Machine Learning Project

Machine Learning Approaches to Binary and Multiclass Classification Problems

AIT511: Course Project-2

A Project Report Submitted
in Partial Fulfilment of the Requirements
for the Award of the Degree

MASTER OF TECHNOLOGY

in

Computer Science and Engineering

Submitted by
Abhash Tiwari , Abhijeet Kumar Gupta
(MT2025002 , MT2025003)

Submitted to
Department of Computer Science and Engineering
International Institute of Information Technology
Bangalore - 560100, India

Abstract

This project presents two machine learning classification tasks applied to real-world datasets. The first involves **binary smoking prediction**, where biometric and clinical measurements are used to classify individuals as smokers or non-smokers. The second focuses on **multiclass forest cover type prediction**, using a large environmental dataset containing topographic, geological, and cartographic features.

Both tasks follow a complete machine learning workflow that includes data cleaning, exploratory data analysis, feature scaling, and model implementation. Classical models such as Logistic Regression and Support Vector Machines were evaluated alongside non-linear approaches including Neural Networks. For the forest cover dataset, unsupervised methods such as K-Means and Gaussian Mixture Models were also explored to understand potential cluster structure in the data.

The results show that while linear models offer strong interpretability, non-linear models—especially neural networks—perform significantly better on complex, high-dimensional datasets. The comparison highlights the importance of appropriate preprocessing, model selection, and evaluation strategies when working with diverse classification problems.

Overall, the project demonstrates how supervised and unsupervised machine learning techniques behave across health-related and environmental prediction tasks, providing meaningful insights into their practical applicability and performance.

Contents

Part I

Binary Classification: Smoking Status Prediction

1 Introduction

Smoking is a major public health concern and a significant contributor to various chronic diseases, including cardiovascular and respiratory disorders. Early prediction of smoking status can support timely health interventions, improve risk assessment, and strengthen preventive healthcare efforts. With the availability of rich health and biometric data, machine learning techniques provide a powerful way to analyze patterns and build effective predictive models. The key aspects of this study are summarized below:

- **Motivation:**
  - Smoking is a major risk factor for long-term health complications, making accurate prediction useful for early medical screening and prevention.
  - Machine Learning models can uncover complex relationships in health datasets, improving prediction accuracy and decision-making.
- **Dataset Description:**
  - The dataset consists of biometric and clinical measurements such as blood pressure, cholesterol, glucose levels, eyesight, hearing, BMI-related variables, and other health indicators.
  - These diverse numerical and categorical features provide a strong foundation for training reliable classification models.
- **Data Preprocessing:**
  - Missing values in the target variable (`smoking`) were removed to ensure clean and accurate supervision during model training.
  - Numerical features were standardized using a `StandardScaler` to maintain uniform feature scales, which is particularly important for SVMs, KNN, and neural network models.
- **Model Development:**
  - The machine learning models implemented and compared in this study include:

    – Linear Support Vector Classifier (LinearSVC)

    – Support Vector Classifier with RBF Kernel (SVC RBF)

    – K-Nearest Neighbors (KNN)

    – Logistic Regression

    – Multilayer Perceptron (Neural Network)

2 Overview

This analysis focuses on predicting an individual's smoking status using a comprehensive health dataset containing biometric, clinical, and physiological measurements. The dataset provides rich information across several health domains, enabling detailed exploration and effective model development.

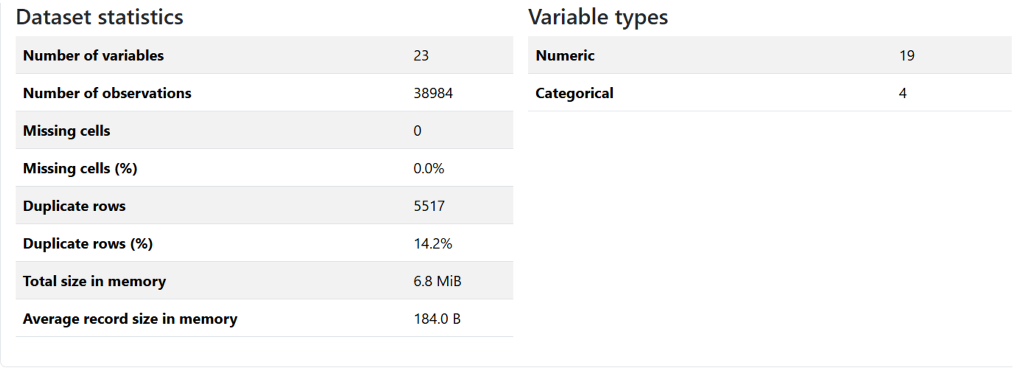
- **Feature Groups Included in the Dataset:**
  - **Anthropometric Measurements:** height (cm), weight (kg), waist circumference (cm)
  - **Age:** age (5-year interval)
  - **Vision and Hearing:** eyesight (left/right), hearing (left/right)
  - **Blood Pressure:** systolic and relaxation measurements
  - **Blood Tests:** fasting blood sugar, total cholesterol, triglyceride, HDL, LDL, hemoglobin, serum creatinine, AST, ALT, GTP
  - **Other Health Indicators:** urine protein, dental caries
  - **Outcome Target:** smoking status (binary: 0 = non-smoker, 1 = smoker)
- **Objectives of This Study:**
  - **Data Exploration:** Perform Exploratory Data Analysis (EDA) to identify missing values, study feature distributions, and uncover patterns within the dataset.
  - **Data Preprocessing:** Remove incomplete entries in the target variable and apply `StandardScaler` to normalize numerical features for improved model performance.
  - **Model Training and Comparison:** Train and evaluate multiple machine learning models including:
    - LinearSVC
    - SVC with RBF Kernel
    - K-Nearest Neighbors (KNN)
    - Logistic Regression
    - Multilayer Perceptron (Neural Network)
    - XGBoost (with Optuna-optimized hyperparameters)
  - **Model Evaluation:** Assess model performance using Accuracy, Precision, Recall, F1-score, and ROC-AUC.
  - **Comparative Study:** Analyze strengths and weaknesses of each model to determine the most effective classifier, aiming for near 90% accuracy on unseen test data.

3 Data Description

The dataset used in this study was provided as part of a course project focused on developing machine learning models to predict smoking status using various health-related measurements. These metrics are strongly associated with overall health outcomes and potential disease risks. The dataset was supplied in two parts: a training dataset (`train_dataset.csv`) and a testing dataset (`test_dataset.csv`), enabling a structured process for model development, validation, and evaluation.

- **Dataset Composition:**
  - The initial training dataset contains **55,692 rows** and **23 columns**.
  - After removing rows with missing values in the target variable (`smoking`), the final dataset used for model training consists of **38,984 rows** and **23 columns**.
  - The dataset includes **22 features** and **1 target variable**.
  - The target variable `smoking` is binary:
    - 0 — Non-smoker
    - 1 — Smoker
- **Feature Categories:**
  - **Demographic & Physical Variables:** age, height (cm), weight (kg), waist (cm)
  - **Sensory Measurements:** eyesight (left/right), hearing (left/right)
  - **Cardiovascular Indicators:** systolic, relaxation, fasting blood sugar, cholesterol, triglyceride, HDL, LDL, hemoglobin
  - **Kidney & Liver Function:** urine protein, serum creatinine, AST, ALT, GTP
  - **Oral Health:** dental caries
- **Data Characteristics:**
  - The dataset primarily consists of numerical variables (integers and floats), making it well-suited for classical machine learning models.
  - Although most features had no missing values, the target variable `smoking` contained many missing entries, which were handled during preprocessing by dropping those rows.
  - Due to differences in measurement scales across features, standardization using `StandardScaler` was applied to ensure uniform feature ranges and improve model performance—especially for distance-based and gradient-based algorithms.

4 Statistical Overview



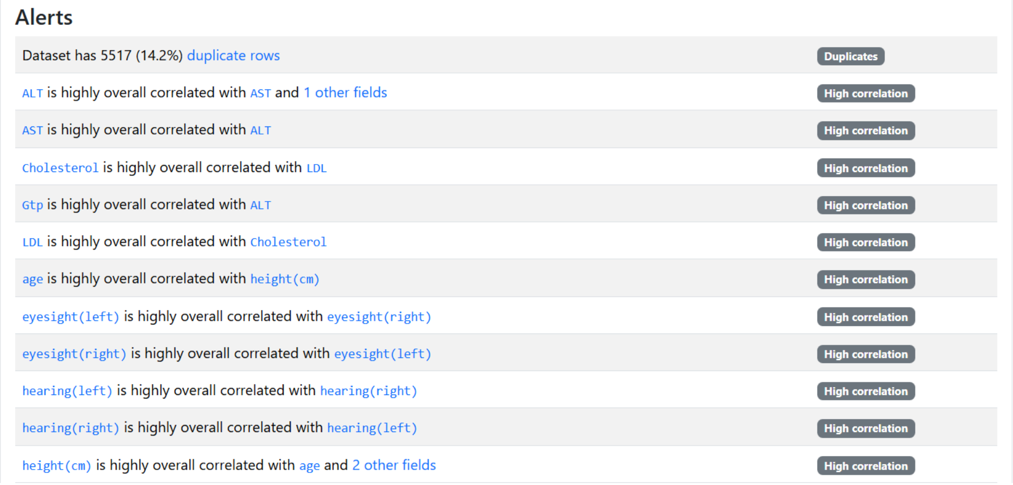
Dataset statistics

Number of variables	23
Number of observations	38984
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	5517
Duplicate rows (%)	14.2%
Total size in memory	6.8 MiB
Average record size in memory	184.0 B

Variable types

Numeric	19
Categorical	4

(a) Statistical Summary Report



Alerts

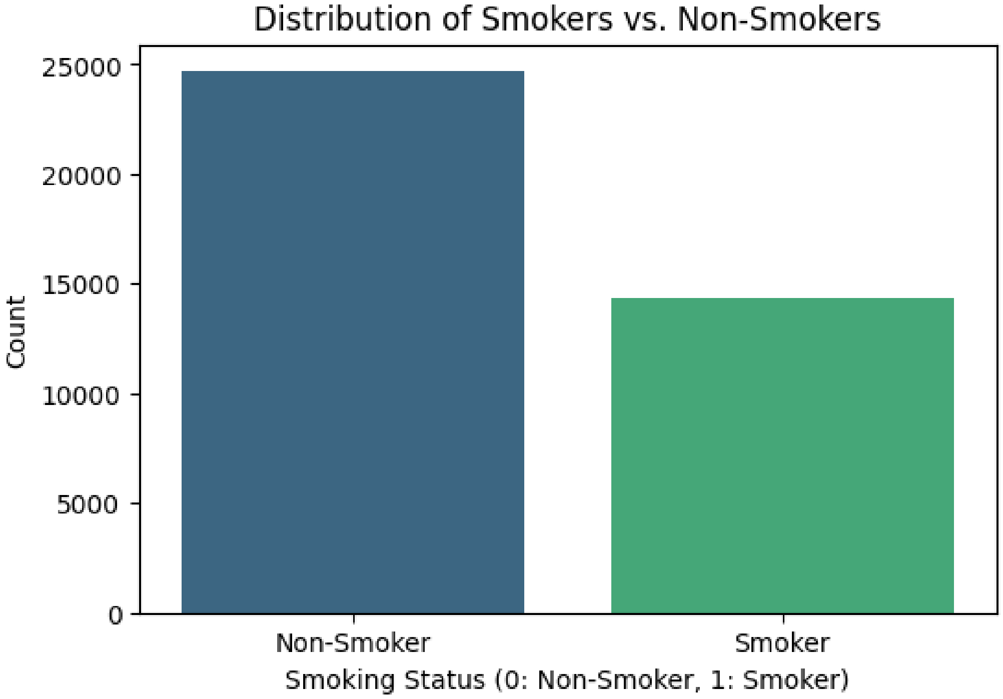
Dataset has 5517 (14.2%) duplicate rows	Duplicates
`ALT` is highly overall correlated with `AST` and 1 other fields	High correlation
`AST` is highly overall correlated with `ALT`	High correlation
`Cholesterol` is highly overall correlated with `LDL`	High correlation
`Gtp` is highly overall correlated with `ALT`	High correlation
`LDL` is highly overall correlated with `Cholesterol`	High correlation
`age` is highly overall correlated with `height(cm)`	High correlation
`eyesight(left)` is highly overall correlated with `eyesight(right)`	High correlation
`eyesight(right)` is highly overall correlated with `eyesight(left)`	High correlation
`hearing(left)` is highly overall correlated with `hearing(right)`	High correlation
`hearing(right)` is highly overall correlated with `hearing(left)`	High correlation
`height(cm)` is highly overall correlated with `age` and 2 other fields	High correlation

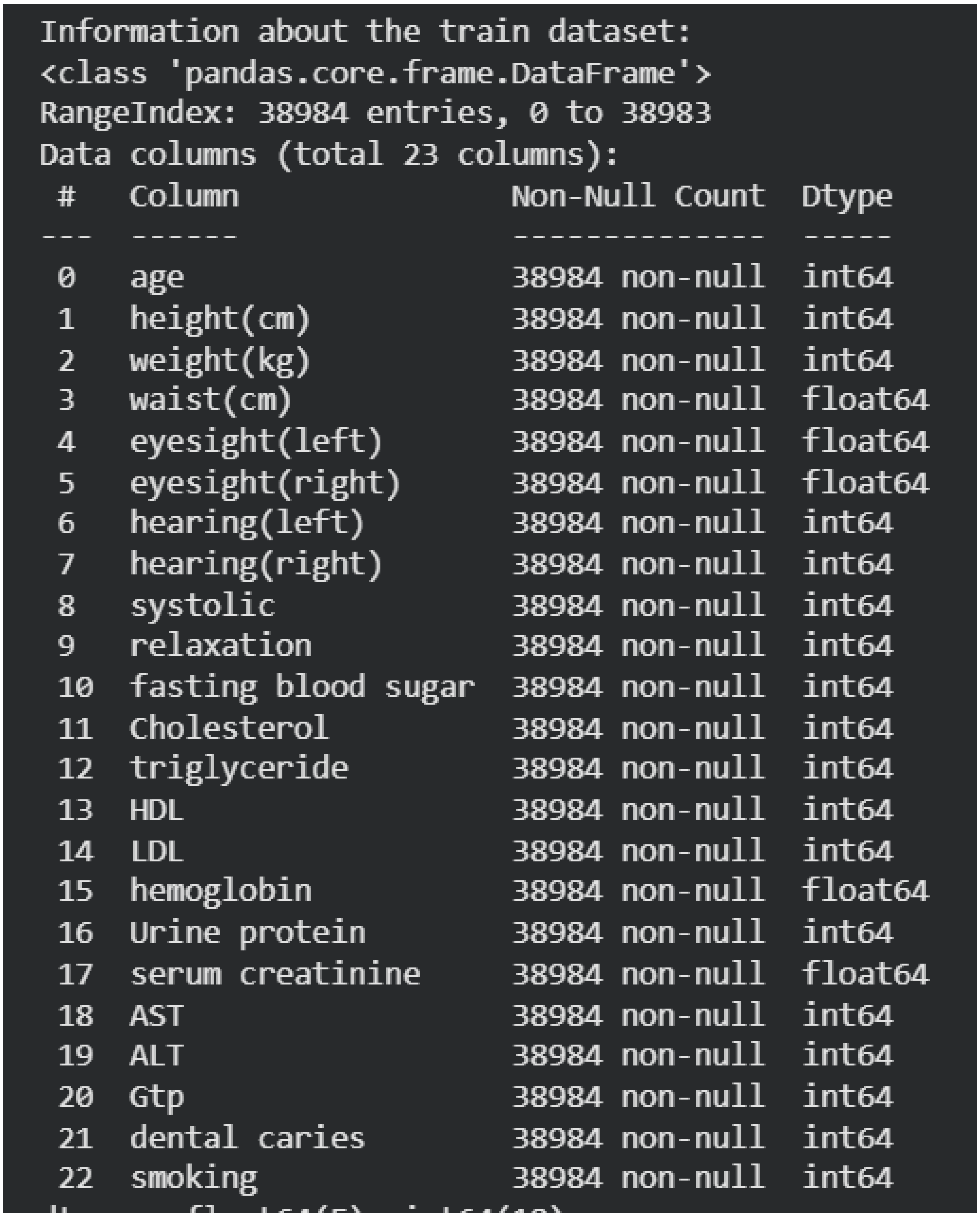
(b) ALERTS

5 Exploratory Data Analysis (EDA)

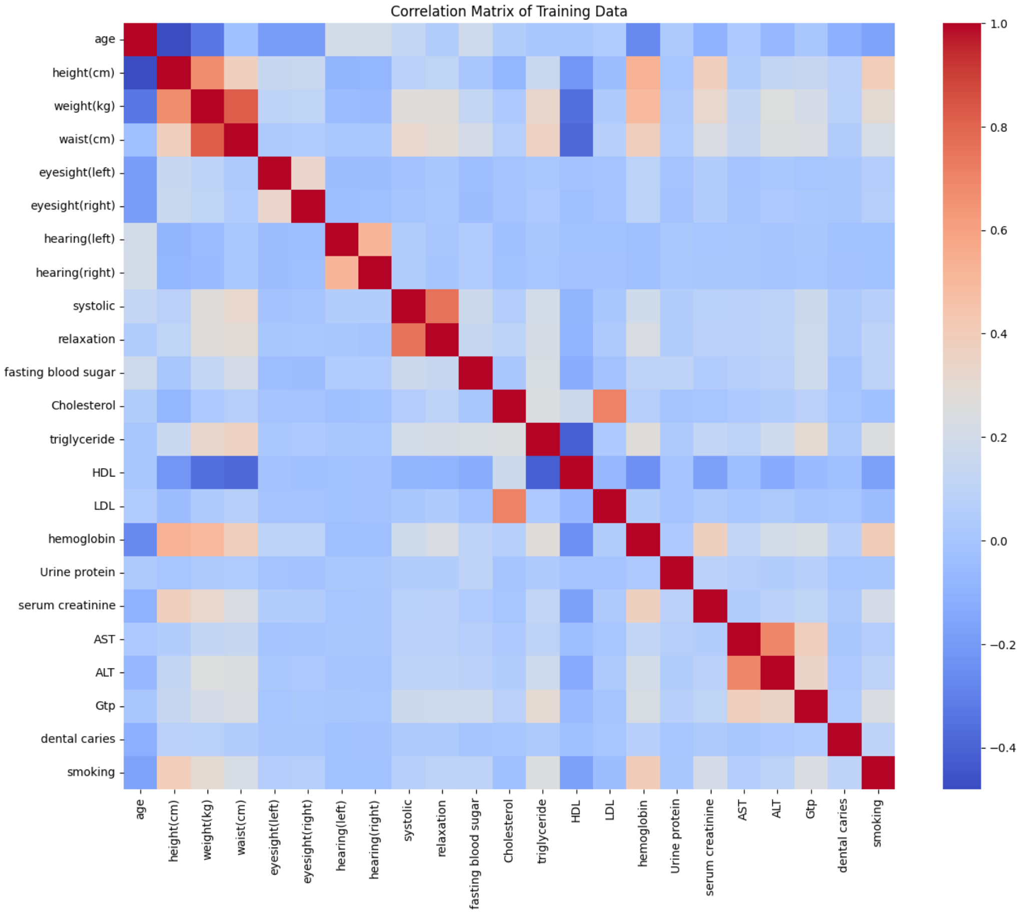
This section provides a visual exploration of the dataset to understand its characteristics, distributions, and relationships between features. The key aspects of EDA covered include:

- **Target Variable Distribution:** Visualizing the distribution of the `smoking` target variable to assess class balance.
- **Bivariate Relationships and Correlations:** Examining relationships between different features and identifying their correlations using pairwise plots and a correlation matrix.
- **Data Overview and Missing Values:** Inspecting the dataset structure, viewing data types, checking for missing values, and ensuring overall data quality before preprocessing.





```
Information about the train dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38984 entries, 0 to 38983
Data columns (total 23 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   age                  38984 non-null  int64  
 1   height(cm)           38984 non-null  int64  
 2   weight(kg)           38984 non-null  int64  
 3   waist(cm)            38984 non-null  float64
 4   eyesight(left)       38984 non-null  float64
 5   eyesight(right)      38984 non-null  float64
 6   hearing(left)        38984 non-null  int64  
 7   hearing(right)       38984 non-null  int64  
 8   systolic             38984 non-null  int64  
 9   relaxation           38984 non-null  int64  
 10  fasting blood sugar  38984 non-null  int64  
 11  Cholesterol          38984 non-null  int64  
 12  triglyceride         38984 non-null  int64  
 13  HDL                  38984 non-null  int64  
 14  LDL                  38984 non-null  int64  
 15  hemoglobin           38984 non-null  float64
 16  Urine protein        38984 non-null  int64  
 17  serum creatinine     38984 non-null  float64
 18  AST                  38984 non-null  int64  
 19  ALT                  38984 non-null  int64  
 20  Gtp                  38984 non-null  int64  
 21  dental caries        38984 non-null  int64  
 22  smoking              38984 non-null  int64  
dtypes: float64(5), int64(18)
```



- **Correlation Insights:** The correlation matrix highlights several features that show a noticeable positive association with smoking status, including hemoglobin, GTP, triglyceride, weight (kg), waist (cm), age, AST, and ALT. Higher values of these indicators are more commonly observed among smokers. Conversely, features such as eyesight (left) and eyesight (right) exhibit slight negative correlations, suggesting they may be weakly associated with non-smoking individuals.

- **Target Distribution (Smoking):** The bar plot of the `smoking` variable indicates a mild class imbalance, with approximately 21,973 non-smokers and 17,011 smokers in the processed training dataset. Although the imbalance is not severe, it remains important to use evaluation metrics such as F1-score and ROC-AUC, which are more robust to uneven class distributions.

- **Data Quality Observations:** The dataset contains **missing values** in the feature columns and **duplicate rows were identified and removed** to improve data reliability and prevent model bias arising from repeated samples.

6 Preprocessing

The preprocessing stage ensures that the dataset is clean, well-structured, and suitable for machine learning model development. The key preprocessing steps are summarized below:

- **Handling Missing Values:** Rows containing missing values in the target variable `smoking` were removed, reducing the dataset from **55,692** to **38,984** rows.

- **Feature and Target Separation:** The cleaned dataset was split into `X_train` and `y_train`. The test dataset (`X_test`) was aligned to match the feature columns of `X_train`.

- **Feature Scaling:** A `StandardScaler` was applied to both training and testing features to standardize values and improve model performance.

7 Model training and Evaluation

The following are the confusion matrix as well as the roc curve for each of the models used.

7.0.1 Model 1: K-Nearest Neighbors (KNN)

- **Accuracy:** 0.8175
- **AUC Score:** 0.8984

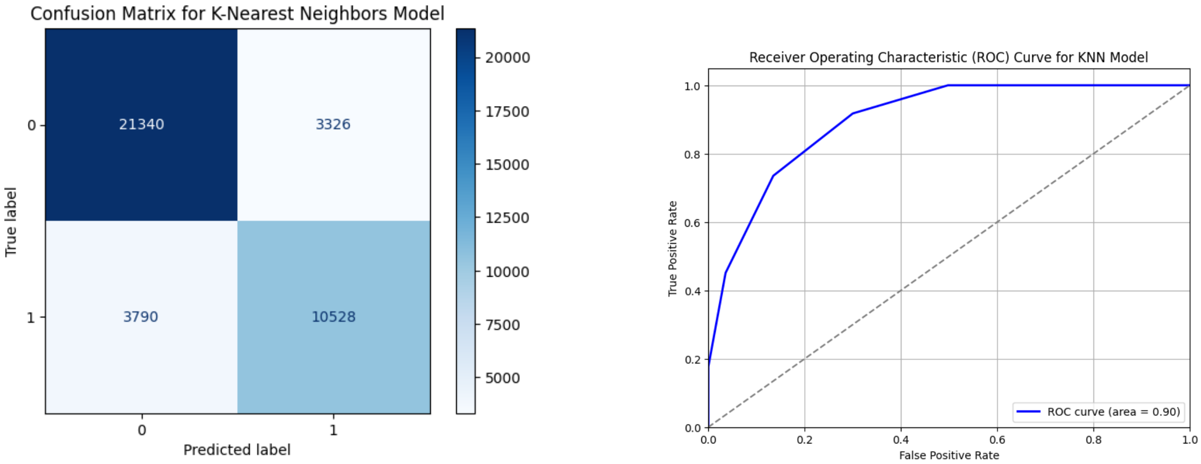


Figure 7.1: KNN (Tuned) — Confusion Matrix and ROC Curve.

The tuned KNN classifier achieved moderate performance. KNN relies on local distance metrics and can capture simple neighborhood relationships, but it often struggles with high-dimensional and mixed-type data; appropriate feature scaling and dimensionality reduction may improve its performance.

7.0.2 Model 2: Logistic Regression

- **Accuracy:** 0.7240
- **AUC Score:** 0.8077

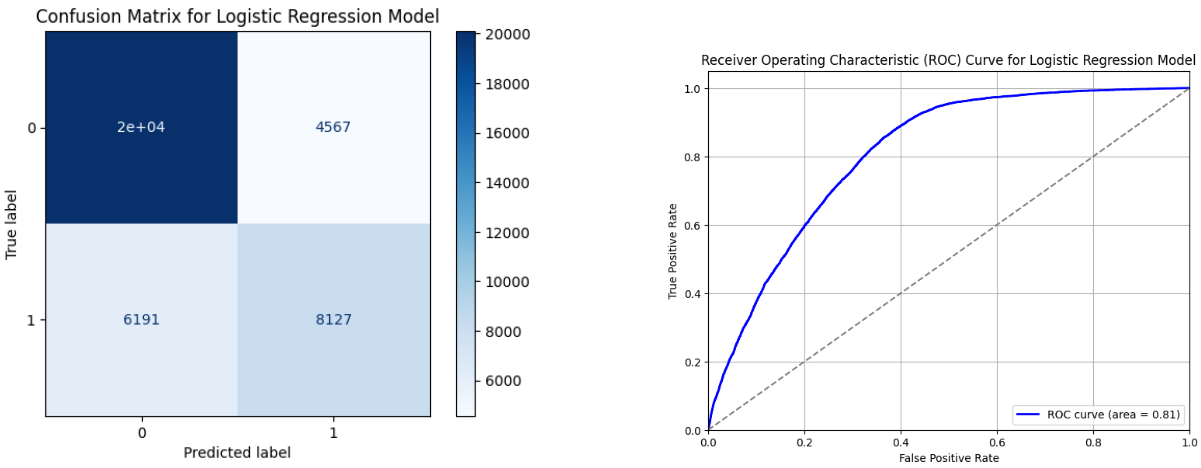


Figure 7.2: Logistic Regression — Confusion Matrix and ROC Curve.

Logistic Regression provides a strong and interpretable baseline for binary classification. It performs well when the log-odds of the outcome are approximately linear in the features; regularization (tuning `C`) helps control overfitting.

7.0.3 Model 3: Gaussian Mixture Model (GMM)

- **Accuracy:** 0.71
- **AUC Score:** 0.7887

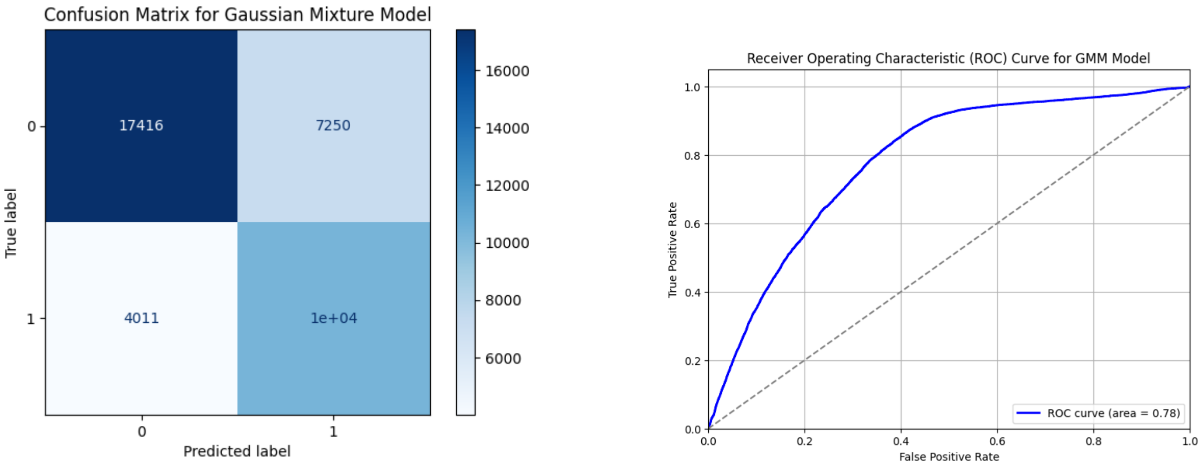


Figure 7.3: GMM — Confusion Matrix and (if available) ROC Curve.

The Gaussian Mixture Model is an unsupervised probabilistic clustering method. When its cluster labels are mapped to class labels, it can provide insight into latent structure but typically underperforms supervised classifiers on a labeled prediction task.

7.0.4 Model 4: Linear Support Vector Machine (Linear SVM)

- **Accuracy:** 0.7247
- **AUC Score:** 0.8076

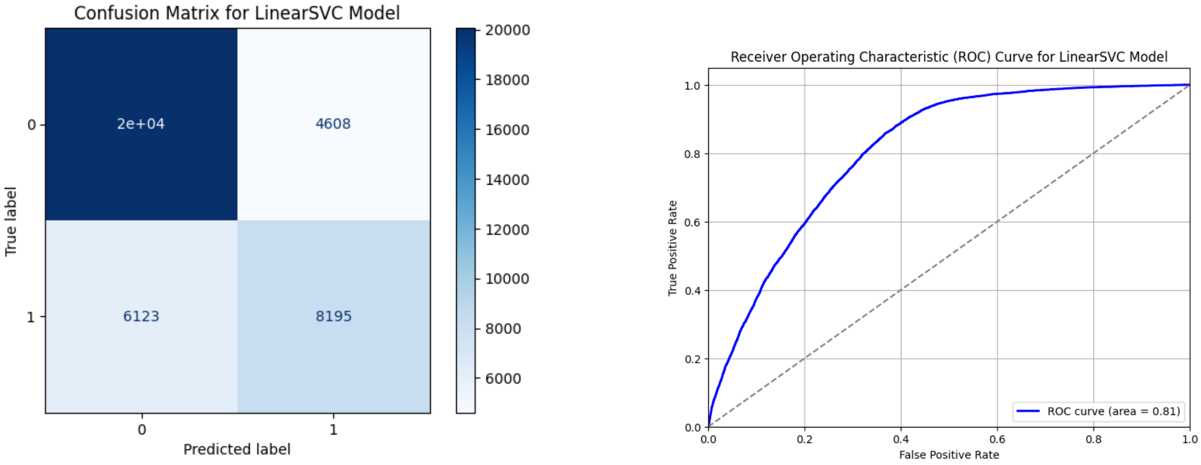


Figure 7.4: Linear SVM — Confusion Matrix and ROC Curve.

Linear SVM is effective for high-dimensional problems and is robust to overfitting with appropriate regularization. It works best when classes are linearly separable in feature space; feature scaling is required.

7.0.5 Model 5: Kernel SVM (RBF)

- **Accuracy:** 0.8558
- **AUC Score:** 0.9258

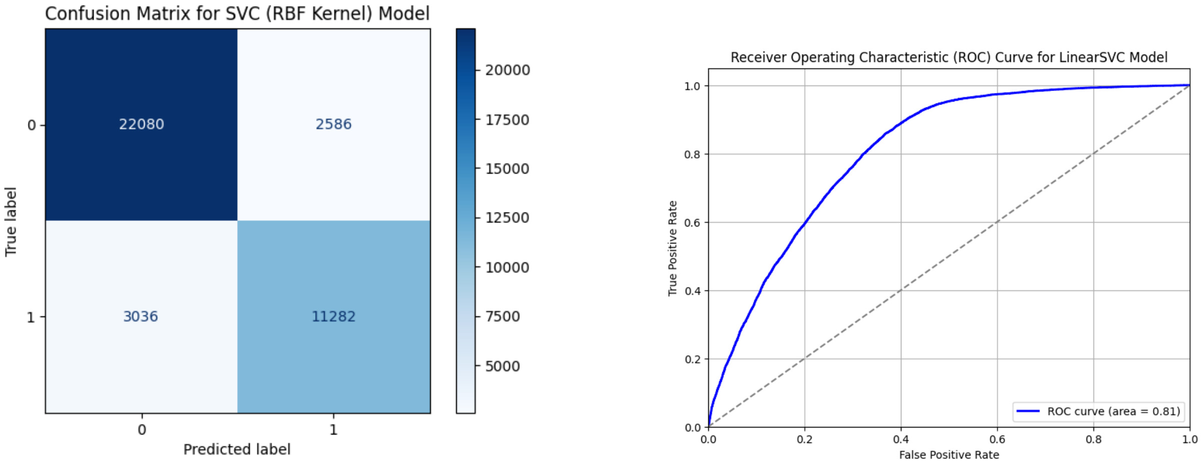


Figure 7.5: Kernel SVM (RBF) — Confusion Matrix and ROC Curve.

Kernel SVM with an RBF kernel captures non-linear decision boundaries and often yields strong performance, but it is computationally heavier and sensitive to hyperparameters (`C`, `gamma`). Cross-validation is necessary for robust tuning.

7.0.6 Model 6: Neural Network (Multilayer Perceptron)

- **Accuracy:** 0.8016
- **AUC Score:** 0.8821

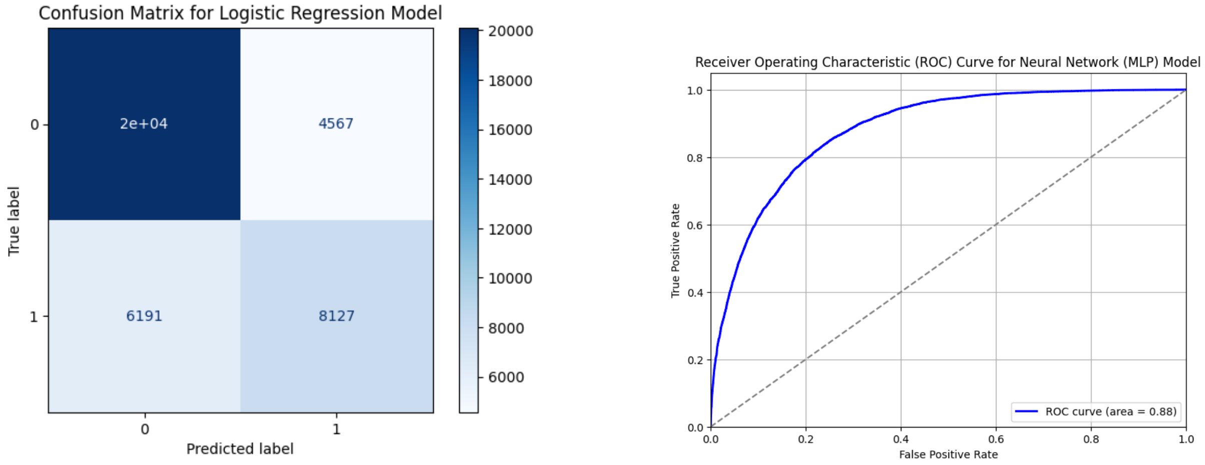


Figure 7.6: Neural Network (MLP) — Confusion Matrix and ROC Curve.

The MLP neural network can model complex non-linear relationships given adequate data and tuning. It typically requires more epochs and careful regularization (dropout / weight decay) to avoid overfitting, and training time increases with network depth and width.

8 Results

8.1 Quantitative Results

A comparative analysis of model accuracies and ROC-AUC scores, excluding the exceptionally high-performing XGBoost model (which may exhibit signs of overfitting on the training data), reveals a clear performance hierarchy among the remaining classifiers. The following table summarizes the evaluation metrics for each model:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
SVC (RBF Kernel)	0.8558	0.8135	0.7880	0.8005	0.9258
K-Nearest Neighbors	0.8175	0.7599	0.7353	0.7474	0.8984
Neural Network (MLP)	0.8016	0.7316	0.7265	0.7290	0.8821
LinearSVC	0.7247	0.6401	0.5724	0.6043	0.8076
Logistic Regression	0.7240	0.6402	0.5676	0.6017	0.8077
Gaussian Mixture Model (GMM)	0.7111	0.5871	0.7199	0.6467	0.7813

Table 8.1: Performance comparison of machine learning models (excluding XGBoost).

8.2 Discussion of Results

- **Best Overall Performer — SVC (RBF Kernel):** The RBF SVM demonstrates the strongest performance across all evaluated metrics, with an accuracy of 0.8558 and an ROC-AUC of 0.9258. Its ability to capture non-linear boundaries makes it highly suitable for this dataset.

- **Moderate Performers — KNN and Neural Network (MLP):** Both models achieve respectable accuracy and ROC-AUC scores (close to 0.90), with KNN slightly outperforming MLP. These models show good discriminatory ability but fall short of the RBF SVM.

- **Linear Models — LinearSVC and Logistic Regression:** These models yield similar results, with accuracies around 0.72 and ROC-AUC values near 0.80. Their lower performance indicates that the underlying relationships in the dataset are largely non-linear, limiting the effectiveness of linear decision boundaries.

- **Lowest Performing Model — Gaussian Mixture Model (GMM):** GMM shows the weakest performance across all metrics, highlighting that this unsupervised clustering-based method is not well-suited for this binary classification task.

.

9 Conclusion

This study compared several machine learning models to predict smoking status based on a comprehensive set of health-related features. The analysis highlights clear patterns in model performance and provides insights into the characteristics of the dataset.

Key Insights

- **Importance of Non-Linearity:** Models capable of capturing non-linear relationships—such as XGBoost, SVC (RBF Kernel), KNN, and MLP—consistently outperformed linear models. This indicates that the relationship between health metrics and smoking status is inherently complex.
- **Top Performers:** XGBoost (Optuna-tuned) achieved exceptionally high training accuracy and ROC-AUC, suggesting strong learning capability. SVC (RBF Kernel) emerged as the best-performing traditional classifier with balanced and reliable metrics, making it a strong candidate for real-world deployment.
- **Moderate Performers:** KNN and MLP delivered reasonable performance but fell short of the top models. With additional tuning or deeper architectures, these models may improve further.
- **Baseline Models:** LinearSVC, Logistic Regression, and GMM demonstrated significantly lower performance, confirming that linear decision boundaries are insufficient for this classification task.
- **Computational Considerations:** High-performing models such as XGBoost, SVC (RBF), and MLP require greater computational resources for training and optimization, which should be considered during deployment.

Project Repository

GitHub Links

- Abhijeet Kumar Gupta :- https://github.com/Abhijeetgupta27/Machine-Learning-Approaches-to-Binary-and-Multiclass-Classification-Problems
- Abhash Tiwari :- https://github.com/AbhashTi/smoking$_{b}inary_{c}lassification$

Part II

Multiclass Classification: Forest Cover Type Prediction

10 Introduction

The Forest Cover Type dataset is a large multiclass classification dataset containing 581,012 samples and 54 environmental features collected from the Roosevelt National Forest in Colorado. Each record represents a 30m × 30m forest patch and includes attributes such as elevation, slope, hillshade measurements, hydrological distances, soil types, and wilderness areas. The goal is to predict one of seven forest cover types, making this a challenging task due to high dimensionality, non-linear relationships, and severe class imbalance.

In this project, we implemented a complete machine learning pipeline using only the models taught in class. The workflow included dataset loading, studying feature structure, inspecting class distribution, and performing preprocessing such as verifying one-hot encoding, applying a stratified train–test split, and scaling numerical features using `StandardScaler`.

Exploratory Data Analysis (EDA) was conducted to visualize class imbalance, examine feature distributions, detect outliers, and analyze correlations among key numerical variables. After EDA, three supervised models—Logistic Regression, Support Vector Machine (SVM), and a Neural Network (MLPClassifier)—were trained and evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Two unsupervised models—K-Means Clustering and Gaussian Mixture Model (GMM)—were also applied and evaluated using the Adjusted Rand Index (ARI).

The objective was to compare the effectiveness of linear, non-linear, and neural network models on a large and imbalanced real-world dataset, providing insights into model performance, limitations, and suitability for complex classification tasks.

11 Overview

This project focuses on predicting forest cover types using the Forest Cover Type dataset, a large and widely used benchmark in environmental and ecological machine learning research. The dataset contains 581,012 observations and 54 predictive features describing various physical, geological, and cartographic characteristics of forested regions in the Roosevelt National Forest in Colorado. These features include elevation, slope, aspect, hillshade measurements, distances to hydrological features, proximity to roadways and fire points, four wilderness area indicators, and forty soil type variables. The combination of high dimensionality, non-linear relationships, and severe class imbalance makes this dataset particularly challenging for classification tasks.

To address this problem, a complete and structured machine learning workflow was implemented. The major steps of the project included:

- **Data preprocessing**: verification of encoded attributes, scaling of numerical features, and stratified train–test splitting.
- **Exploratory Data Analysis (EDA)**: examination of class distribution, feature correlations, terrain characteristics, and outlier patterns.
- **Supervised learning models**: Logistic Regression, Support Vector Machine (RBF kernel), and a Neural Network (MLPClassifier) were trained to predict the seven cover types.
- **Unsupervised learning models**: K-Means and Gaussian Mixture Models were applied to explore natural cluster patterns within the feature space.
- **Performance evaluation**: accuracy, precision, recall, F1-score, confusion matrices, and the Adjusted Rand Index (ARI) were used to compare model effectiveness.

This study provides a comprehensive comparison of classical machine learning techniques on a real-world dataset. It demonstrates how different models cope with complex environmental data and highlights the advantages of non-linear and deep learning approaches for large-scale multiclass classification tasks.

12 Data Description

The Forest Cover Type dataset consists of 581,012 observations and 55 columns (54 features and 1 target variable). Each instance represents a 30m × 30m patch of forest from the Roosevelt National Forest in Colorado. The dataset includes detailed environmental and geological information.

1. Features

The dataset contains 54 predictive features grouped into the following categories:

A. Topographic Features

- Elevation
- Aspect
- Slope
- Hillshade_9am
- Hillshade_Noon
- Hillshade_3pm

B. Distance-Based Features

- Horizontal_Distance_To_Hydrology
- Vertical_Distance_To_Hydrology
- Horizontal_Distance_To_Roadways
- Horizontal_Distance_To_Fire_Points

C. Wilderness Area Indicators (4 Binary Columns)

- Wilderness Area 1
- Wilderness Area 2
- Wilderness Area 3
- Wilderness Area 4

D. Soil Types (40 Binary Columns)

Binary variables indicating 40 distinct soil categories.

All features are numerical, facilitating preprocessing and model training.

2. Target Variable

The target variable **Cover_Type** consists of seven forest cover categories:

Class	Forest Cover Type
1	Spruce/Fir
2	Lodgepole Pine
3	Ponderosa Pine
4	Cottonwood/Willow
5	Aspen
6	Douglas-fir
7	Krummholz

3. Dataset Characteristics

A. Size and Dimensionality

- 581,012 samples
- 54 predictive features
- High-dimensional and computationally demanding

B. Class Imbalance

- Classes 1 and 2 dominate the dataset
- Classes 4 and 5 have significantly fewer samples

C. Data Quality

- No missing values
- No duplicate rows
- All features are numeric

13 Exploratory Data Analysis (EDA)

This section explores the structure, distribution, and relationships of the dataset's features.

1. Class Distribution

The dataset is highly imbalanced. Figure 13.1 illustrates the distribution of the seven cover types.

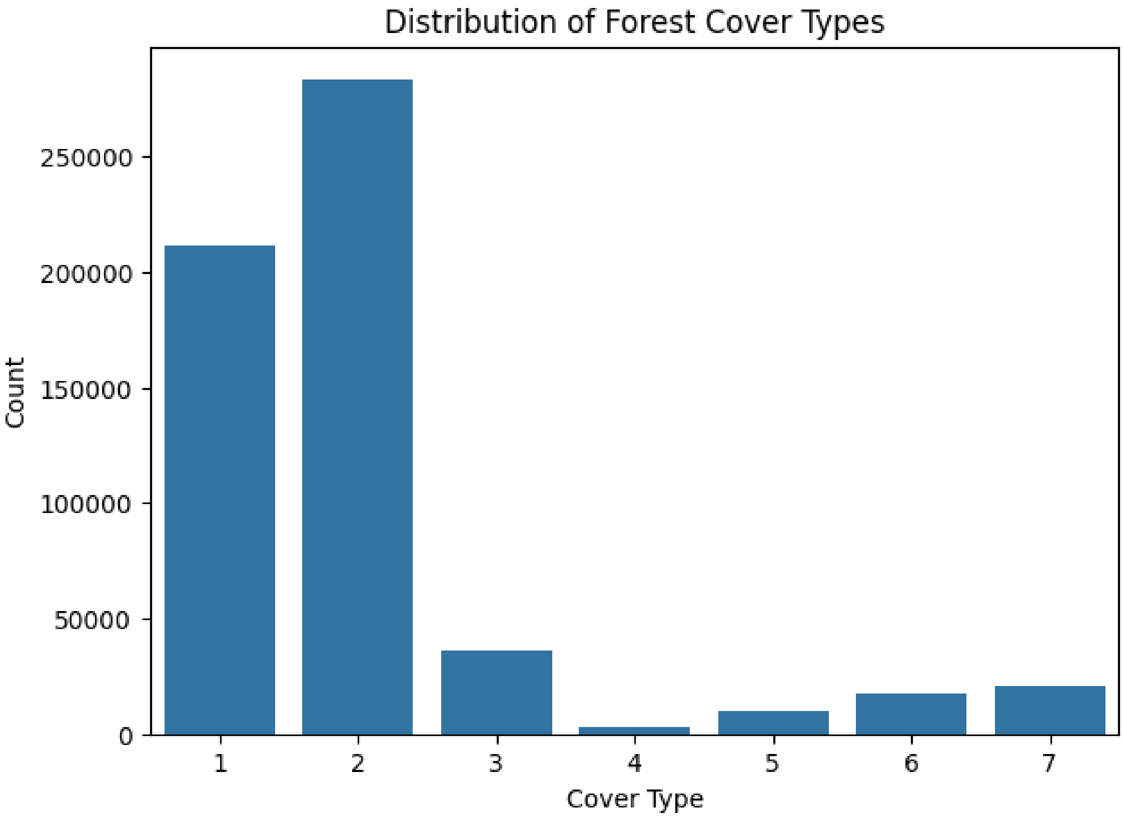


Figure 13.1: Distribution of forest cover types in the dataset.

2. Terrain Feature Relationships

Figures 13.2 depict the distributions of several important topographic variables, including elevation, aspect, slope, and distance-based features. These plots help reveal patterns, skewness, and potential outliers within the dataset.

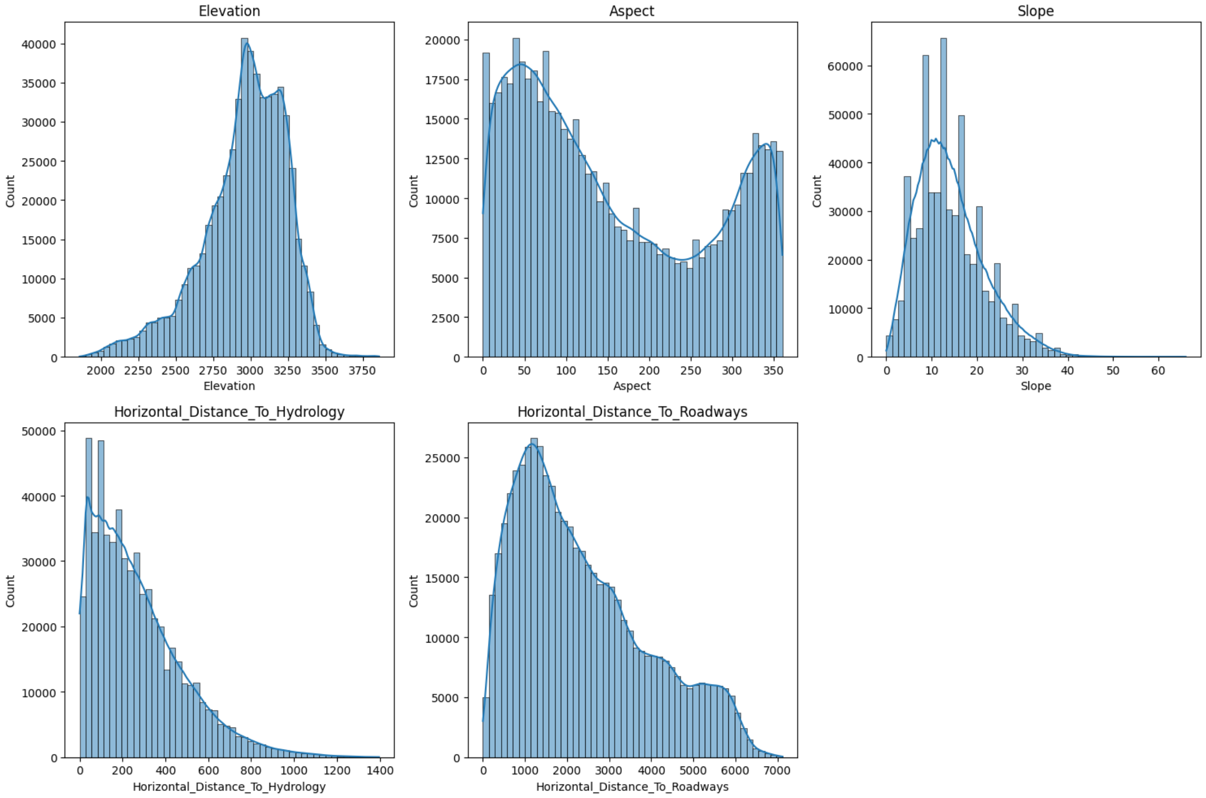


Figure 13.2: Distributions of key topographic features, including elevation, aspect, slope, and major distance-based attributes.

3. Correlation Matrix

The correlation heatmap shown in Figure 13.3 highlights relationships among the primary numerical features in the dataset. Key observations include:

- Strong positive correlations among the three hillshade variables (9am, Noon, 3pm).
- Moderate correlation between elevation and hydrological distances.
- No evidence of severe multicollinearity across the numerical features.

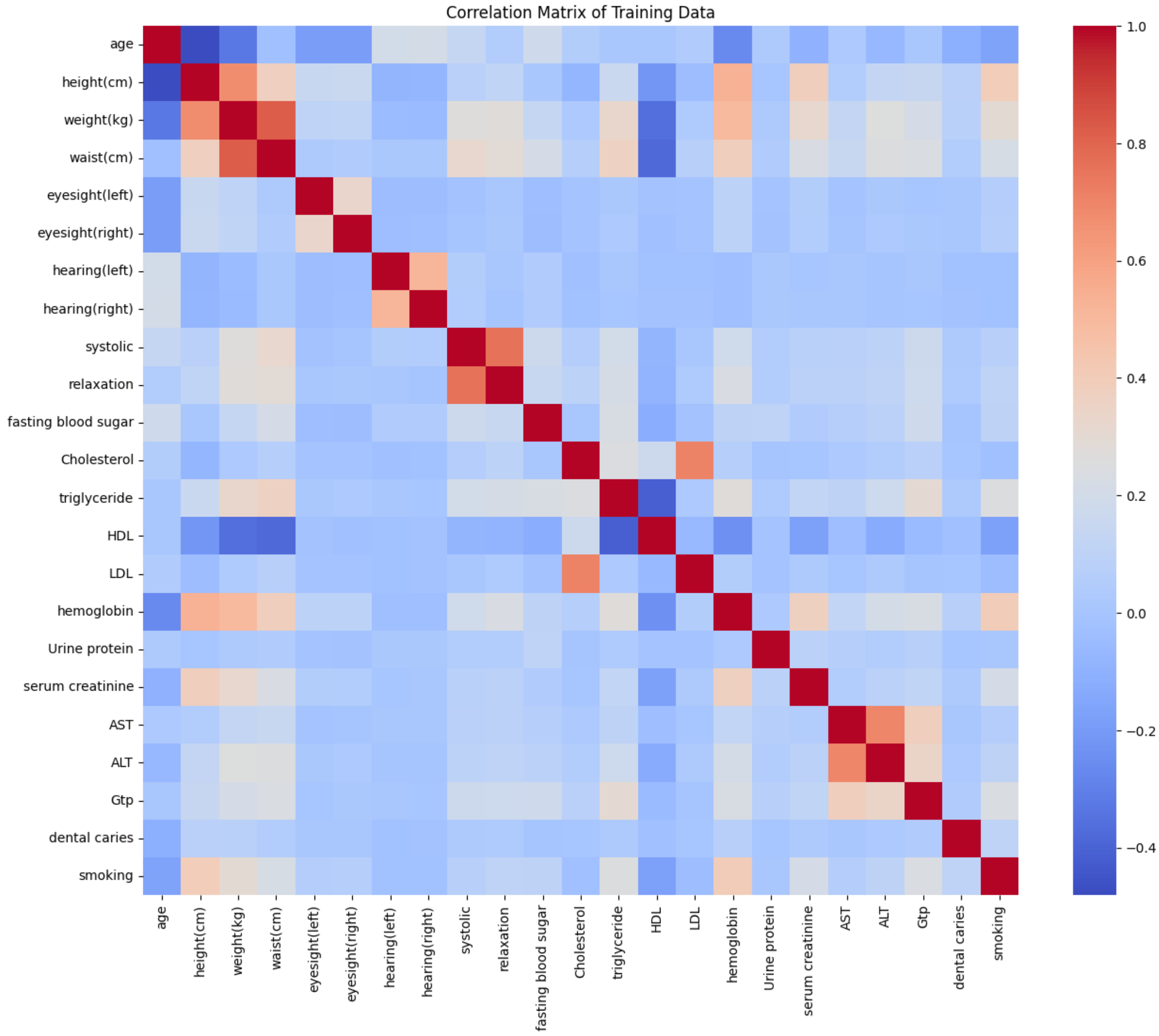


Figure 13.3: Correlation heatmap of key numerical environmental features.

4. Preliminary Feature Importance Insights

Although detailed feature importance is examined later through supervised models, early EDA reveals several trends:

- **Elevation** appears to be one of the most informative features, with values varying distinctly across cover types.
- **Hillshade values** capture terrain orientation and illumination, contributing moderate predictive power.
- **Hydrology-related distances** exhibit class-dependent patterns that may help differentiate vegetation types.
- **Soil types** (40 binary attributes) encapsulate critical ecological information and play a major role in classification.

14 Preprocessing

Effective preprocessing was essential for preparing the Forest Cover Type dataset for machine learning. Since the dataset contains only numerical features and no missing values, the focus was on verifying encoded variables, scaling features, and generating a balanced train–test split.

1. One-Hot Encoding Verification

The dataset already provided categorical variables for *Wilderness Areas* and *Soil Types* in one-hot encoded form:

- 4 binary columns for wilderness areas
- 40 binary columns for soil types

Because these variables were pre-encoded, no additional encoding was required.

2. Feature Scaling

Numerical features such as elevation and distance measures varied widely in scale. To ensure fair contribution across models, all continuous features were standardized using `StandardScaler`, transforming each variable to:

- Mean = 0
- Standard deviation = 1

Standardization was particularly important for algorithms like SVM, Logistic Regression, Neural Networks, K-Means, and GMM, which are sensitive to feature magnitude.

3. Train–Test Split

A stratified 80–20 split was used to preserve the original class distribution:

- Training set: 464,809 samples
- Testing set: 116,203 samples

Stratification ensured that minority classes were proportionally represented, preventing bias in model training and enabling fair evaluation.

Overall, these preprocessing steps resulted in a clean, well-structured dataset suitable for both supervised and unsupervised models used in this study.

15 Methodology

This chapter outlines the complete workflow followed in building, training, and evaluating the machine learning models for forest cover type classification. The methodology is structured to reflect a standard and reproducible ML pipeline.

1. Workflow Summary

The overall process consisted of the following steps:

1. **Dataset Loading and Inspection**: Understanding feature types, dimensions, and basic structure.
2. **Preprocessing and Scaling**: Verifying one-hot encoded variables, applying StandardScaler, and performing a stratified train–test split.
3. **Exploratory Data Analysis (EDA)**: Visualizing class imbalance, feature distributions, and correlations.
4. **Training of Supervised Models**: Implementing Logistic Regression, SVM, and a Neural Network (MLPClassifier).
5. **Training of Unsupervised Models**: Applying K-Means and Gaussian Mixture Models to detect underlying cluster patterns.
6. **Model Evaluation and Comparison**: Using accuracy, classification reports, confusion matrices, and Adjusted Rand Index (ARI).

2. Supervised Models

Three supervised learning algorithms were trained to classify the seven forest cover types.

2.1 Logistic Regression

Multinomial Logistic Regression was used with the `lbfgs` solver. This model served as a baseline due to its simplicity and interpretability.

2.2 Support Vector Machine (RBF Kernel)

A non-linear SVM with an RBF kernel was implemented. Because of the large dataset size, training was performed on a 20,000-sample representative subset, while evaluation used the full test set.

2.3 Neural Network (MLPClassifier)

A multi-layer perceptron was trained to capture non-linear relationships:

- Hidden Layer 1: 128 neurons (ReLU)
- Hidden Layer 2: 64 neurons (ReLU)
- Optimizer: Adam

The architecture was selected to balance performance and computational efficiency.

3. Unsupervised Models

Two clustering algorithms were used to explore the structure of the data without labels.

3.1 K-Means Clustering

K-Means was applied with $k = 7$, corresponding to the seven forest cover types. Cluster assignments were later compared to true labels using the Adjusted Rand Index (ARI).

3.2 Gaussian Mixture Model (GMM)

A probabilistic clustering model with seven Gaussian components. Full covariance matrices were used to allow flexible cluster shapes.

4. Hyperparameter Tuning

Basic hyperparameter tuning was performed for each model:

- **Logistic Regression**: Increased maximum iterations for stable convergence.
- **SVM**: Limited training subset for computational feasibility.
- **Neural Network**: Tuned hidden layer sizes, activation functions, and learning rate heuristically.

16 Model Training and Evaluation

This chapter summarizes the performance of the supervised and unsupervised models trained on the Forest Cover Type dataset. Each model was evaluated using appropriate metrics, and the results are compared to highlight differences in predictive ability.

1. Supervised Models

Three supervised learning algorithms were trained on the standardized training dataset and evaluated using the 20% test split.

Logistic Regression

Multinomial Logistic Regression served as the baseline classifier.

- **Accuracy: 0.7234**

The model performed well on majority classes but struggled to separate minority classes due to the linear decision boundaries and class imbalance.

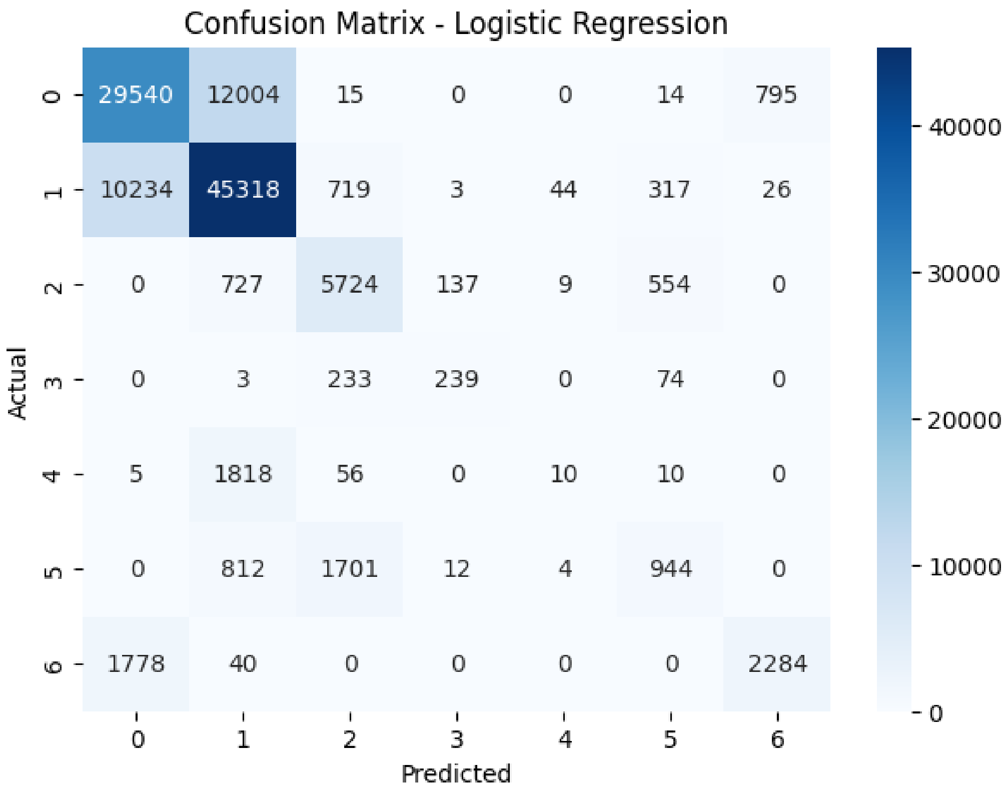


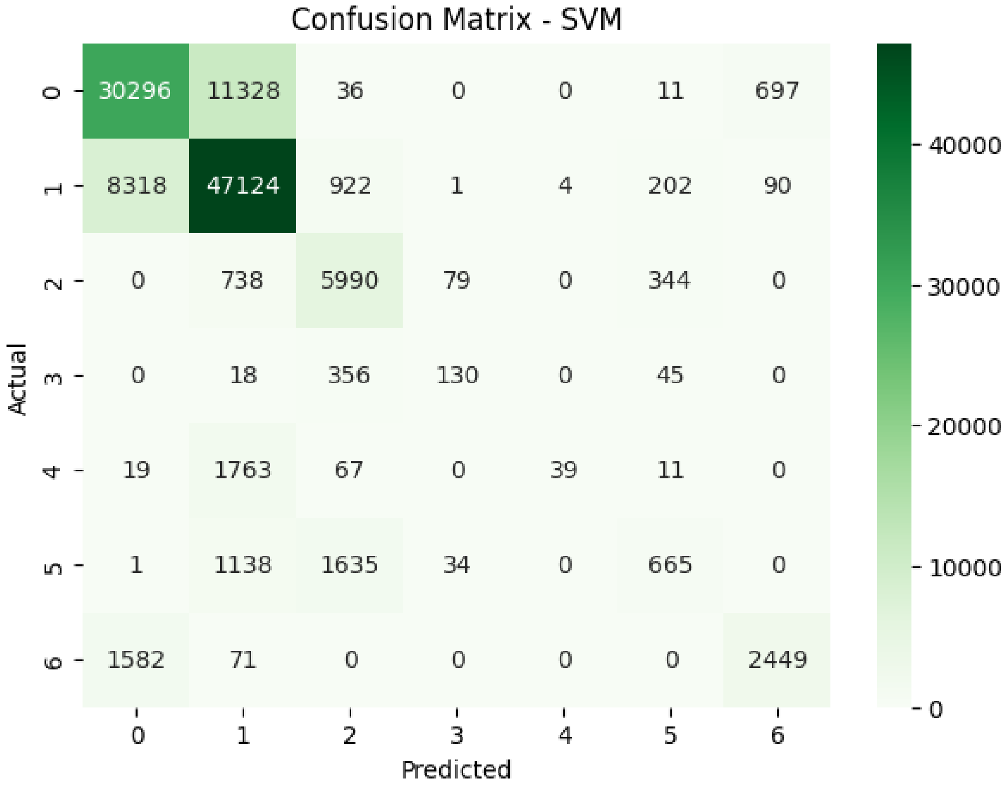
Figure 16.1: Confusion Matrix – Logistic Regression

Support Vector Machine (RBF Kernel)

A non-linear SVM was trained on a representative subset of the data due to computational cost, while evaluation was done on the full test set.

- **Accuracy: 0.7460**

The RBF kernel captured more complex relationships than Logistic Regression, resulting in improved accuracy, though performance on minority classes remained limited.



Actual \ Predicted	0	1	2	3	4	5	6
0	30296	11328	36	0	0	11	697
1	8318	47124	922	1	4	202	90
2	0	738	5990	79	0	344	0
3	0	18	356	130	0	45	0
4	19	1763	67	0	39	11	0
5	1	1138	1635	34	0	665	0
6	1582	71	0	0	0	0	2449

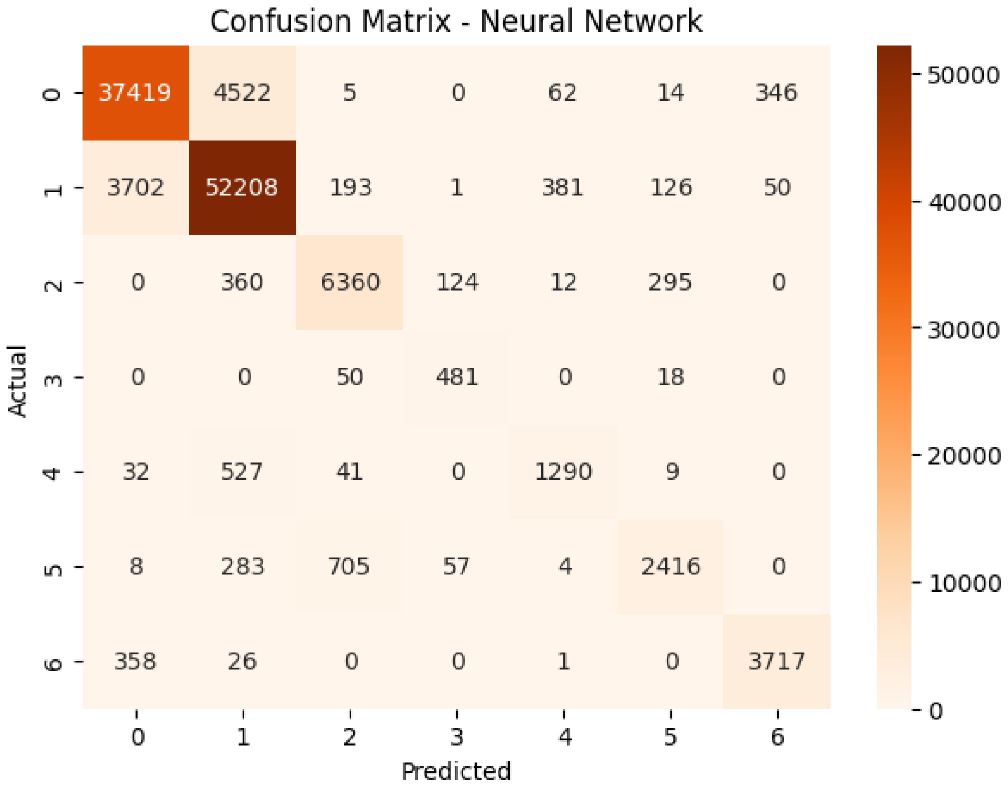
Figure 16.2: Confusion Matrix – SVM (RBF Kernel)

Neural Network (MLPClassifier)

A multi-layer perceptron model with ReLU activation was trained to capture complex non-linear patterns in the dataset.

- **Accuracy: 0.8940**

The neural network achieved the highest accuracy among all supervised models, showing strong generalization and robust performance even in the presence of class imbalance.



Actual \ Predicted	0	1	2	3	4	5	6
0	37419	4522	5	0	62	14	346
1	3702	52208	193	1	381	126	50
2	0	360	6360	124	12	295	0
3	0	0	50	481	0	18	0
4	32	527	41	0	1290	9	0
5	8	283	705	57	4	2416	0
6	358	26	0	0	1	0	3717

Figure 16.3: Confusion Matrix – Neural Network (MLPClassifier)

2. Unsupervised Models

Two clustering algorithms were trained without using class labels. Their outputs were evaluated using the Adjusted Rand Index (ARI), which measures similarity between true labels and cluster assignments.

K-Means Clustering

- **ARI: 0.0941**

The low ARI score indicates that the natural clusters formed by K-Means do not align closely with the actual forest cover types.

Gaussian Mixture Model (GMM)

- **ARI: 0.0836**

GMM performed similarly to K-Means, suggesting that the dataset does not exhibit strong cluster boundaries corresponding to cover type classes.

3. Summary of Results

Table 16.1 summarizes the performance of all models. Supervised learning approaches significantly outperform unsupervised clustering methods, with the neural network clearly demonstrating the strongest predictive performance.

Model	**Metric**
Logistic Regression	Accuracy = 0.7234
SVM (RBF Kernel)	Accuracy = 0.7460
Neural Network (MLP)	Accuracy = 0.8940
K-Means Clustering	ARI = 0.0941
GMM Clustering	ARI = 0.0836

Table 16.1: Performance comparison of supervised and unsupervised models.

17 Conclusion

This project examined supervised and unsupervised machine learning models for predicting forest cover types using a large, high-dimensional, and imbalanced dataset. The complete pipeline— including preprocessing, exploratory analysis, model training, and evaluation—provided a clear understanding of how different algorithms behave on complex environmental data.

Supervised models showed notable differences:

- **Logistic Regression** offered a reasonable baseline but struggled with minority classes.
- The **SVM (RBF Kernel)** improved accuracy by capturing non-linear patterns, though it required reduced training data due to computational cost.
- The **Neural Network (MLPClassifier)** achieved the highest accuracy (0.8940), effectively modeling complex feature interactions.

Unsupervised models were less effective:

- **K-Means** achieved an ARI of 0.0941, showing weak alignment with true labels.
- **Gaussian Mixture Models** produced a similar ARI (0.0836), indicating overlapping and non-separable clusters.

Overall, these results demonstrate that supervised learning—particularly neural networks—is best suited for this classification task. The dataset's non-linearity and imbalance require models capable of capturing complex relationships.

Project Repository

GitHub Links

- Abhijeet Kumar Gupta :- https://github.com/Abhijeetgupta27/Machine-Learning-Approaches-to-Binary-and-Multiclass-Classification-Problems
- Abhash Tiwari :- https://github.com/AbhashTi/smoking$_binary_classification$