International Institute of Information Technology
Bangalore (IIITB)

Comparative Analysis of Supervised and Unsupervised Learning Models for Forest Cover Type Classification

Rahul Raman (MT2025100) and Aayank Singhai (MT2025001)

Machine Learning Project Report

December 12, 2025

Abstract

This report presents a comprehensive machine learning analysis aimed at predicting forest cover types based on cartographic variables derived from US Geological Survey (USGS) data. The study addresses a multi-class classification problem with seven distinct cover types characterized by significant class imbalance. The experimental pipeline involves rigorous data preprocessing, including standardization and dimensionality reduction via Principal Component Analysis (PCA) retaining 95% variance. We evaluate four supervised learning models—Logistic Regression, Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Random Forest—alongside unsupervised clustering methods (K-Means and DBSCAN). Our results demonstrate that the Random Forest classifier achieves the highest performance with an accuracy of 89.60%, effectively handling the non-linear complexities of the topographical data.

Contents

1 Introduction

Forest cover type classification is a critical task in environmental science and resource management. The objective of this study is to predict the dominant tree species (e.g., Spruce/Fir, Lodgepole Pine, Ponderosa Pine) in 30x30 meter cells based on attributes such as elevation, slope, soil type, and distance to water bodies.

The dataset presents several challenges:

1. **High Dimensionality:** The inclusion of 40 binary columns for soil types creates a sparse feature space.
2. **Class Imbalance:** Two cover types (Spruce/Fir and Lodgepole Pine) constitute the majority of the data, potentially biasing models.
3. **Non-linear Decision Boundaries:** Topographical dependencies are complex and rarely linearly separable.

This report details the methodology adopted to overcome these challenges, comparing linear models against non-linear and ensemble approaches.

2 Dataset Characterization

2.1 Data Source and Sampling

The dataset utilized is the standard `covtype.csv`. To optimize computational resources for hyperparameter tuning, a stratified random sample of 50% of the original data was generated.

- **Total Samples Analyzed:** 290,506
- **Feature Count:** 54 (10 continuous, 44 binary)
- **Target Classes:** 7 (Integers 1 through 7)

2.2 Feature Description

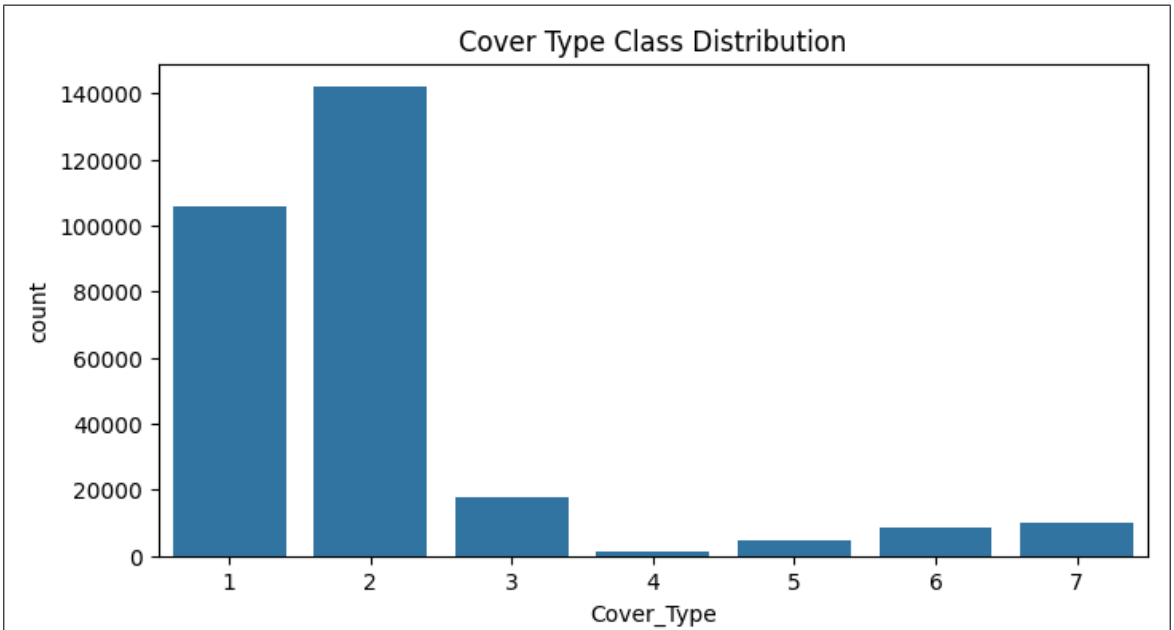
The features capture various physical characteristics of the terrain:

- **Elevation, Aspect, Slope:** Fundamental topographical metrics.
- **Hydrology Distances:** Both horizontal and vertical distances to the nearest water source, critical for vegetation growth.
- **Hillshade Indices:** Measures of shade at 9am, Noon, and 3pm, serving as proxies for sunlight exposure.
- **Wilderness Areas:** 4 binary columns representing different geological zones.
- **Soil Types:** 40 binary columns indicating specific soil substrata.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Target Distribution

As visualized in Figure 1, the dataset is heavily skewed. Cover Types 1 and 2 account for nearly 70% of the observations, while Type 4 is extremely rare (¡1%).



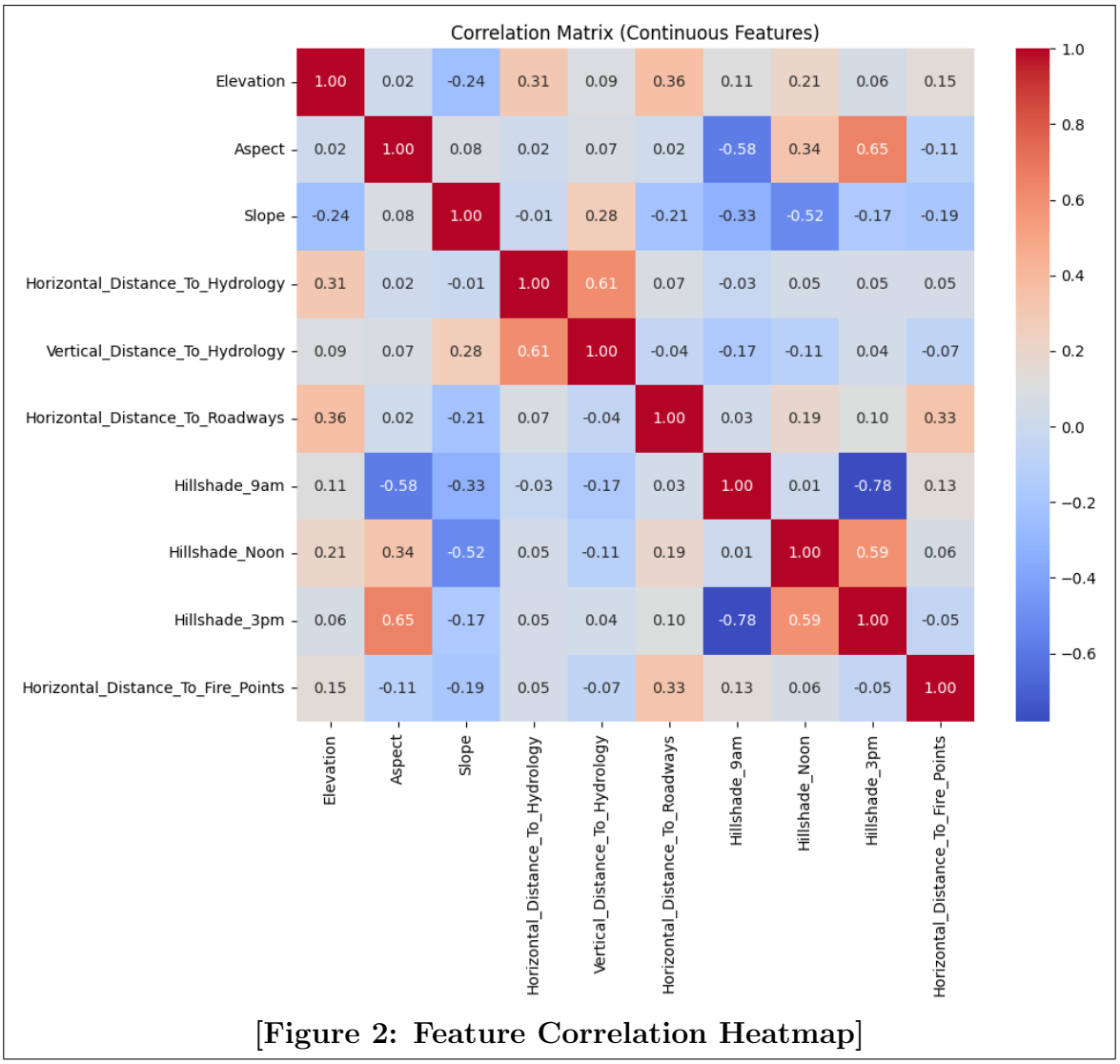
[Figure 1: Target Class Distribution (Countplot)]

Figure 1: Distribution of Forest Cover Types. Note the significant imbalance between Types 1/2 and Type 4.

2.3.2 Feature Correlations

The correlation matrix (Figure 2) reveals significant multicollinearity among certain features:

- **Hillshade Attributes:** Strong correlations exist between Hillshade_9am and Hillshade_3pm due to the sun's trajectory.
- **Hydrology:** Horizontal and Vertical distances to hydrology are positively correlated, suggesting a relationship between distance and elevation gain relative to water.



[Figure 2: Feature Correlation Heatmap]

Figure 2: Heatmap displaying Pearson correlation coefficients between continuous features.

3 Data Preprocessing Pipeline

3.1 Feature Scaling

Machine learning algorithms such as SVM and Neural Networks are sensitive to the magnitude of input features. Features like `Elevation` (range 1800-4000) dominate features like `Slope` (range 0-60) in distance calculations.

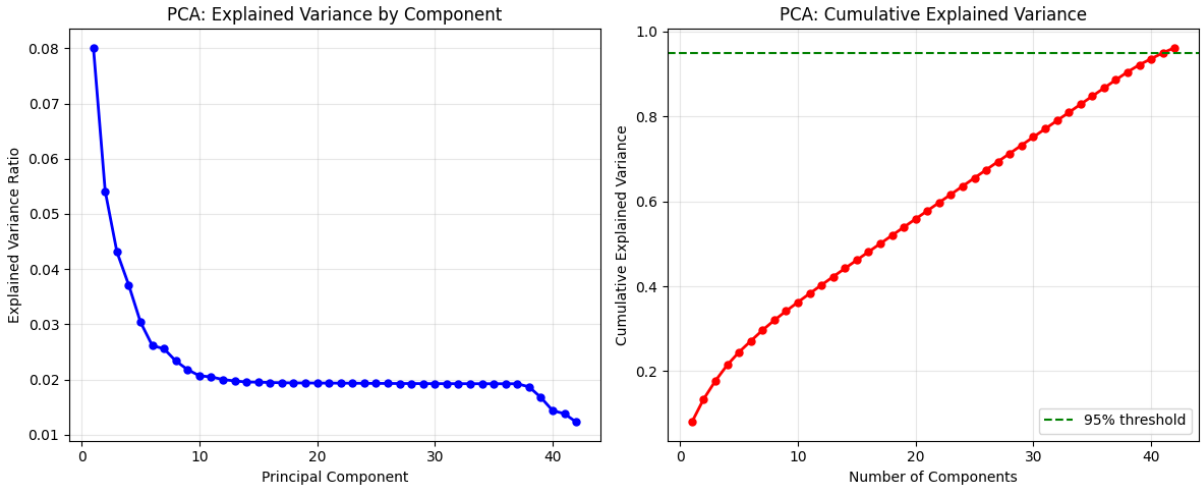
- **Method:** `StandardScaler`
- **Effect:** All features were transformed to have a mean of 0 and a standard deviation of 1.

3.2 Dimensionality Reduction

The dataset contains 44 sparse binary columns (Soil Types and Wilderness Areas). To reduce computational complexity and potential noise, Principal Component Analysis (PCA)

was employed.

- **Configuration:** `n_components=0.95` (Retain 95% variance).
- **Result:** Dimensionality reduced from 54 to 42 components, achieving a 22% reduction in feature space while preserving the majority of information.



3.3 Data Splitting

The reduced dataset was split into training and testing subsets:

- **Split Ratio:** 70% Training / 30% Testing.
- **Strategy:** Stratified split to ensure the class distribution in the train/test sets mirrors the original population.

4 Model Methodologies

Hyperparameter tuning was performed using `RandomizedSearchCV` to efficiently explore the parameter space.

4.1 Logistic Regression (Baseline)

A linear model was chosen as a baseline to test linear separability.

- **Parameters Tuned:** Regularization strength (C), Solver (`lbfgs`, `liblinear`).
- **Best Configuration:** $C = 10$, `solver='lbfgs'`.

4.2 Support Vector Machine (SVM)

SVM was selected for its ability to create non-linear decision boundaries using the Kernel trick.

- **Parameters Tuned:** Kernel type (`rbf`, `poly`), Regularization (C).
- **Best Configuration:** $C = 100$, `kernel='rbf'`. The Radial Basis Function proved effective for the complex topography.

4.3 Neural Network (MLP)

A Multi-Layer Perceptron was trained to capture high-order feature interactions.

- **Architecture:** Two hidden layers (100 neurons, 50 neurons).
- **Optimization:** Adam optimizer, ReLU activation, with early stopping to prevent overfitting.

4.4 Random Forest Classifier

An ensemble of decision trees was used to handle the non-linear nature of the data and provide robustness against noise.

- **Parameters Tuned:** Number of estimators, Max depth, Max features.
- **Best Configuration:** `n_estimators=100`, `max_depth=20`.

4.5 Unsupervised Clustering

To analyze the natural grouping of the data, K-Means and DBSCAN were applied. Clusters were mapped to the most frequent true label to calculate classification metrics.

- **K-Means:** $k = 4$ (determined via Elbow Method).
- **DBSCAN:** Density-based clustering with `eps=0.75`.

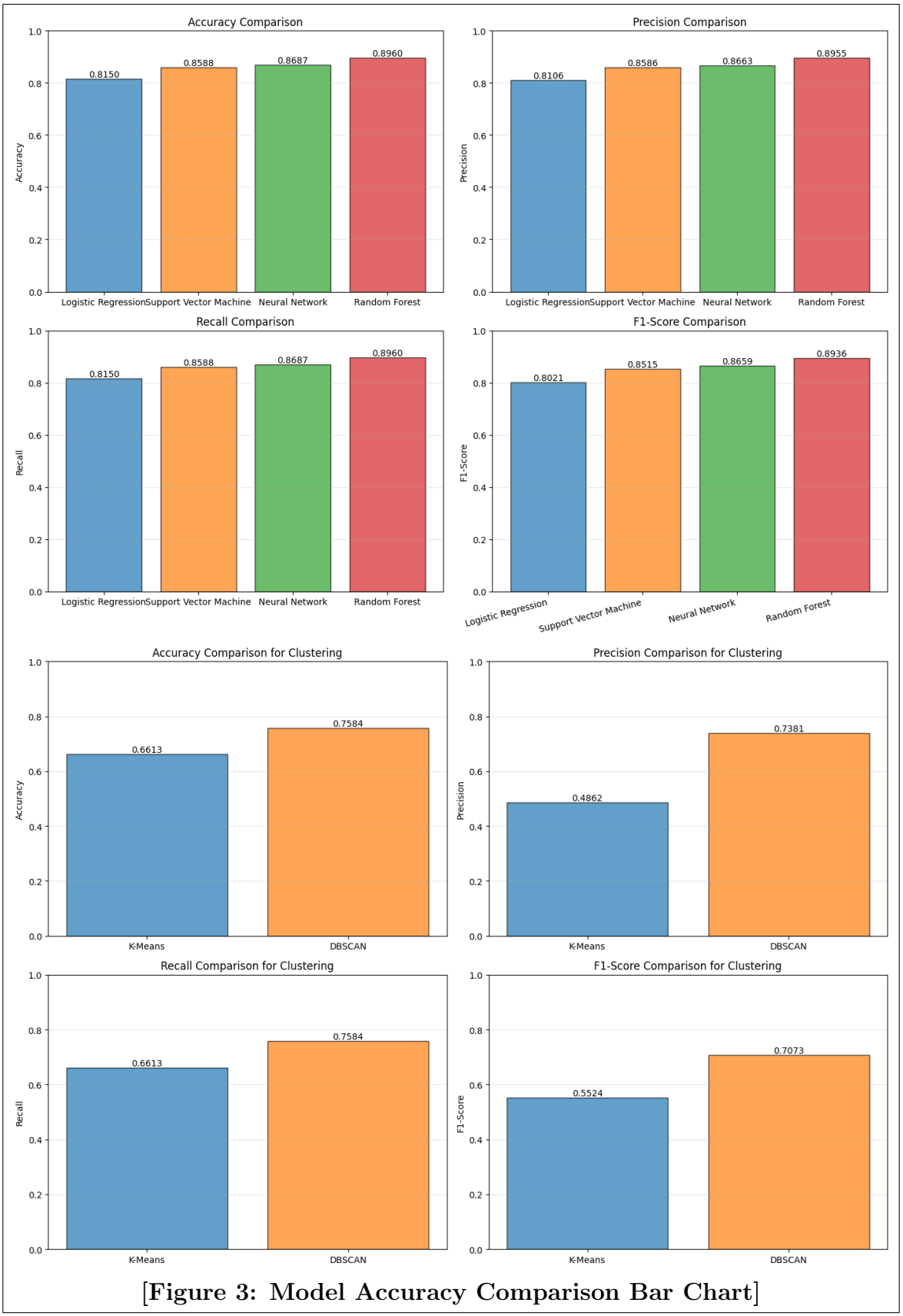
5 Experimental Results and Analysis

5.1 Performance Summary

Table 1 presents the performance metrics on the unseen test set. The Random Forest classifier outperformed all other models significantly.

Table 1: Comparative Performance Metrics (Test Set)

Model	Accuracy	W. Precision	W. Recall	W. F1-score
Logistic Regression	0.8150	0.8106	0.8150	0.8021
SVM (RBF)	0.8588	0.8586	0.8588	0.8515
Neural Network (MLP)	0.8687	0.8663	0.8687	0.8659
Random Forest	**0.8960**	**0.8955**	**0.8960**	**0.8936**
K-Means (Clustering)	0.6613	0.4862	0.6613	0.5524
DBSCAN (Clustering)	0.7584	0.7381	0.7584	0.7073



[Figure 3: Model Accuracy Comparison Bar Chart]

Figure 3: Visual comparison of model accuracies. Random Forest leads the supervised models.

5.2 Deep Dive Analysis

5.2.1 Why Random Forest Won?

The Random Forest model achieved an accuracy of **89.60%**. Its superior performance can be attributed to:

1. **Feature Interactions:** Topographical rules are often hierarchical (e.g., "If Elevation ¿ 3000m AND Soil is Type 10..."). Decision trees model these interactions natively.

2. **Robustness:** By averaging 100 trees, the model reduced variance and resisted overfitting better than the single Decision Tree or MLP.

3. **Handling Sparsity:** Even though PCA was applied, the underlying data structure remains distinct across cover types, which the non-linear splits of the trees exploited effectively.

5.2.2 Linear vs. Non-Linear Models

The gap between Logistic Regression (81.5%) and SVM/MLP (86%+) confirms that the decision boundaries between forest cover types are highly non-linear. The linear model struggled to differentiate between overlapping classes (specifically Cover Types 1 and 2), while the RBF kernel of the SVM and the hidden layers of the MLP could map these complex boundaries.

5.2.3 Clustering Limitations

Unsupervised learning proved ineffective for this classification task.

- **K-Means (66.1%):** Failed because it assumes spherical clusters. Forest cover zones are often irregular and elongated along elevation lines.

- **DBSCAN (75.8%):** Performed better by recognizing density-based shapes but still lagged behind supervised methods because the classes are contiguous and lack clear density separation gaps in the feature space.

6 Conclusion and Future Scope

6.1 Conclusion

This study successfully evaluated multiple machine learning approaches for forest cover classification. **Random Forest** emerged as the optimal choice, delivering high accuracy and a balanced F1-score across all classes. The analysis highlighted the necessity of non-linear models for topographical data and demonstrated that while dimensionality reduction (PCA) improves efficiency, the ensemble tree methods provide the necessary complexity to model the environment accurately.

6.2 Future Recommendations

- **Feature Engineering:** Creating a "Euclidean Distance to Hydrology" feature (combining vertical and horizontal distances) may help linear models.
- **Advanced Ensembles:** Implementing Gradient Boosting (XGBoost or LightGBM) could potentially push accuracy beyond 90% by correcting errors of previous trees sequentially.
- **Raw Data RF:** Testing Random Forest *without* PCA. Trees handle sparse binary data well; PCA might have obscured some sharp categorical distinctions in Soil Types.

Link GitHub Repository