AIT511: Machine Learning

Group Project Report

Predicting Smoker Status and Forest Cover Types using Machine Learning

Group Members:

Jenish Niteshbhai Vekariya (MT2025055)

Bhautik Pravinbhai Vekariya (MT2025029)

Department of Computer Science
IIITB

Contents

Chapter 1

Introduction

1.1 Background

Predictive modeling plays a pivotal role across diverse domains, ranging from public health to environmental resource management. This study analyzes two distinct datasets to address classification challenges in these fields. The first focuses on the healthcare sector, specifically examining the correlation between physiological bio-signals and smoking habits. Identifying smokers through bio-signals is crucial for preventive healthcare and insurance underwriting.

The second area of focus is environmental science, specifically the forestry sector. Accurately cataloging forest cover types using cartographic data is essential for ecosystem monitoring and wildfire resource planning. Both tasks present unique challenges in feature engineering and multi-class versus binary classification.

1.2 Problem Statement

This project addresses two separate classification problems. The first is a binary classification task aimed at predicting an individual's smoking status. The target variable is *smoking*, where the model must distinguish between smokers and non-smokers based on health metrics.

The second task is a multi-class classification problem designed to predict the forest cover type for a given 30x30 meter cell. The target variable *Cover_Type* consists of seven distinct classes: *Spruce/Fir*, *Lodgepole Pine*, *Ponderosa Pine*, *Cottonwood/Willow*, *Aspen*, *Douglas-fir*, and *Krummholz*.

1.3 Dataset Description

The first dataset, "Smoker Status Prediction" is sourced from Kaggle. It contains physiological records including basic characteristics such as *age*, *height*, *weight*, and *waist* circumference. It also includes detailed medical metrics like *systolic* and *relaxation* blood pressure, *fasting blood sugar*, *Cholesterol* (total, HDL, and LDL), *triglyceride* levels, *hemoglobin*, and liver enzymes (*AST*, *ALT*, *Gtp*).

The second dataset, "Forest Cover Type," is sourced from the UCI Machine Learning Repository and Kaggle. It represents cartographic variables derived from the Roosevelt National Forest of Northern Colorado. The attributes include *Elevation*, *Aspect*, *Slope*, and distances to key landmarks such as hydrology, roadways, and fire points. Additionally, it features hillshade indices at different times of the day and binary columns representing *Wilderness_Area* and *Soil_Type*.

Chapter 2

Data Overview and Exploratory Data Analysis

2.1 Dataset 1: Smoker Status

2.1.1 Dataset Overview

The first dataset focuses on predicting smoking status based on physiological bio-signals. It consists of 38,984 records with 23 columns, including basic health indicators (age, height, weight) and detailed medical metrics. The analysis begins with a comprehensive examination of the target variable and feature distributions.

2.1.2 Target Variable Analysis

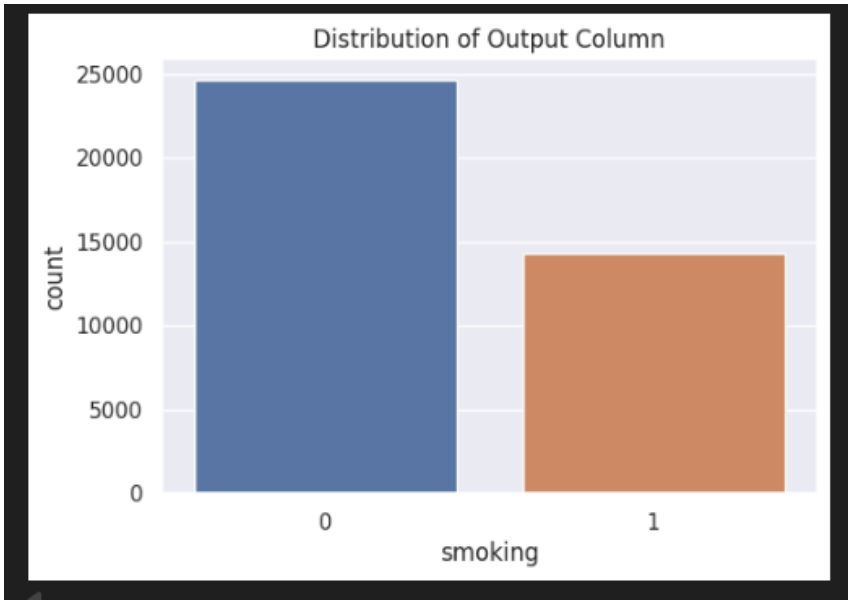

Figure 2.1: Distribution of Smoking Status

The target variable, *smoking*, represents a binary classification task. The analysis reveals a class imbalance:

- **Non-smokers (0):** 63.27% (24,666 individuals)
- **Smokers (1):** 36.73% (14,318 individuals)

While imbalanced, the minority class is sufficiently represented to train effective machine learning models without immediate need for oversampling.

2.1.3 Numerical Features Analysis

The dataset contains numerous numerical features. Histograms were plotted to analyze their distributions.

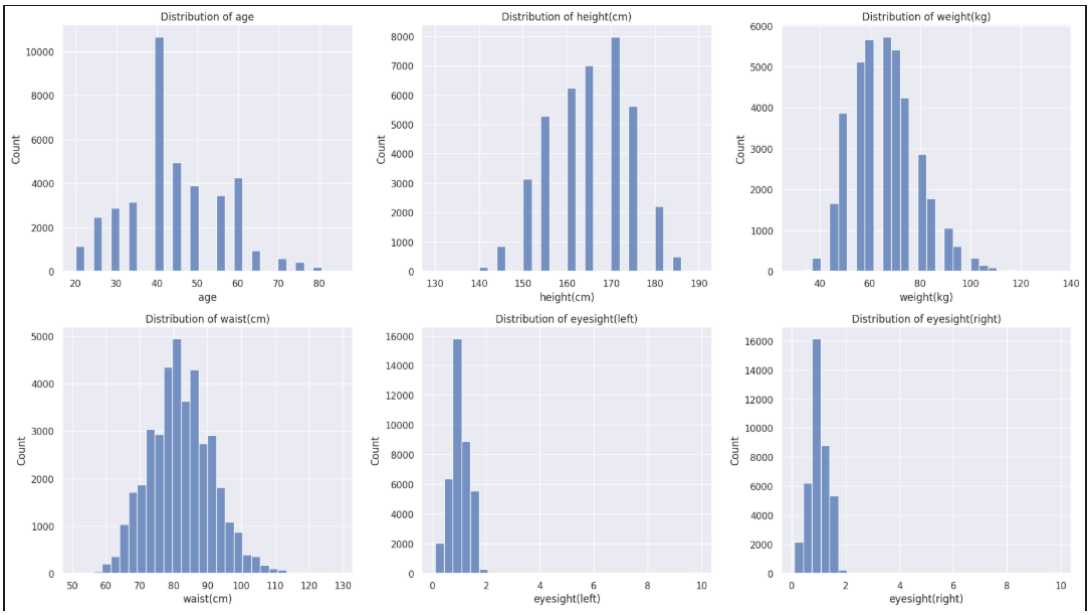

Figure 2.2: Distribution of Numerical Features (Part 1)

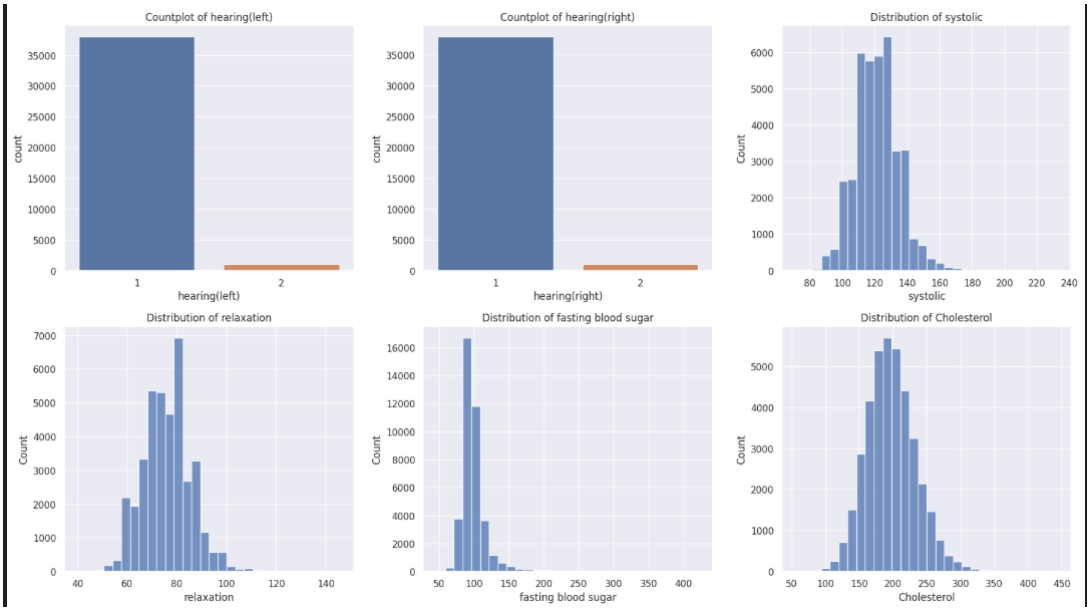

Figure 2.3: Distribution of Numerical Features (Part 2)

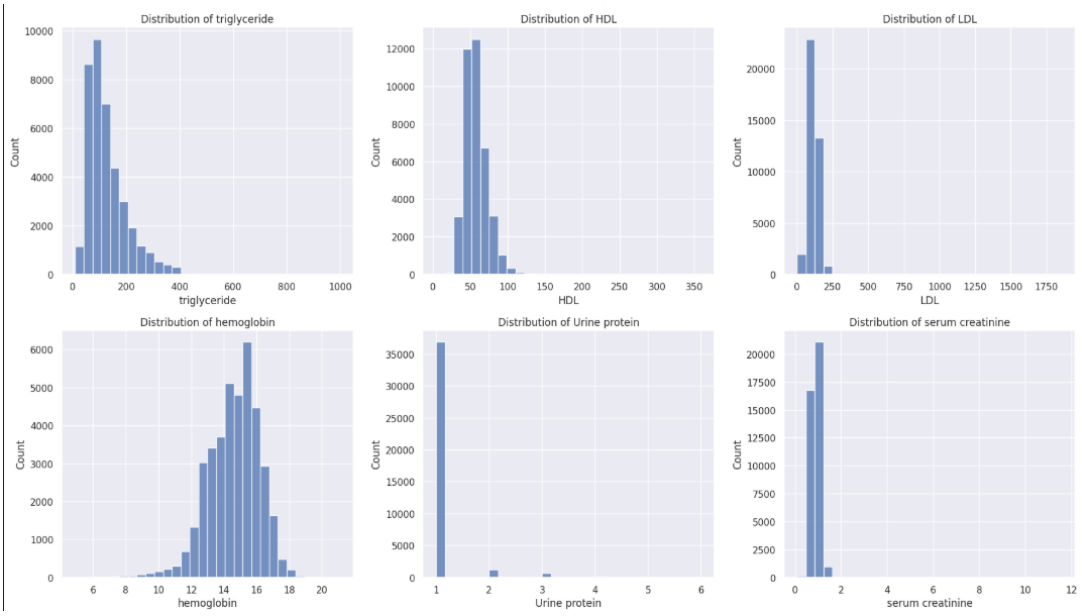

Figure 2.4: Distribution of Numerical Features (Part 3)

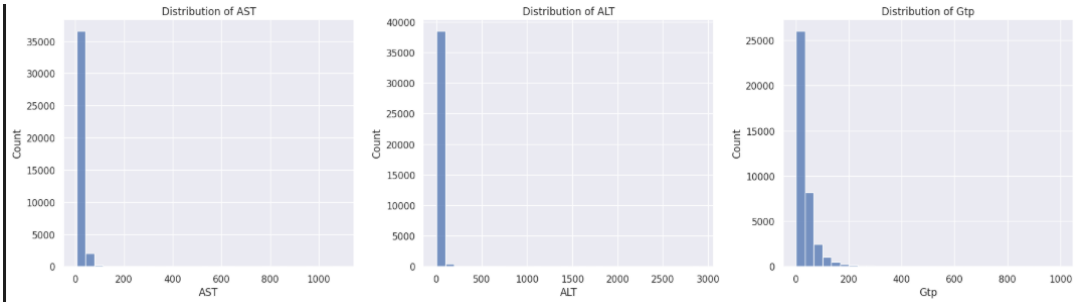

Figure 2.5: Distribution of Numerical Features (Part 4)

2.1.4 Correlation Analysis

To identify relationships between physiological features and potential multicollinearity, a correlation heatmap was generated.

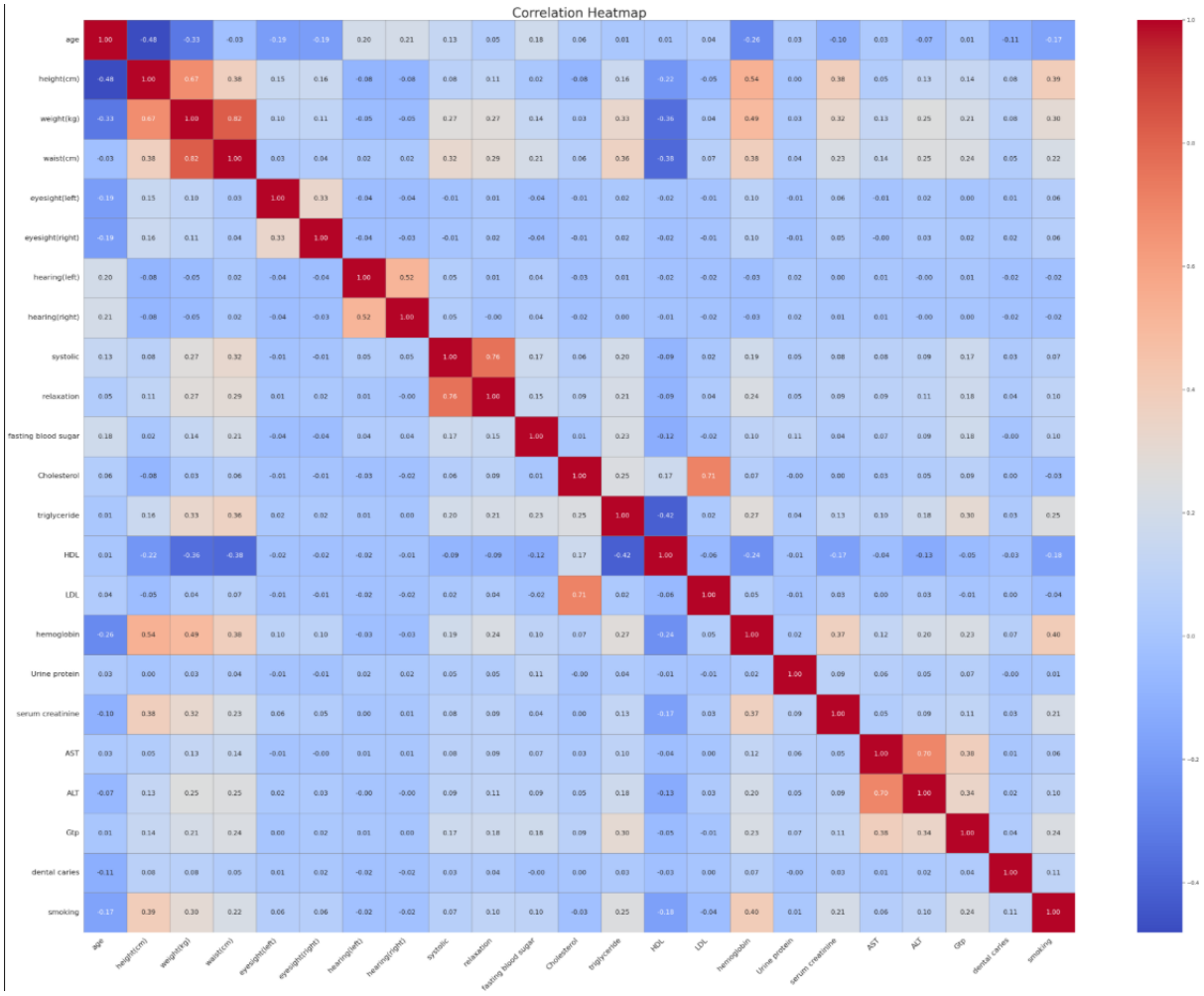

Figure 2.6: Correlation Heatmap

The analysis highlights expected physiological correlations, such as the relationship between *Systolic* and *Relaxation* blood pressure, as well as *Weight* and *Waist* circumference. These strong linear relationships suggest that feature selection or dimensionality reduction could benefit the model.

2.1.5 Outlier Detection

Boxplots were generated to identify extreme values in data.

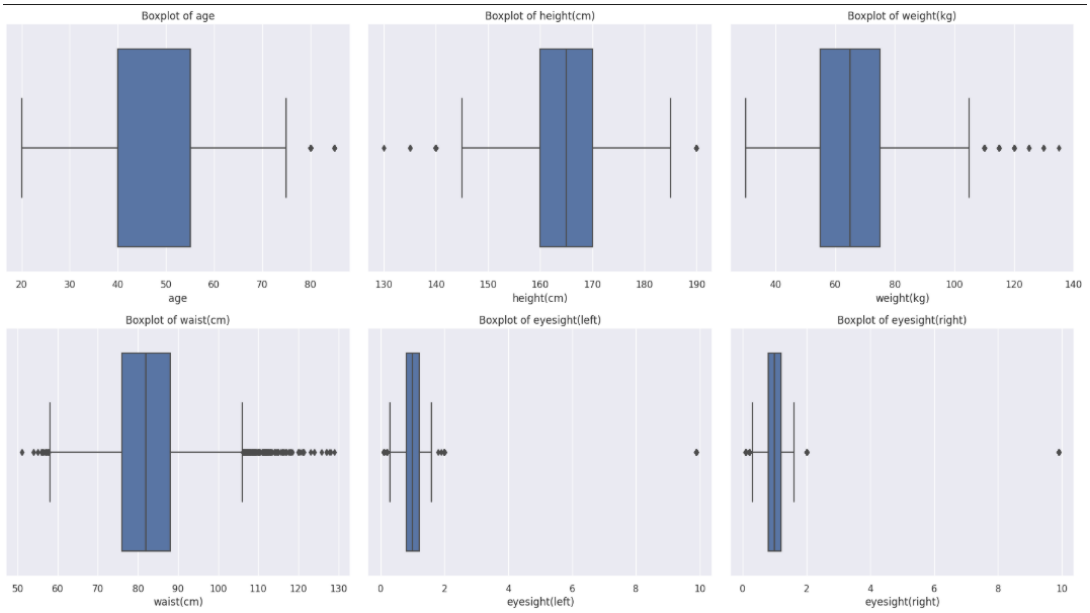

Figure 2.7: Outlier Detection (Part 1)

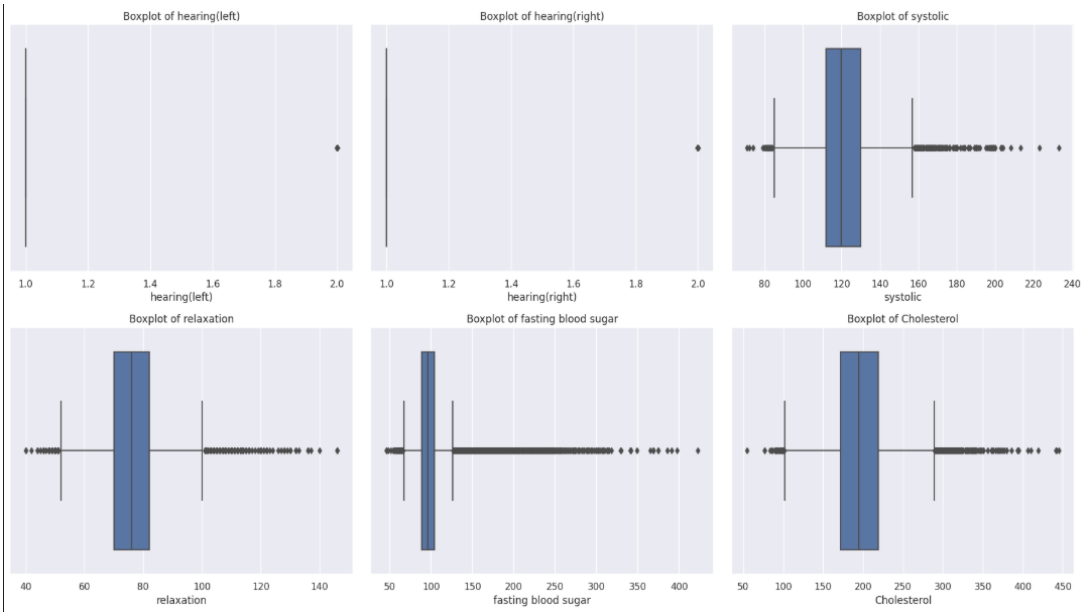

Figure 2.8: Outlier Detection (Part 2)

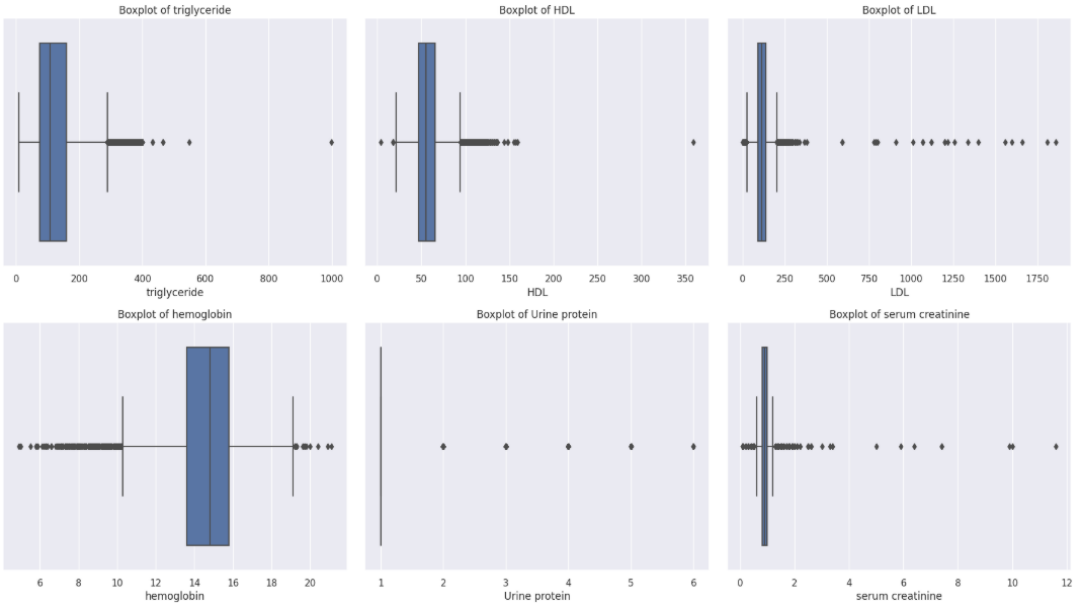

Figure 2.9: Outlier Detection (Part 3)

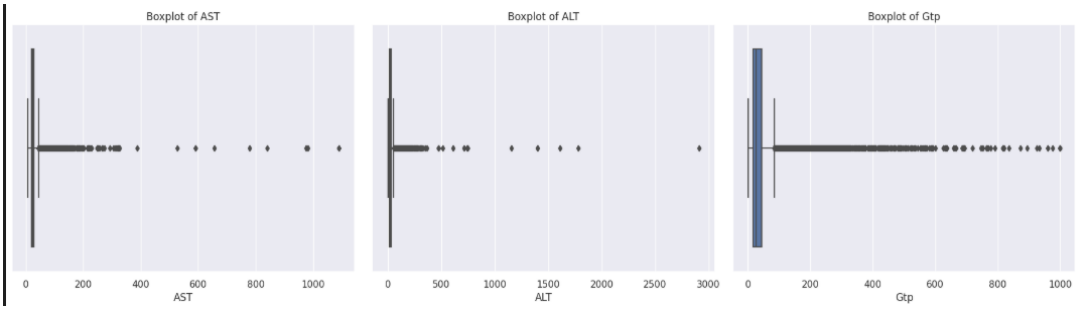

Figure 2.10: Outlier Detection (Part 4)

2.2 Dataset 2: Forest Cover Type

2.2.1 Dataset Overview

The second dataset involves predicting forest cover types from cartographic variables. It is significantly larger, containing 581,012 entries. The dataset includes numerical topographical features and binary categorical features for wilderness areas and soil types.

2.2.2 Target Variable Analysis

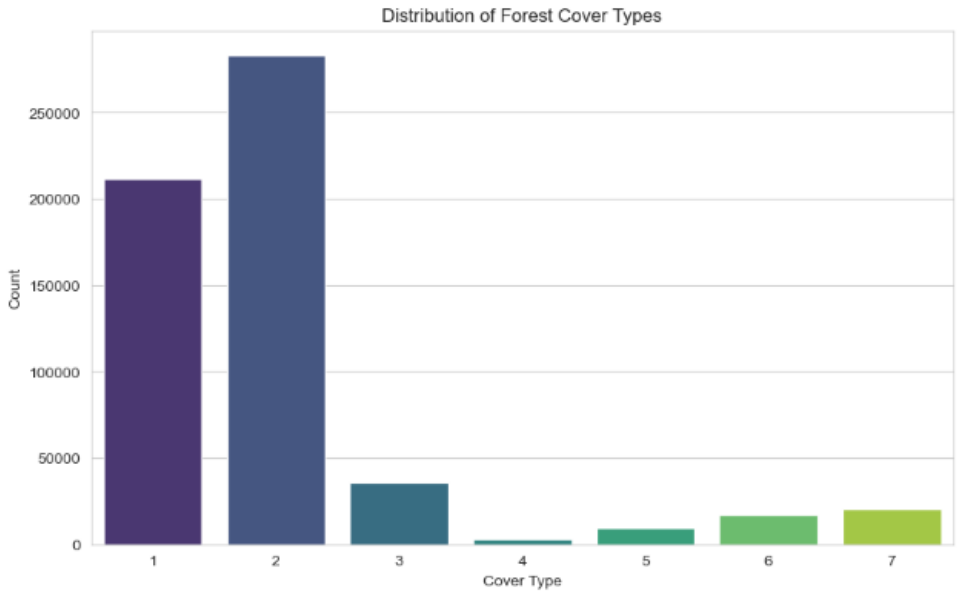

Figure 2.11: Distribution of Forest Cover Types

The target variable, *Cover_Type*, exhibits a severe class imbalance across its 7 classes:

- **Type 2 (Lodgepole Pine):** 48.76%
- **Type 1 (Spruce/Fir):** 36.46%
- **Minority Classes:** Types 3, 4, 5, 6, and 7 combined make up less than 15% of the data, with Type 4 being the rarest (0.47%).

2.2.3 Numerical Features Analysis

Histograms were plotted for the continuous topographical variables (Elevation, Aspect, Slope, etc.) to understand their distributions.

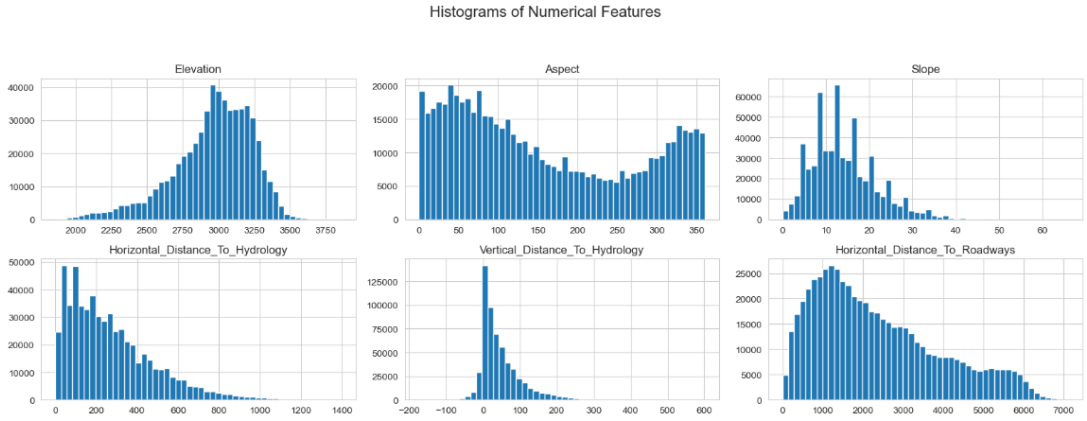

Figure 2.12: Histograms of Numerical Features (Part 1)

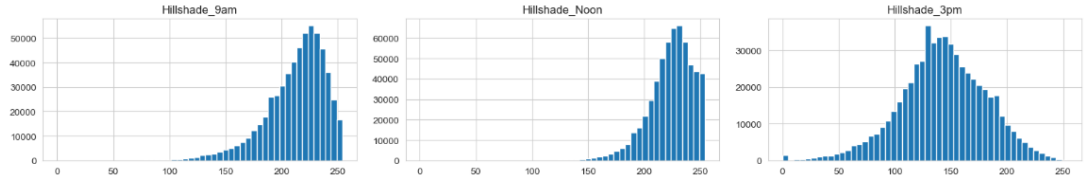

Figure 2.13: Histograms of Numerical Features (Part 2)

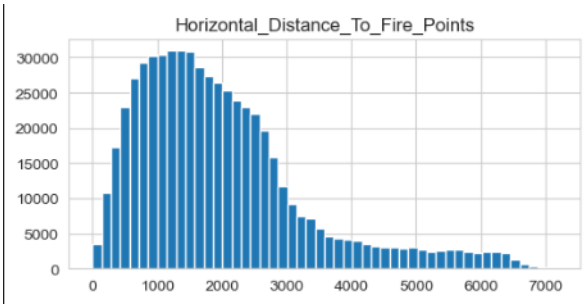

Figure 2.14: Histograms of Numerical Features (Part 3)

2.2.4 Correlation Analysis

(Placed here to maintain consistent rank with Dataset 1)

A correlation matrix was computed for the numerical features.

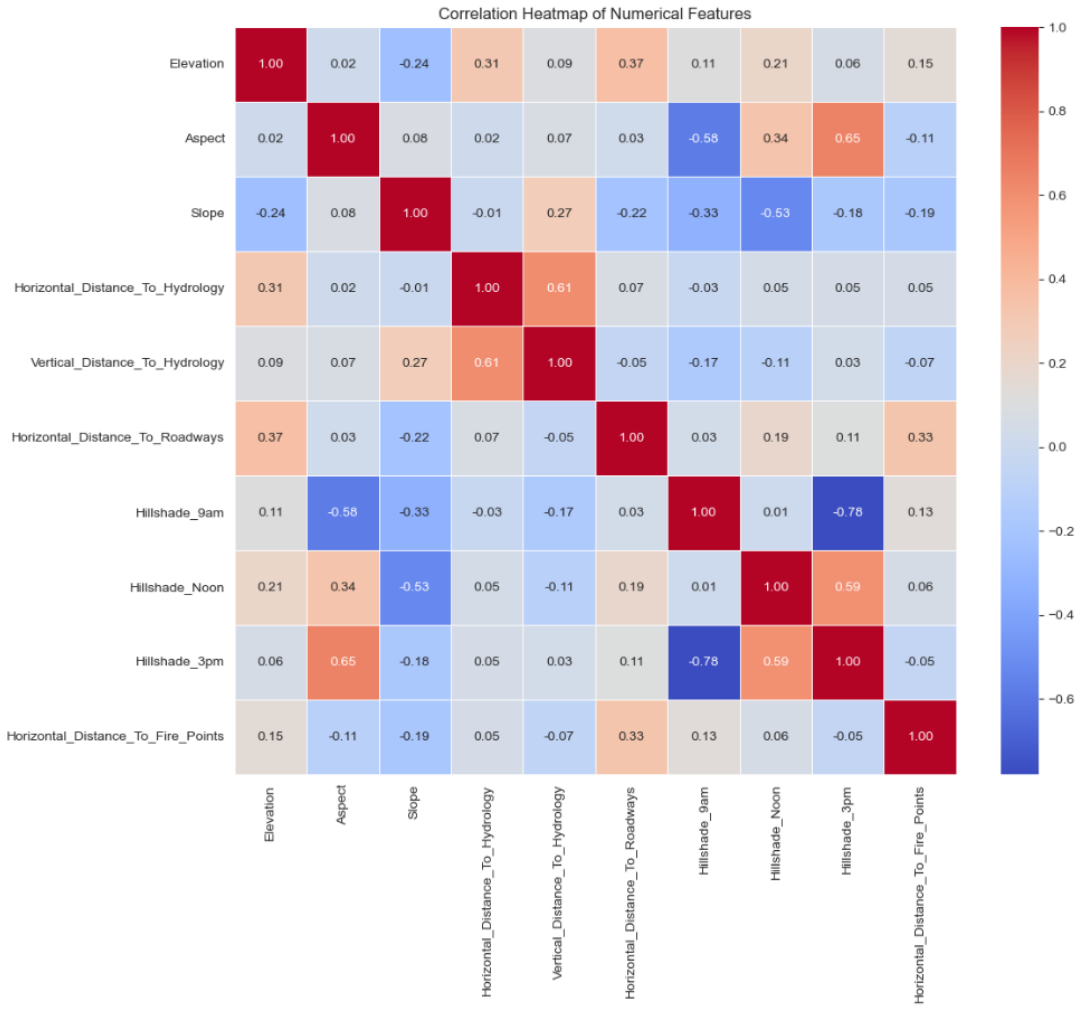

Figure 2.15: Correlation Heatmap of Numerical Features

The analysis revealed distinct characteristics compared to the first dataset. Notably, no feature pairs exhibited a correlation coefficient greater than 0.8, indicating a lower degree of multicollinearity among the topographical variables.

2.2.5 Outlier Detection

Boxplots were used to visualize the spread and detect outliers in the numerical features.

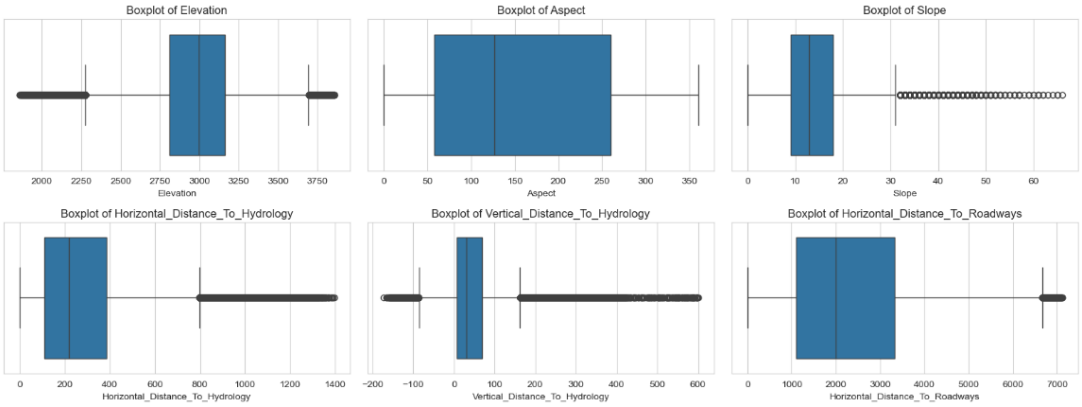

Figure 2.16: Boxplots of Numerical Features (Part 1)

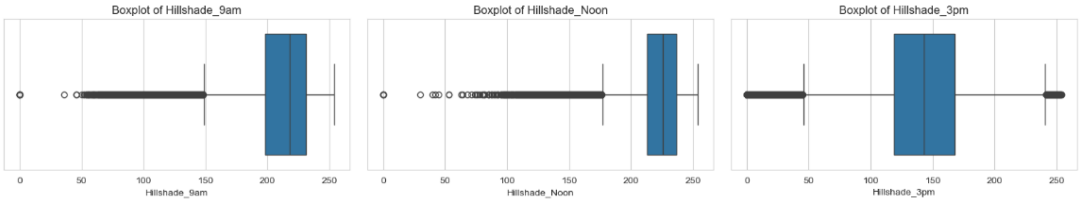

Figure 2.17: Boxplots of Numerical Features (Part 2)

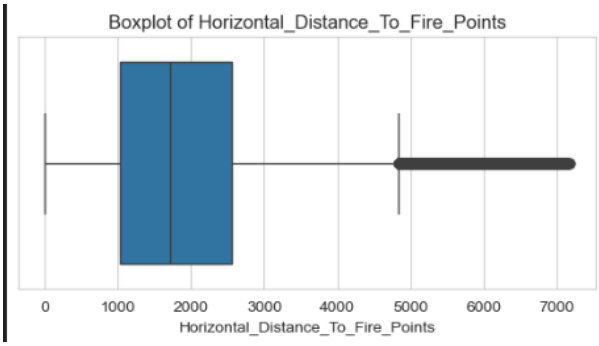

Figure 2.18: Boxplots of Numerical Features (Part 3)

An outlier removal process using Modified Z-Score was applied, which identified and removed approximately 63,090 rows (roughly 10% of the data) to improve data quality.

2.2.6 Categorical Features Analysis

This dataset includes specific binary categorical features representing Wilderness Areas and Soil Types.

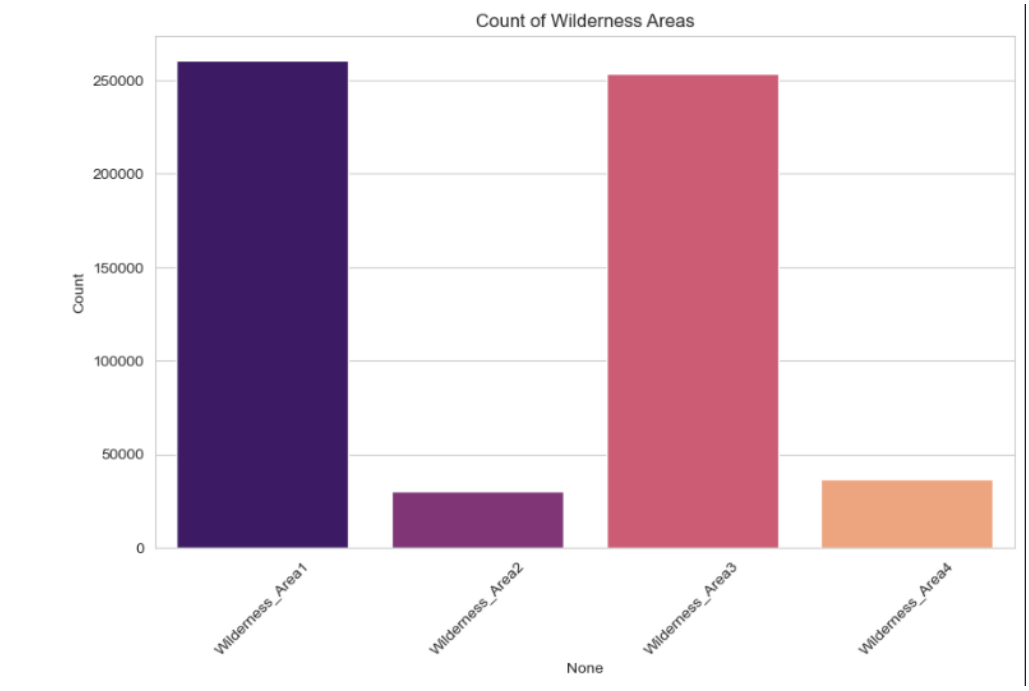

Figure 2.19: Count of Wilderness Areas

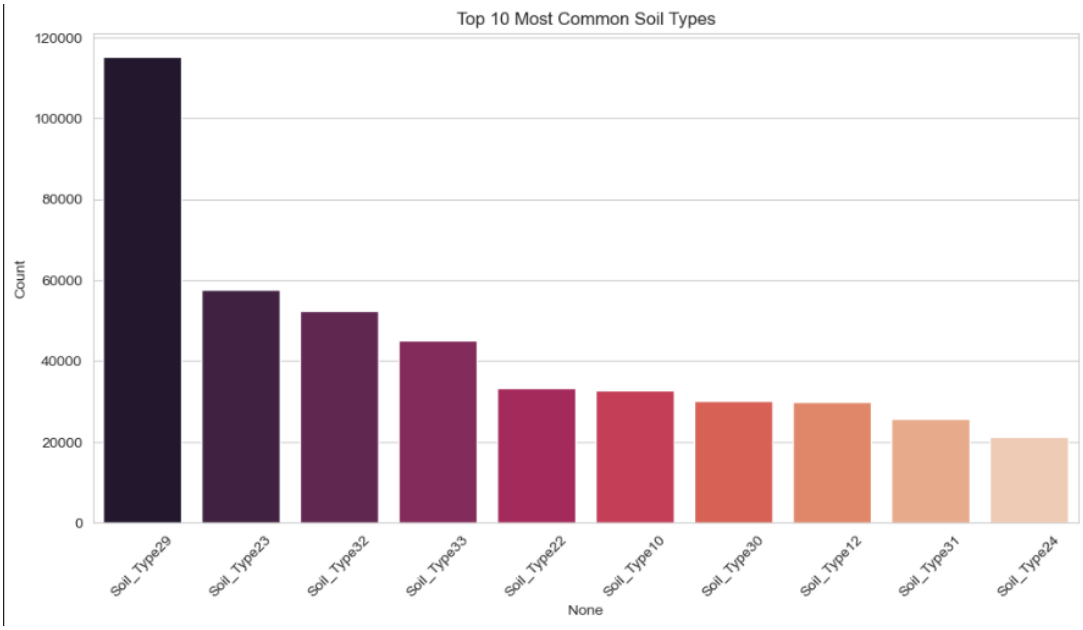

Figure 2.20: Top 10 Most Common Soil Types

Chapter 3

Data Preprocessing and Feature Engineering

3.1 Overview

Data preprocessing is a critical step in the machine learning pipeline, essential for ensuring data quality and model performance. This chapter outlines the distinct preprocessing strategies applied to the Smoker Status Prediction dataset (Dataset 1) and the Forest Cover Type dataset (Dataset 2).

3.2 Dataset 1: Smoker Status Prediction

3.2.1 Data Quality Assessment

A comprehensive check for missing values was conducted on the smoker status dataset. The analysis confirmed that the dataset was complete, with no null values found across its 38984 records, eliminating the need for imputation strategies.

3.2.2 Data Cleaning

To prevent the model from memorizing duplicate patterns, a cleaning step was performed.

- **Duplicate Removal:** A total of 5,517 duplicate entries were detected and removed. This reduced the dataset dimension from 38,984 to 33,467 records, ensuring that each data point represents a unique patient record.

3.2.3 Feature Transformation

The distributions of physiological features often exhibit skewness. To conform the data more closely to a Gaussian distribution, the **Yeo-Johnson transformation** was applied.

- **Methodology:** Yeo-Johnson method is used for feature transformation.
- **Application:** This transformation was applied to numerical columns to stabilize variance and minimize skewness, thereby improving the predictive power of linear and neural network models.

3.2.4 Feature Scaling

Following the skewness correction, **Robust Scaling** was selected to normalize the feature range.

- **Technique:** The *RobustScaler* scales features using robust statistics (the median and the interquartile range) rather than the mean and standard deviation.
- **Benefit:** This approach effectively mitigates the influence of any remaining outliers in the physiological data.

3.3 Dataset 2: Forest Cover Type Prediction

3.3.1 Data Quality Assessment

A comprehensive check for missing values was conducted on the Forest Cover dataset. The analysis confirmed that the dataset was complete, with no null values found across its 581,012 records, eliminating the need for imputation strategies.

3.3.2 Outlier Detection and Removal

Because map data varies so much, standard outlier removal might delete real, extreme values. So, we used a stronger statistical method instead.

- **Method: Modified Z-Score:** This method uses the Median Absolute Deviation (MAD) to detect outliers. It is defined as:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \tag{3.1}$$

  where $\tilde{x}$ denotes the median.

- **Thresholding:** Data points possessing a Modified Z-Score greater than 3.5 were classified as outliers.
- **Result:** This cleaning process removed approximately 63,090 data points, resulting in a cleaner dataset (517,922 records) that better represents general forest cover patterns without the noise of extreme anomalies.

3.3.3 Feature Scaling

Similar to the first dataset, **Robust Scaling** was applied to the numerical topographical features (e.g., Elevation, Aspect, Slope).

- **Implementation:** The *RobustScaler* was fit to the numerical columns, transforming them to a common scale based on their percentiles. This ensures that features with large ranges (like Elevation) do not dominate distance-based algorithms compared to features with smaller ranges (like Slope).

Chapter 4

Methodology

This chapter details the experimental framework employed to address the classification tasks for both the Smoking Status dataset (Dataset 1) and the Forest Cover Type dataset (Dataset 2). While the core algorithmic selection remains consistent to allow for comparative analysis, the training and optimization strategies were tailored to the specific characteristics of each dataset.

4.1 Model Selection

For both classification tasks, we selected three distinct families of algorithms to evaluate performance across linear and non-linear decision boundaries:

- **Logistic Regression (LR):** Selected as a robust linear baseline. It provides interpretability and computational efficiency, particularly valuable for high-dimensional data.
- **Support Vector Machine:** Chosen for its effectiveness in high-dimensional spaces. We utilized the `LinearSVC` implementation which scales better than kernel-based SVMs for larger datasets.
- **Neural Network:** A artificial neural network selected to capture complex, non-linear feature interactions that linear models might miss.

4.2 Dataset 1: Smoker Status Prediction

The first task involves binary classification to predict smoking status. For this dataset, we employed an **80-20 train-test split** to ensure the class distribution remained consistent between training and testing sets.

4.2.1 Hyperparameter Optimization Configuration

We utilized the Optuna framework with Bayesian optimization to tune hyperparameters over 100 trials. The optimization was guided by the **Accuracy** score using Stratified Cross-Validation.

Logistic Regression

The optimization focused on the regularization strength (C) and iteration limits.

Table 4.1: Hyperparameter Search Space: Logistic Regression

Hyperparameter	**Search Range**	**Distribution**
C (Inverse Reg.)	10^{-3} to 10	Log-uniform
max_iter	100 to 500	Integer

SVM

We utilized the `LinearSVC` implementation, optimizing the margin hardness and loss function.

Table 4.2: Hyperparameter Search Space: Linear SVC

Hyperparameter	**Search Range**	**Distribution**
C	10^{-4} to 100	Log-uniform
loss	`hinge`, `squared_hinge`	Categorical
tol	10^{-5} to 10^{-2}	Log-uniform

Neural Network

A dynamic architecture search was performed to determine the optimal network depth and width.

Table 4.3: Hyperparameter Search Space: MLP

Hyperparameter	**Search Range**	**Distribution**
n_layers (Depth)	1 to 4	Integer
neurons_per_layer	32 to 256	Integer
activation	`relu`, `tanh`	Categorical
alpha (L2 Penalty)	10^{-6} to 10^{-2}	Log-uniform
learning_rate_init	10^{-4} to 10^{-2}	Log-uniform

4.3 Dataset 2: Forest Cover Type Prediction

The second task involves multi-class classification on a significantly larger dataset ($> 500,000$ instances). We utilized a **70-30 stratified split** to create a substantial hold-out set. To accelerate tuning, optimization was performed on a stratified subsample of 100,000 instances, while final models were retrained on the full training set.

4.3.1 Hyperparameter Optimization Configuration

The optimization objective for this multi-class problem was **Accuracy**.

Logistic Regression

Configured with the `lbfgs` solver and `multinomial` multi-class option.

Table 4.4: Hyperparameter Search Space: Logistic Regression

Hyperparameter	Search Range	Distribution
C	10^{-3} to 10	Log-uniform
max_iter	100 to 500	Integer

SVM

For computational efficiency on the large dataset, we forced the `dual=True` parameter.

Table 4.5: Hyperparameter Search Space: Linear SVC

Hyperparameter	Search Range	Distribution
C	10^{-3} to 50	Log-uniform
loss	`hinge`, `squared_hinge`	Categorical
intercept_scaling	0.5 to 5	Log-uniform
tol	10^{-5} to 10^{-2}	Log-uniform

Neural Network

The neural network search space for Dataset 2.

Table 4.6: Hyperparameter Search Space: MLP

Hyperparameter	Search Range	Distribution
n_layers (Depth)	1 to 4	Integer
neurons_per_layer	32 to 256	Integer
activation	`relu`, `tanh`	Categorical
alpha (L2 Penalty)	10^{-6} to 10^{-2}	Log-uniform
learning_rate_init	10^{-4} to 10^{-2}	Log-uniform

Chapter 5

Experiments and Results

This chapter presents the quantitative results of our predictive models. We analyze the optimal hyperparameters found by the optimization process and evaluate model performance based on Accuracy on the unseen test sets.

5.1 Results for Dataset 1: Smoking Status

5.1.1 Optimal Hyperparameters

The Optuna optimization process over 100 trials yielded the following best configurations for the smoker status prediction task:

Table 5.1: Optimal Hyperparameters

Model	**Best Parameters**
Logistic Regression	$C \approx 0.078$, max_iter = 469
Linear SVM	$C \approx 27.81$, loss = 'hinge', fit_intercept = True, tol $\approx 3.2e^{-4}$
Neural Network	Layers: 3, Neurons: [171, 156, 171], Activation: 'tanh', $\alpha \approx 1.48e^{-3}$, lr_init $\approx 1.20e^{-4}$

5.1.2 Performance Metrics

All three models achieved highly similar performance, with the Neural Network showing a negligible advantage. The linear models (LR and SVM) performed nearly as well as the non-linear MLP, suggesting the decision boundary for smoking status is largely linear in the transformed feature space.

Table 5.2: Test Set Performance

Model	**Accuracy**
Logistic Regression	73.53%
Linear SVM	73.50%
Neural Network	**73.99%**

5.1.3 Confusion Matrix Analysis

For the smoker status dataset, all three models exhibit similar confusion patterns. The matrices reveal a slight bias towards the majority class (non-smokers), which is typical for physiological datasets where the signal-to-noise ratio is low. The False Negative rate (classifying smokers as non-smokers) remains the primary challenge across all architectures.

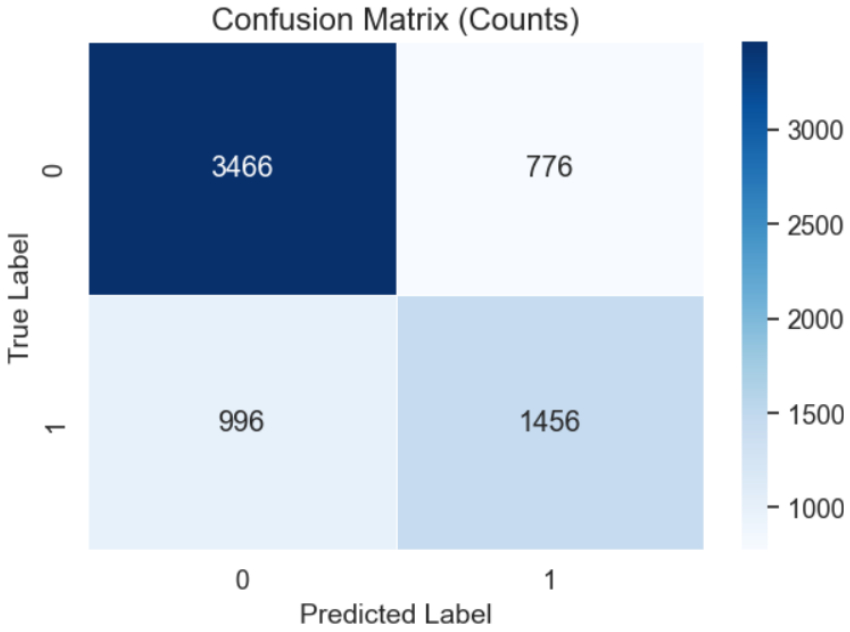

Figure 5.1: Logistic Regression

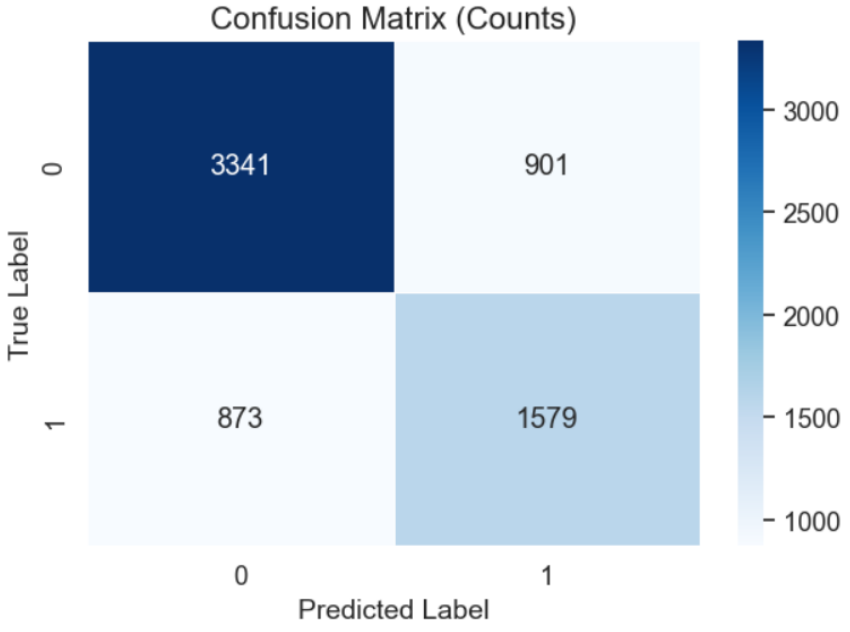

Figure 5.2: Linear SVM

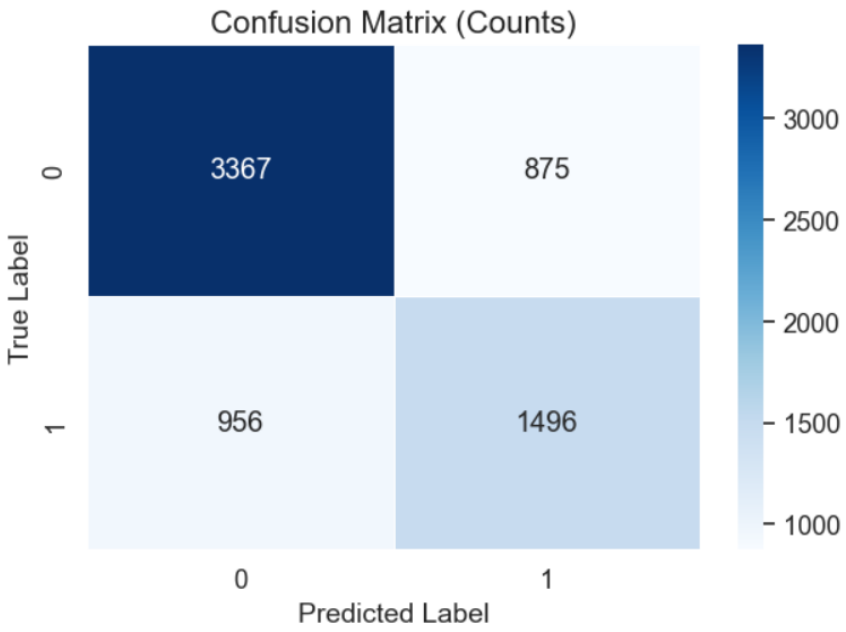

Figure 5.3: Neural Network

5.2 Results for Dataset 2: Forest Cover Type

5.2.1 Optimal Hyperparameters

Table 5.3: Optimal Hyperparameters

Model	**Best Parameters**
Logistic Regression	$C \approx 3.99$, max_iter = 500
Linear SVM	$C \approx 30.46$, loss = 'squared_hinge', fit_intercept = False, tol $\approx 3.78e^{-5}$
Neural Network	Layers: 4, Neurons: [183, 99, 242, 135], Activation: 'tanh', $\alpha \approx 1.28e^{-6}$, lr_init $\approx 1.04e^{-3}$

5.2.2 Performance Metrics

The results highlight a significant performance gap between linear and non-linear models. The MLP achieved near-perfect classification, demonstrating its ability to capture the complex topographical dependencies of forest cover types.

Table 5.4: Test Set Performance

Model	**Accuracy**
Logistic Regression	72.25%
Linear SVM	80.04%
Neural Network	**95.45%**

5.2.3 Confusion Matrix Analysis

The confusion matrices for the Forest Cover dataset highlight the critical difference between the linear and non-linear approaches:

- **Linear Models (LR & SVM):** These models show significant off-diagonal elements, particularly between Class 1 (Spruce/Fir) and Class 2 (Lodgepole Pine). This misclassification suggests that the decision boundary between these cover types is highly non-linear and cannot be separated by a hyperplane.
- **Neural Network (MLP):** The MLP matrix shows a strong, clean diagonal with very few off-diagonal errors. This visualizes the network's superior ability to disentangle the complex feature space, correctly classifying the vast majority of instances across all 7 classes with high precision.

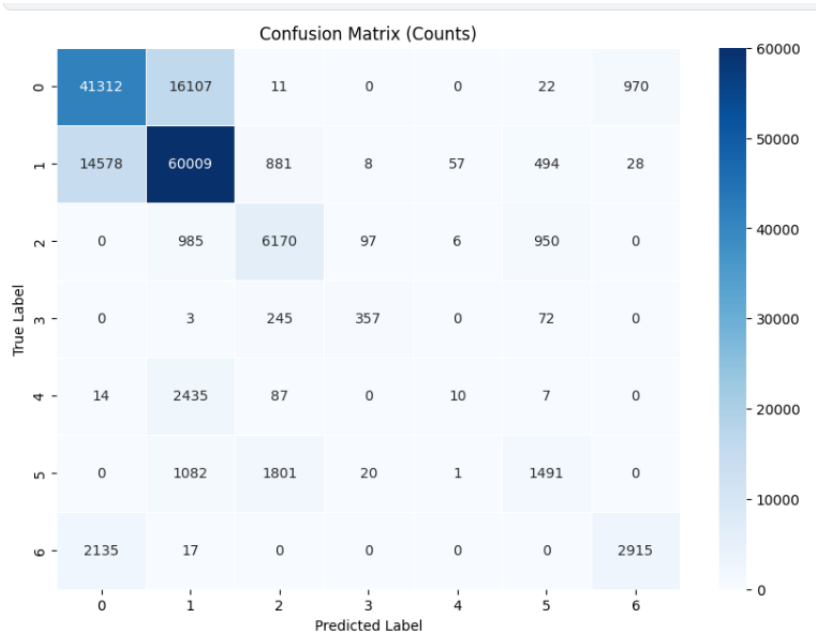

Figure 5.4: Logistic Regression

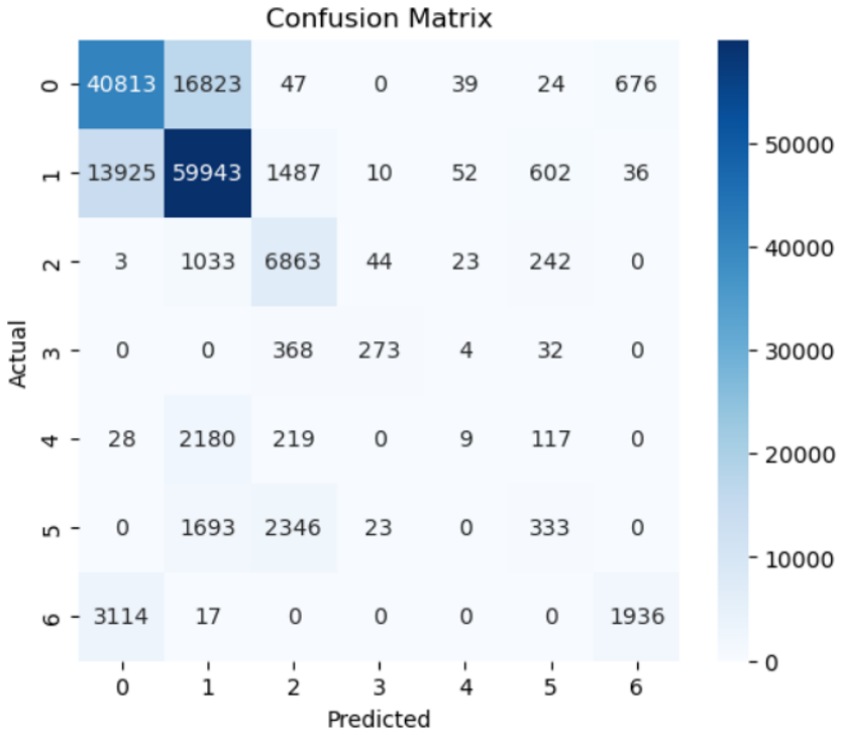

Figure 5.5: Linear SVM

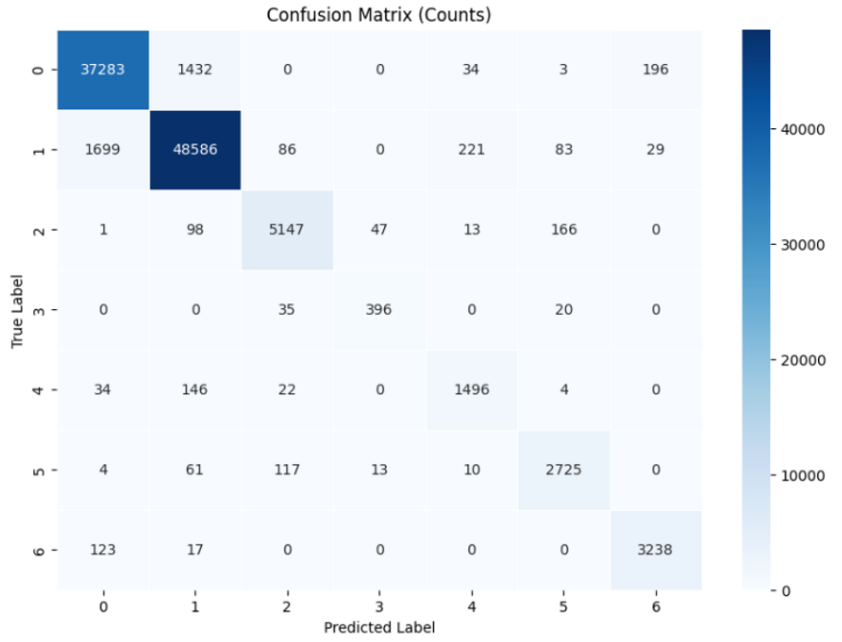

Figure 5.6: Neural Network

Chapter 6

Conclusion

This project successfully implemented and evaluated machine learning pipelines for two distinct classification challenges: predicting Smoker Status and classifying Forest Cover Types using cartographic variables. Through rigorous data preprocessing, feature engineering, and hyperparameter optimization using Optuna, we analyzed the performance of linear (Logistic Regression, Linear SVM) and non-linear (Neural Network) models.

6.1 Summary of Findings

The experimental results highlighted a fundamental difference in the complexity and decision boundaries of the two datasets.

6.1.1 Dataset 1: Smoker Status Prediction

The analysis of the Smoker Status dataset revealed that the relationship between physiological and smoking habits is largely linear in the transformed feature space.

- Performance Consistency: All three models achieved highly comparable accuracy scores, ranging from **73.50%** (Linear SVM) to **73.99%** (Neural Network).
- Linear Separability: The negligible performance gap between the complex Neural Network and the simpler linear models suggests that the decision boundary separating smokers from non-smokers is relatively simple.
- Challenges: The primary challenge identified was the False Negative rate. The confusion matrices indicated a tendency to classify smokers as non-smokers, likely due to the inherent overlap between the two groups.

6.1.2 Dataset 2: Forest Cover Type Prediction

In contrast, the Forest Cover Type dataset demonstrated significant non-linear dependencies, heavily favoring the Neural Network architecture.

- Model Disparity: A substantial performance gap was observed. While Logistic Regression plateaued at 71% and Linear SVM plateaued at 80% accuracy, the Neural Network achieved a superior accuracy of **95.45%**.

- Complexity of Topography: The confusion matrices for linear models showed high misclassification rates between dominant classes (Spruce/Fir and Lodgepole Pine). This indicates that topographical features interact in complex, non-linear ways that cannot be captured by a simple hyperplane.
- Efficacy of MLP: The Multi-Layer Perceptron successfully disentangled these complex feature interactions, resulting in a clean diagonal on the confusion matrix and high precision across all seven distinct forest cover types.

6.2 Comparative Analysis

The study provides a clear case study on the importance of model selection based on data characteristics:

1. Linear vs. Non-Linear: For datasets with lower dimensional complexity (like the Smoker Status), linear models offer a distinct advantage in terms of interpretability and computational efficiency without sacrificing predictive performance.
2. Scalability and Complexity: For large-scale, high-dimensional data with complex interactions (like the Forest Cover dataset), non-linear architectures such as Neural Networks are essential to achieve acceptable classification standards.

Chapter 7

GitHub Repository

The complete source code, datasets, and resources for this project are available in our GitHub repository:

`Smoker-Status-and-Forest-Cover-Types-using-Machine-Learning`

This repository contains all scripts, documentation, and materials necessary to reproduce our analysis and results.