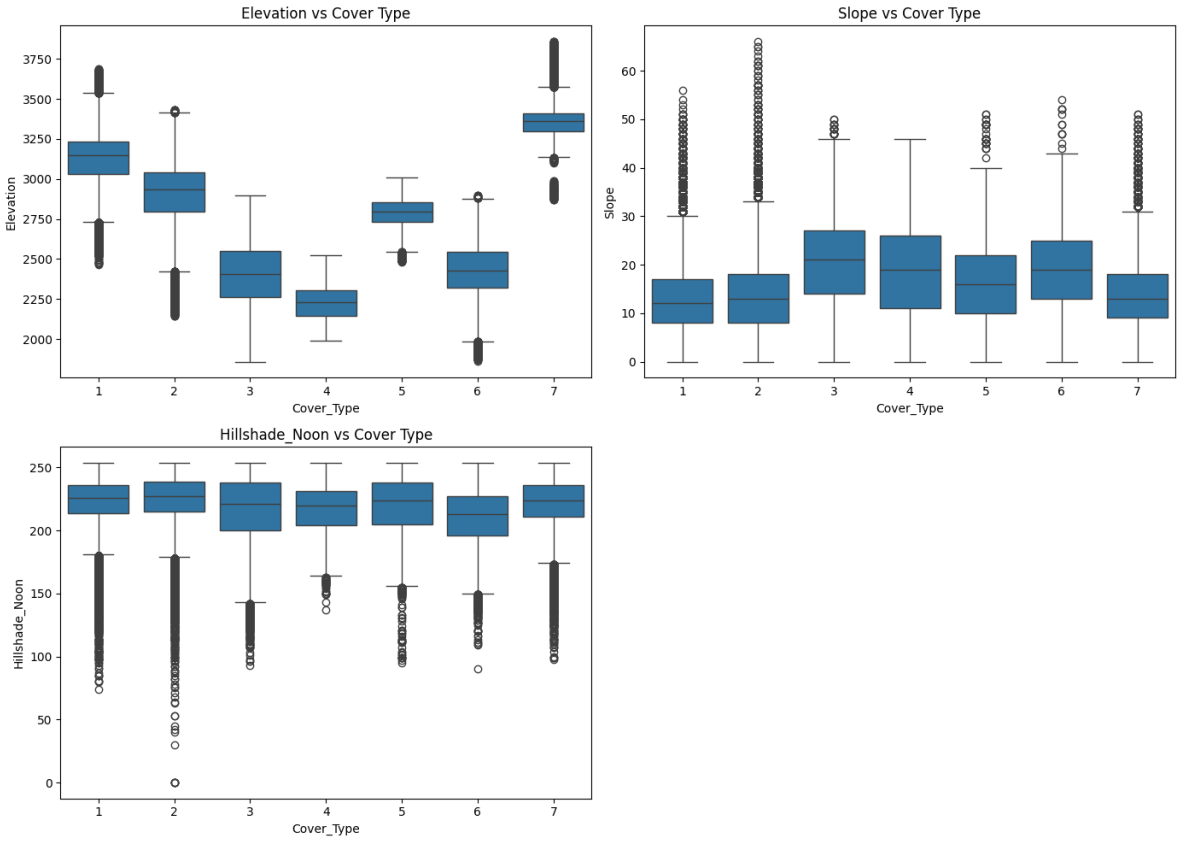

Figure 3: Elevation, Slope, and Hillshade distributions across different cover types.

Elevation serves as a key discriminatory factor. Cover Types 1 and 2 typically occur at higher altitudes, while Types 3 and 4 appear at lower elevations. Slope shows moderate separation across classes, especially for Type 3. Hillshade Noon exhibits substantial overlap, suggesting limited discriminative utility. Overall, elevation emerges as the most influential terrain predictor.

4.4 Correlation Analysis

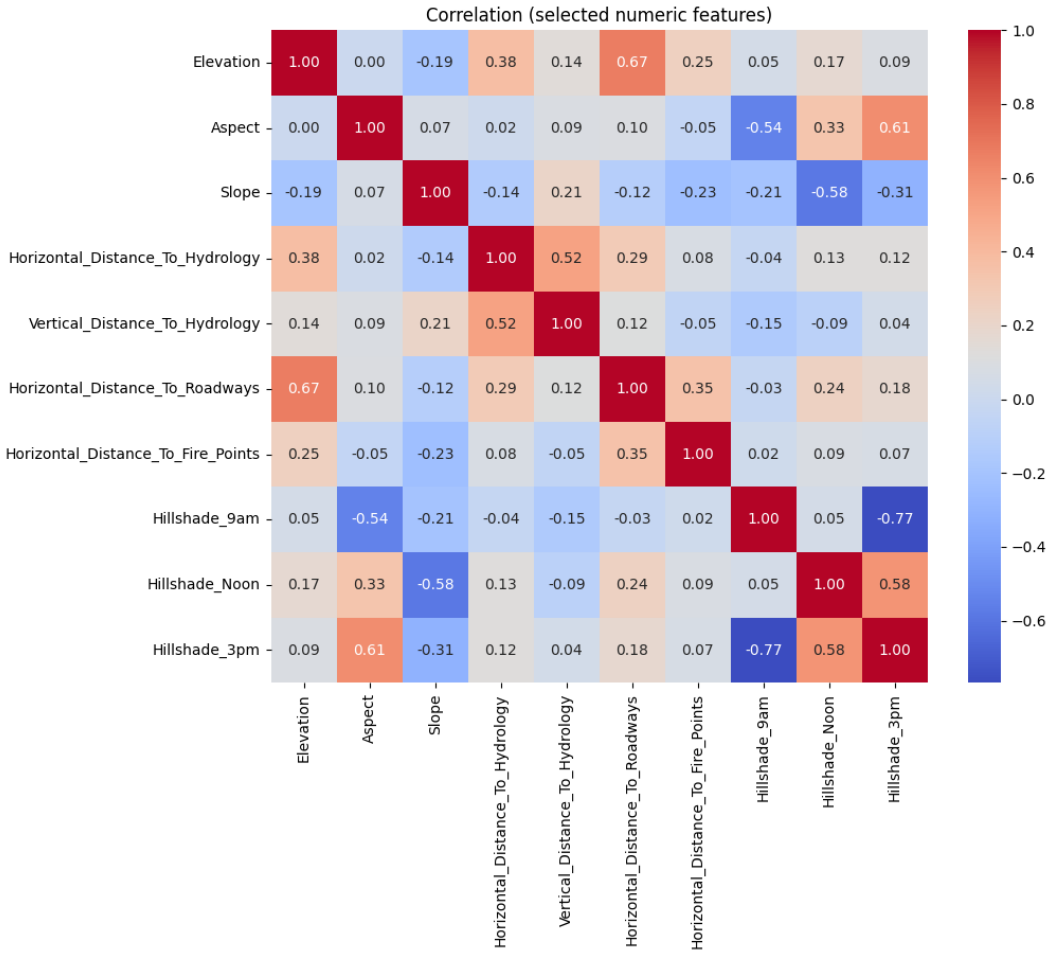

Figure 4: Correlation matrix of selected continuous features.

Feature correlations are generally weak, with the exception of negative correlation between Hillshade at 9am and 3pm (due to solar geometry) and moderate correlation between elevation and road distance. The limited correlation structure indicates a high-dimensional, nonlinear classification landscape.

4.5 Scatter Plot Analysis

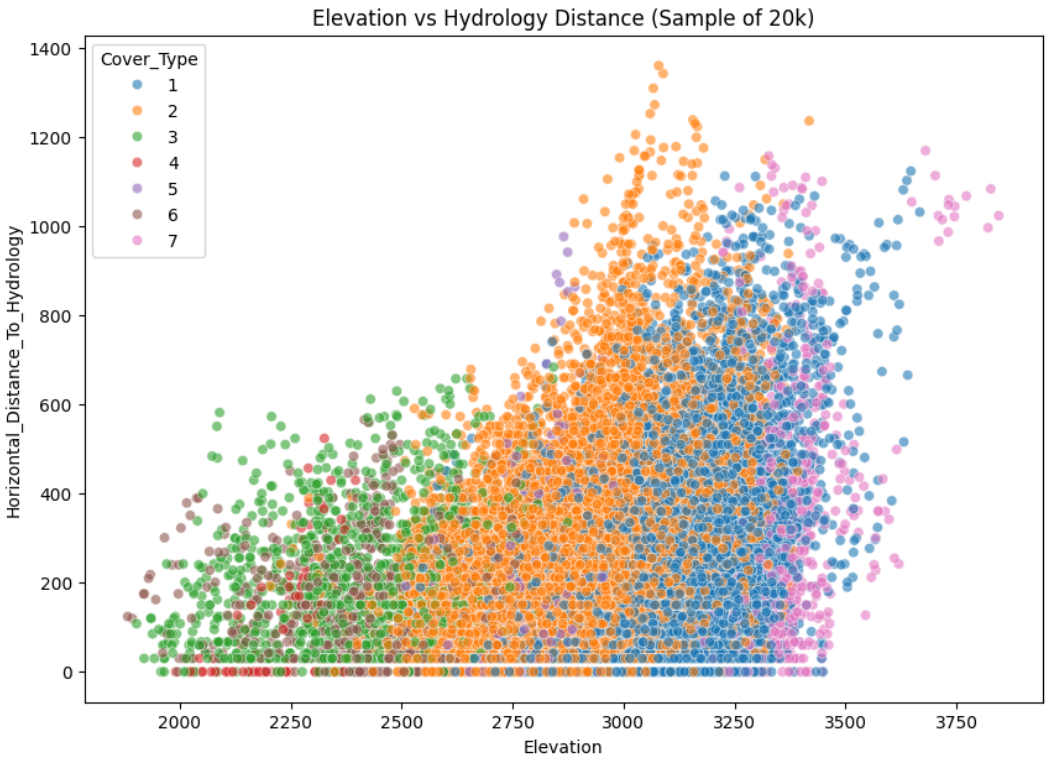

Figure 5: Elevation vs. Hydrology Distance for a 20,000-sample subset, colored by cover type.

Distinct ecological clusters emerge across elevation–hydrology space. Higher cover types tend to occupy elevated regions with greater hydrology distances. Types 3 and 4 cluster closer to lower elevations, whereas Type 7 occupies the highest range. This visualization illustrates nonlinear separability that benefits ANN-based modeling.

5 Data Preprocessing & Feature Engineering

- Missing labels in the target column were removed.
- Numerical features were scaled using `StandardScaler`.
- SMOTE oversampling was used on the training split to correct class imbalance.
- Labels were shifted from $\{1,\ldots,7\}$ to $\{0,\ldots,6\}$.

These preprocessing steps improved model stability and fairness across classes.

6 Model Training and Evaluation

6.1 Artificial Neural Network

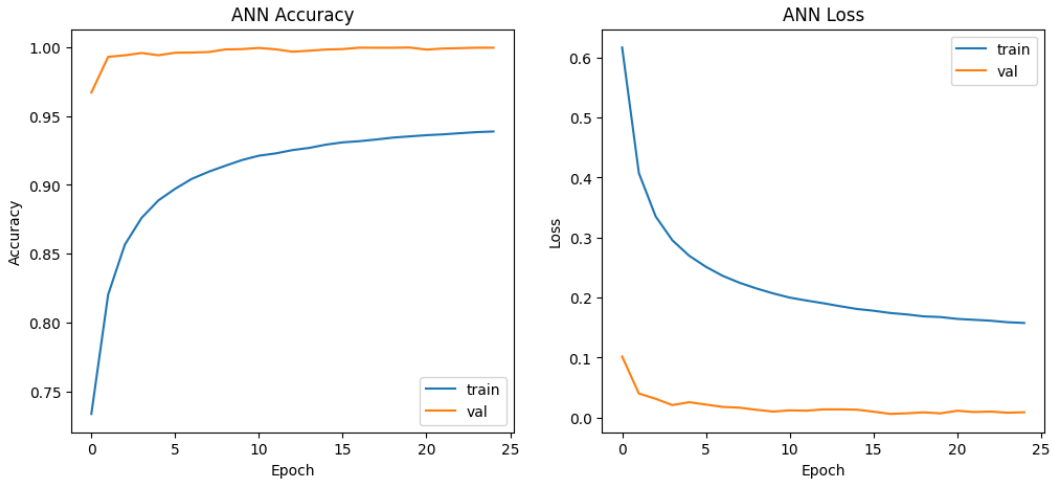

Figure 6: Training and validation accuracy/loss curves for the ANN.

The ANN demonstrates smooth convergence, with training accuracy rising consistently and validation accuracy stabilizing near unity. The loss curves do not diverge, indicating that regularization methods such as BatchNorm and Dropout effectively control overfitting.

Confusion Matrix

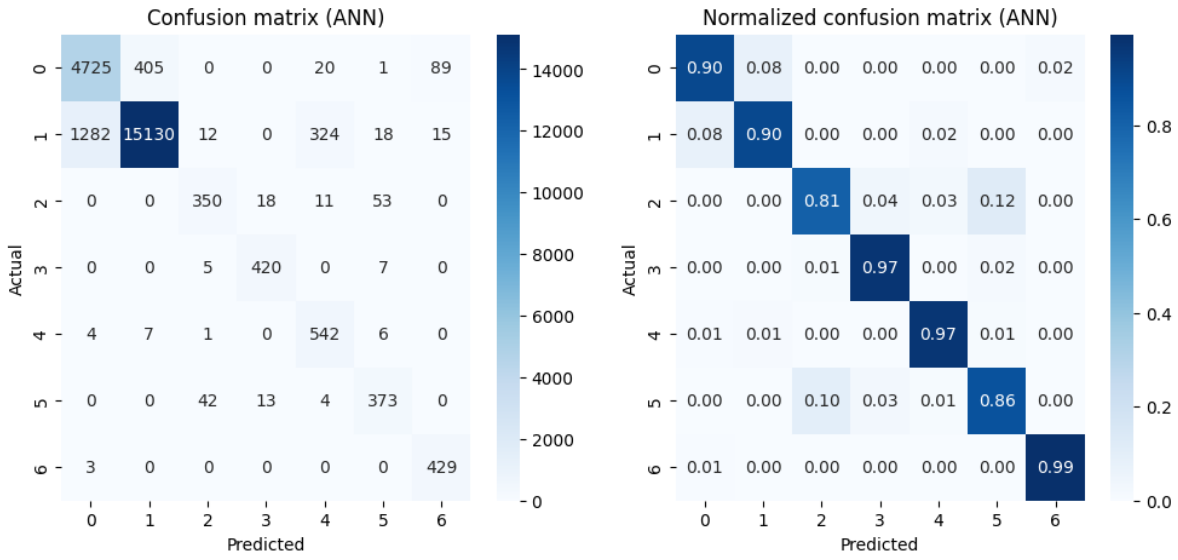

Figure 7: Confusion matrix for the ANN model.

The ANN successfully identifies minority classes that linear models struggle with. Certain classes with similar ecological profiles exhibit occasional misclassification, but overall cross-class separation is strong.

6.2 Logistic Regression

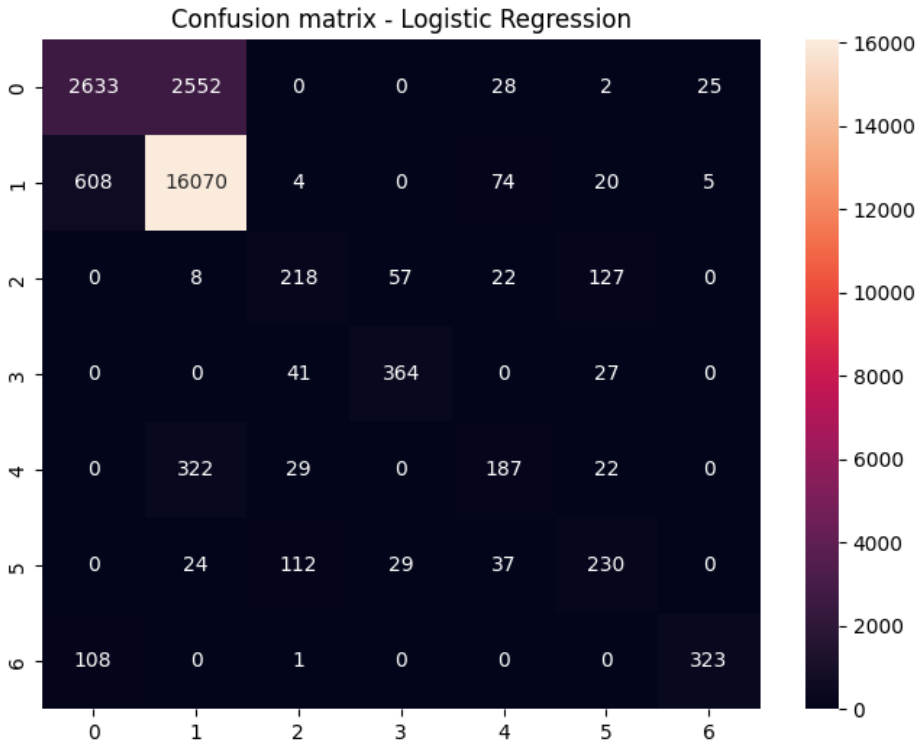

Figure 8: Confusion matrix for Logistic Regression.

Logistic Regression is heavily biased toward majority classes due to linear decision boundaries and class imbalance. Minority classes are often merged into a few dominant categories, yielding low macro-level performance.

6.3 Linear SVM (SGD)

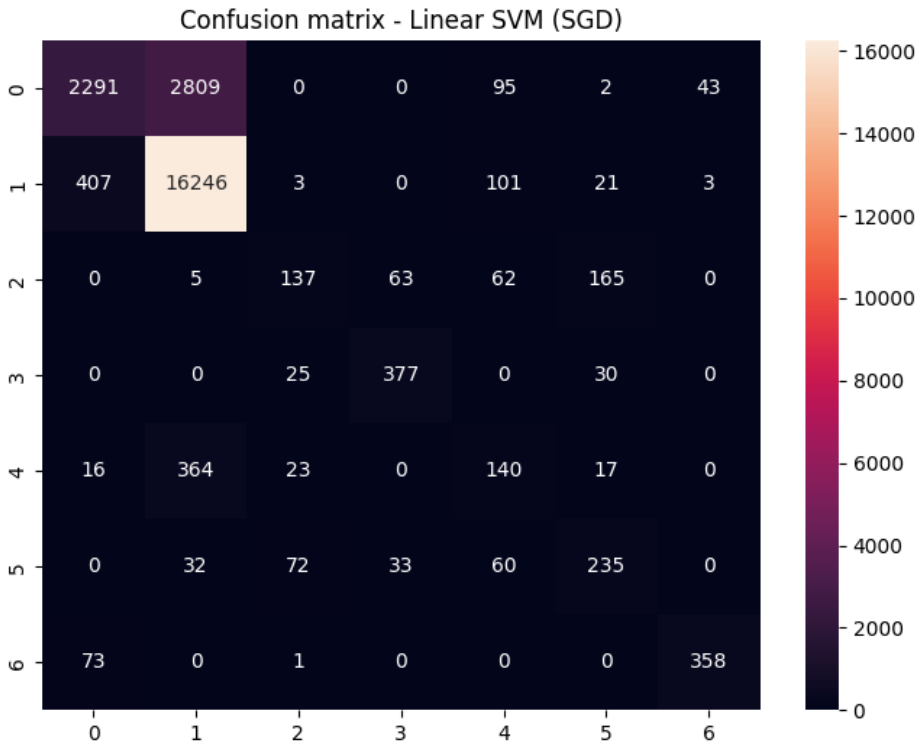

Figure 9: Confusion matrix for Linear SVM (SGD).

Linear SVM offers modest improvement over Logistic Regression but still fails to differentiate complex ecological patterns. Like LR, it relies on linear separability that is not present in this dataset.

7 Results and Discussion

7.1 Model Comparison

Model	Accuracy	Macro F1	Weighted F1
ANN	0.9037	0.8611	0.9067
Logistic Regression	0.8238	0.6611	0.8102
Linear SVM (SGD)	0.8139	0.6212	0.7947

Table 1: Comparison of the three models across key evaluation metrics.

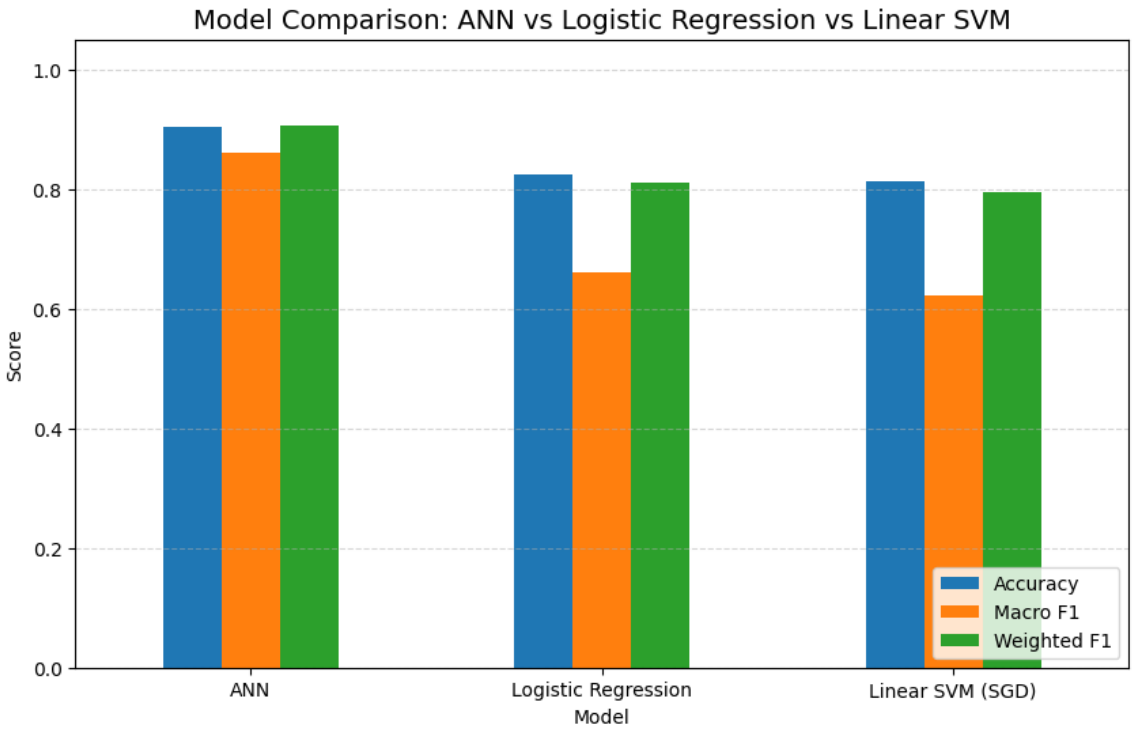

Figure 10: Accuracy and F1-score comparison across models.

The ANN significantly outperforms both linear models across all evaluated metrics. Its ability to model nonlinear interactions, combined with SMOTE balancing and architectural regularization, supports more accurate and generalizable predictions. Logistic Regression and SVM are limited by linear boundaries and disproportionate sensitivity to majority classes.

8 Conclusion

This study applied a comprehensive machine learning workflow to classify forest cover types based on terrain and environmental indicators. The analysis demonstrated that the dataset comprises complex, nonlinear relationships that cannot be captured effectively by classical linear models. The ANN achieved superior accuracy and balanced performance across all cover types, illustrating its suitability for ecological classification tasks of this scale and complexity.

Key findings include the substantial predictive value of elevation and hydrology-based features, the structural necessity of class balancing, and the importance of nonlinear learning architectures. Future extensions may explore ensemble tree-based models, spatial modeling approaches, or deep architectures specifically adapted to terrain data.

Resources

- Pandas: `https://pandas.pydata.org/`

- Scikit-learn: `https://scikit-learn.org/`
- TensorFlow/Keras: `https://www.tensorflow.org/`
- Matplotlib: `https://matplotlib.org/`
- Seaborn: `https://seaborn.pydata.org/`