International Institute of Information Technology, Bangalore

Computer Science and Engineering

Comparative Analysis of Classification Models

Smoker Status Prediction & Forest Cover Type Classification

Course Name: Machine Learning

Course Code: AIT 511

Submitted By:
Ruturaj Dnyandeo Wairkar
Roll No: MT2025108

Vivek Joshi
Roll No: MT2025133

Submitted To:
Prof Sushree Behera

Project Type:
End Semester Project

GitHub Repository:
https://github.com/vivekvj18/ML-PROJECT-2

Contents

List of Figures

List of Tables

Abstract

This report presents a comparative study of machine learning models applied to two classification tasks of different complexities: **Smoker Status Prediction** (binary) and **Forest Cover Type Classification** (multiclass). The objective is to evaluate how classical machine learning algorithms and deep learning models perform under varying feature distributions, dataset sizes, and class imbalances.

For the Smoker dataset consisting of 38,984 bio-signal records, we experimented with Logistic Regression, Linear SVM, RBF SVM, and a Neural Network. The Neural Network achieved the highest test accuracy of **75.41%**, outperforming the traditional linear models, indicating the presence of non-linear physiological relationships.

For the Forest Cover dataset containing 581,012 samples and seven target classes, we evaluated Logistic Regression, Linear SVM, and a Deep Neural Network (MLP). Owing to its ability to model complex interactions among topographical and soil-related features, the MLP achieved a significantly higher accuracy of **91.03%** compared to linear baselines.

Overall, the study highlights that while linear methods offer strong baselines and interpretability, deep neural networks provide substantial performance gains on large, heterogeneous, and inherently non-linear datasets. The analysis demonstrates how preprocessing strategies, feature characteristics, and model capacity collectively influence classification performance across tasks.

1 Introduction

1.1 Problem Statement

Machine learning models behave differently depending on the nature of the target variable and the feature space. This project aims to compare standard classification algorithms across two fundamental types of problems:

1. **Binary Classification:** Predicting whether an individual is a smoker based on bio-signals.
2. **Multiclass Classification:** Categorizing forest cover types based on geological and cartographical features.

1.2 Dataset Description

1.2.1 Dataset 1: Smoker Status Prediction

The competition dataset was sourced from Kaggle (Bio-signals).

- **Size:** 38,984 samples with 23 features.
- **Target:** Smoking Status (0: Non-smoker, 1: Smoker).
- **Features:** Age, height, weight, eyesight, hearing, and various blood test metrics (cholesterol, hemoglobin, etc.). See the smoking notebook for full EDA and code.

1.2.2 Dataset 2: Forest Cover Type

The dataset consists of cartographic variables derived from US Geological Survey data.

- **Size:** 581,012 samples with 55 features.
- **Target:** Cover_Type (Integers 1-7).
- **Features:** Elevation, aspect, slope, distances to hydrology/roadways, and binary soil type indicators. Full forest experiments (logistic/SVM and NN) are in the uploaded Colab notebooks.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical step in understanding the dataset's structure. We performed separate analyses for both datasets (plots and code in the notebooks).

2.1 Dataset 1: Smoker Status Prediction

2.1.1 Univariate Analysis: Distributions

We analyzed the distributions of continuous variables. The histograms below reveal the spread of physical attributes like age, height, and weight, as well as biochemical markers.

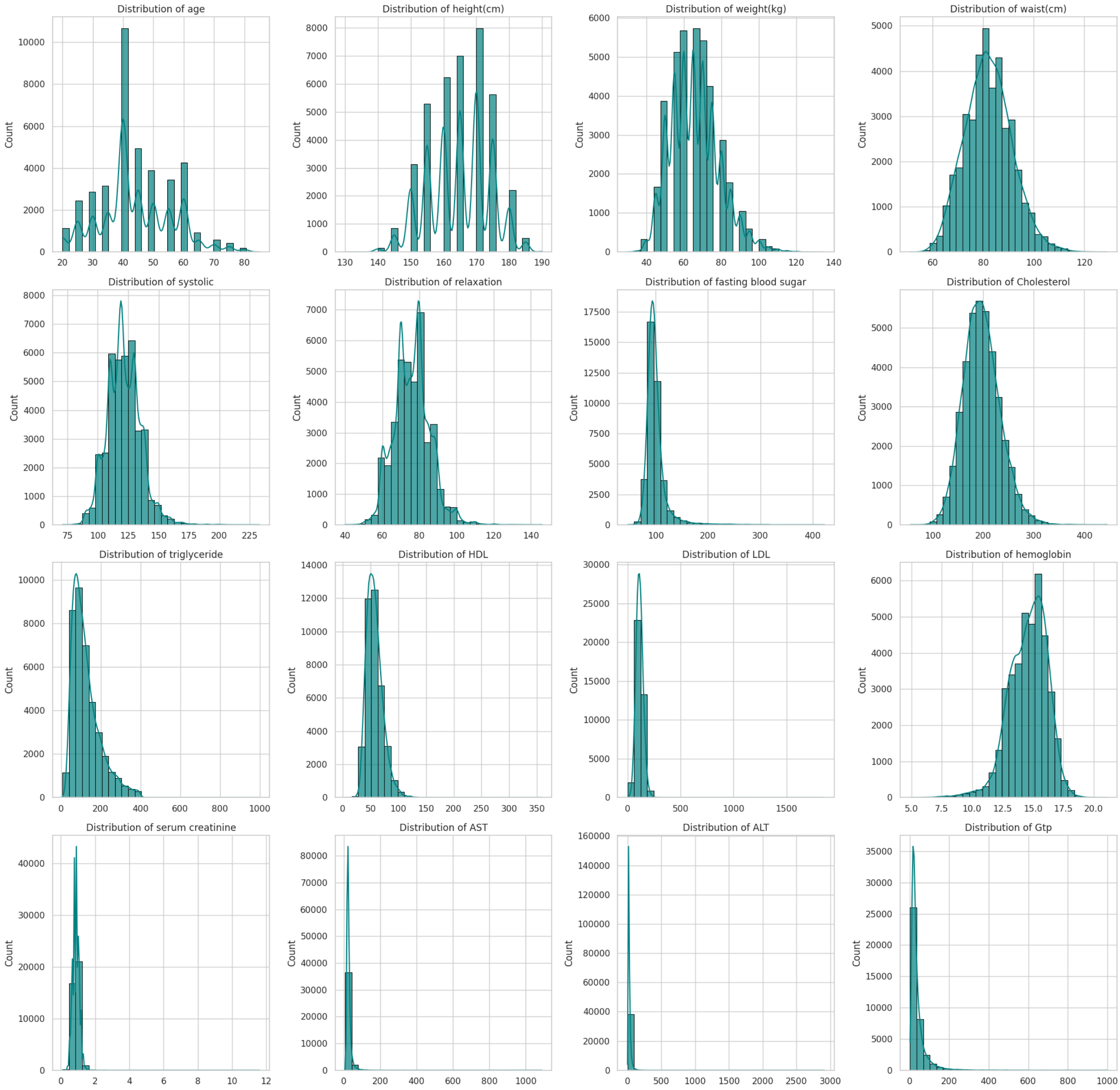

Figure 1: Distribution of Numerical Bio-Signals.

2.1.2 Outlier Analysis

To identify anomalous physiological data points, we utilized boxplots. Significant outliers were observed in features like AST, ALT, and Gtp (IQR capping applied in preprocessing). See smoking notebook for code and figures.

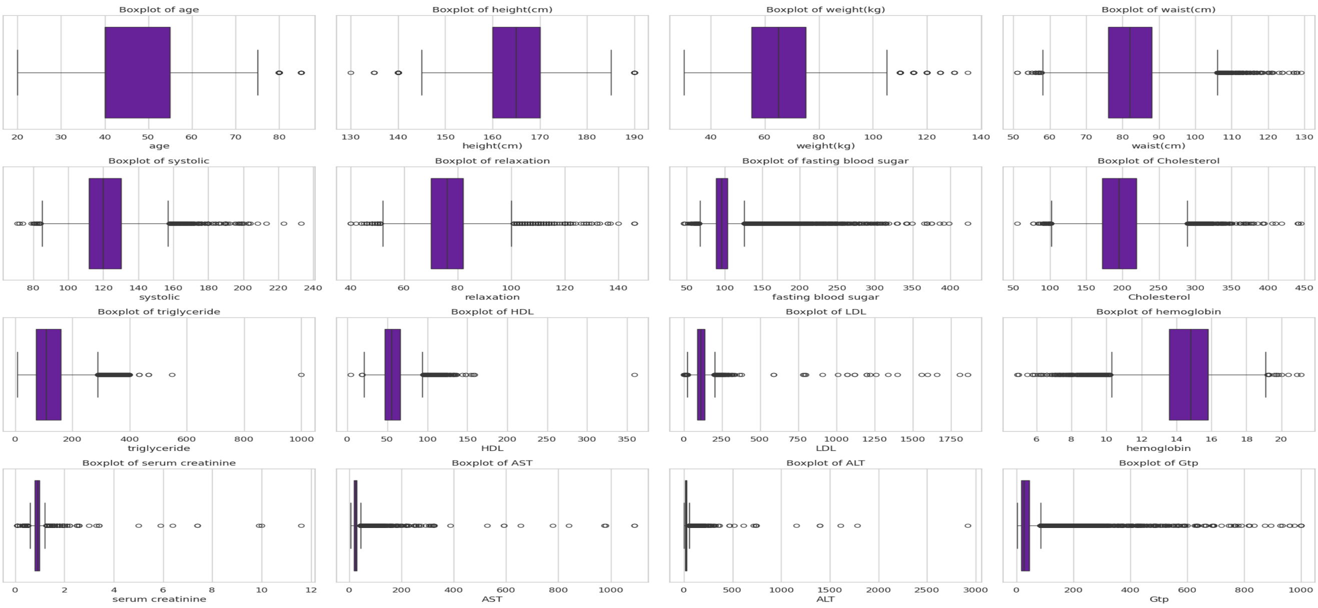

Figure 2: Boxplots for Outlier Detection (Smoker Dataset).

2.1.3 Categorical Feature Analysis

We examined the counts of categorical variables such as hearing ability and dental caries to understand the population demographics (plots in notebook).

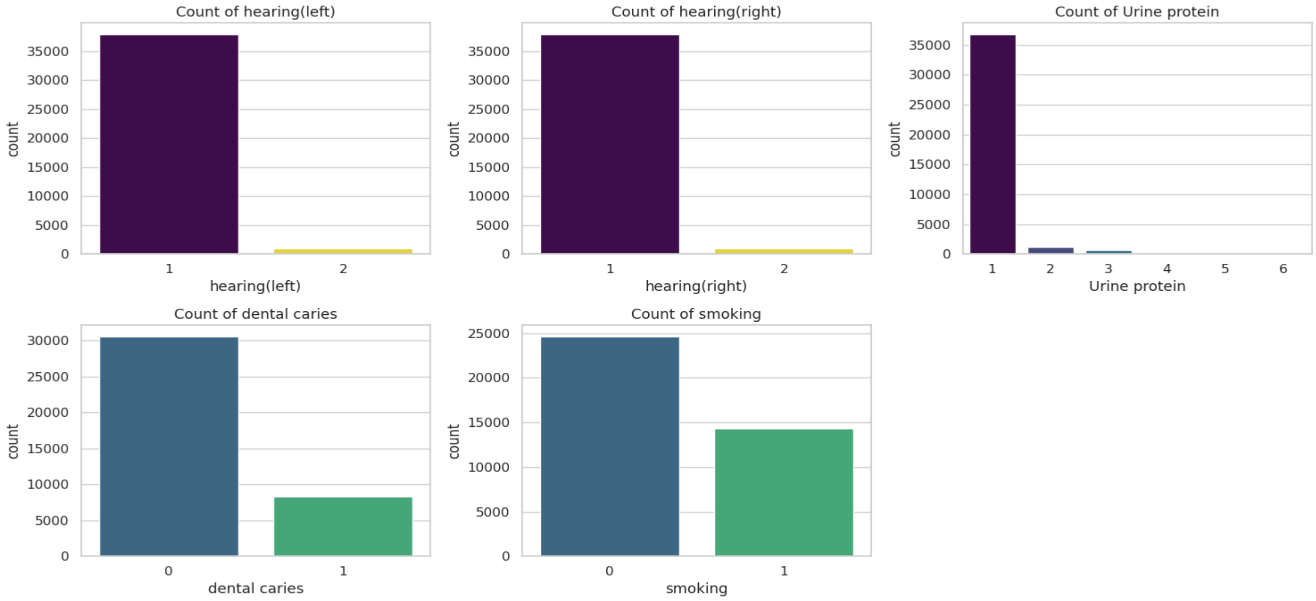

Figure 3: Counts of Categorical Variables (Smoker Dataset).

2.1.4 Correlation Analysis

A correlation heatmap was generated to identify multicollinearity. Strong correlations were found between Weight/Waist and Cholesterol/LDL; these guided feature selection (drop Cholesterol, waist).

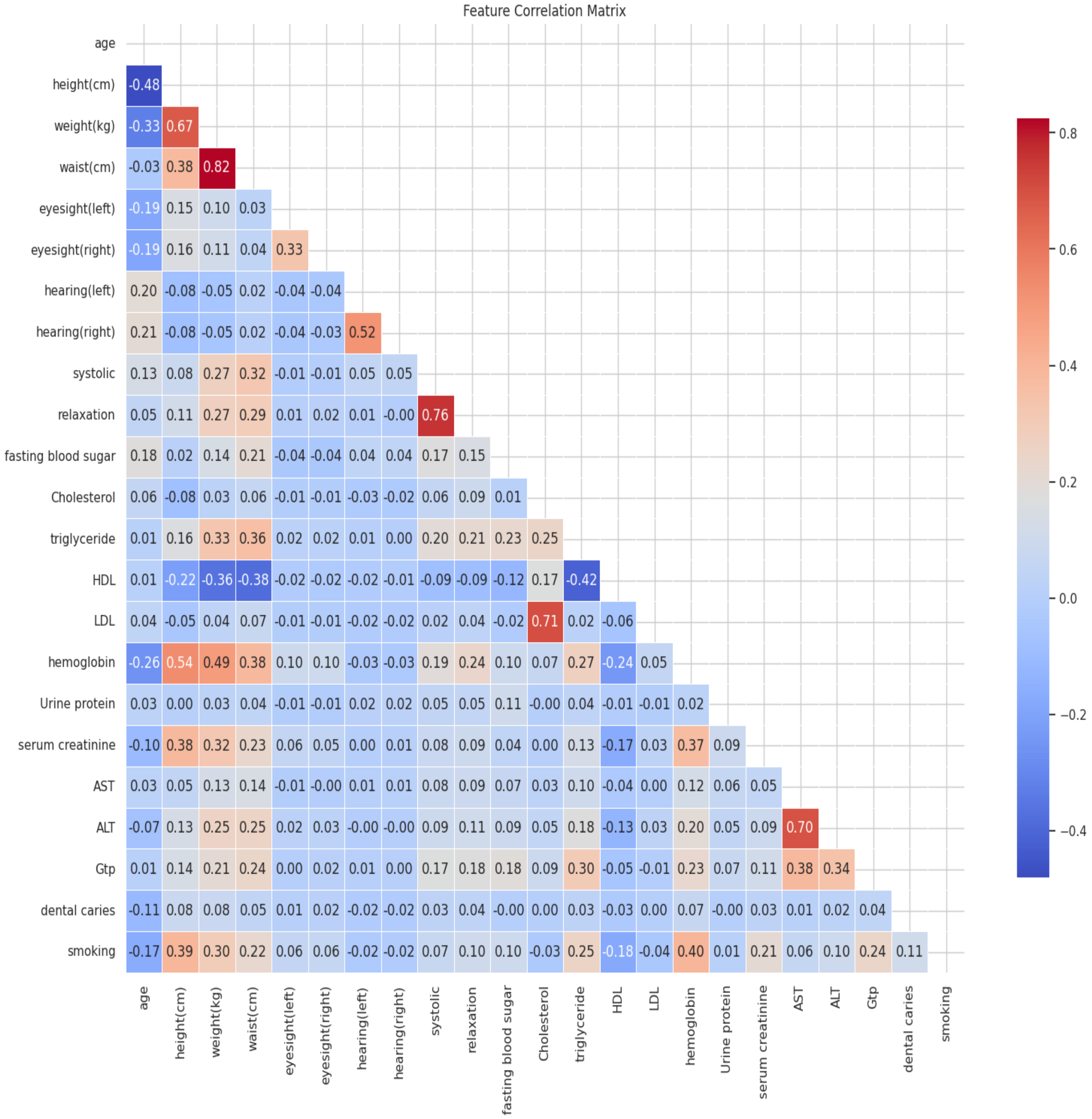

Figure 4: Correlation Matrix of Smoker Dataset Features.

2.2 Dataset 2: Forest Cover Type

2.2.1 Target Class Distribution

The dataset exhibits severe class imbalance, with Classes 1 and 2 dominating the distribution (visualized in notebook). Class weights were computed for NN training.

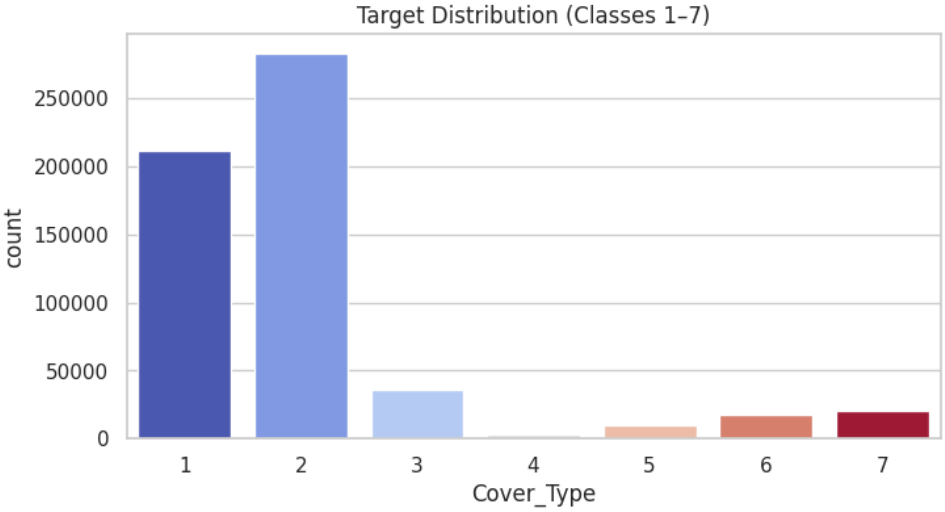

Figure 5: Distribution of Forest Cover Types.

2.2.2 Continuous Feature Analysis

Histograms of continuous features like Elevation and Aspect reveal the topographical characteristics of the regions. See forest notebook for full plots.

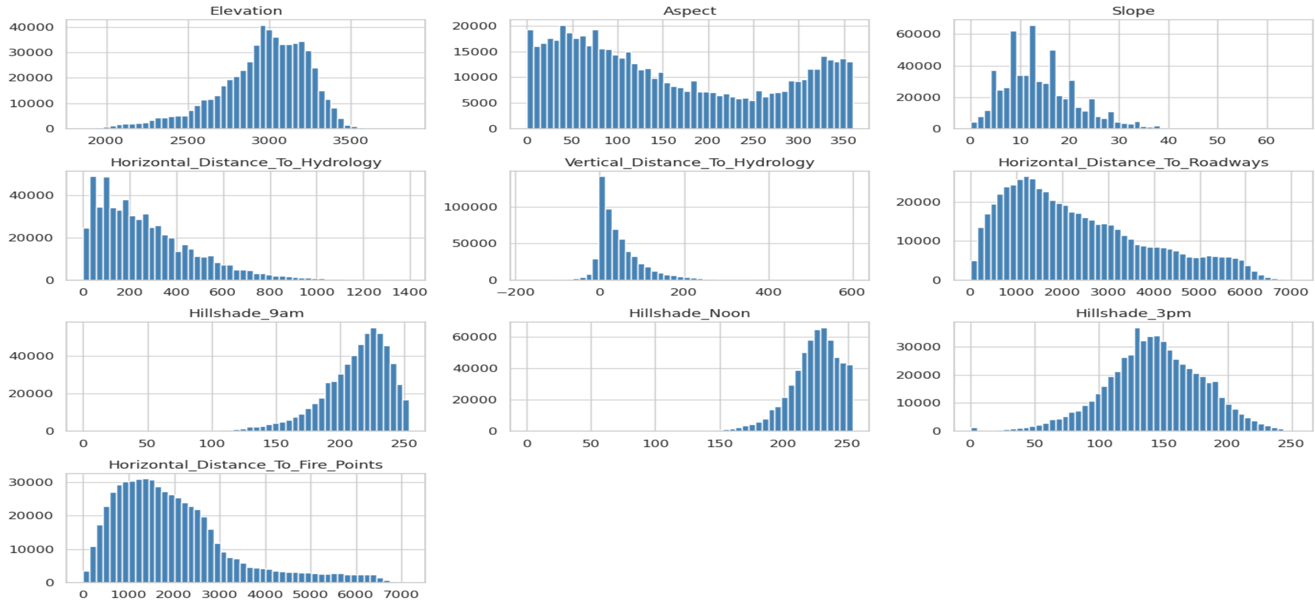

Figure 6: Distribution of Continuous Features (Forest Dataset).

2.2.3 Feature Correlations

The correlation matrix examines the relationships between numerical features to find strong associations and potential redundancies used for feature selection.index=12

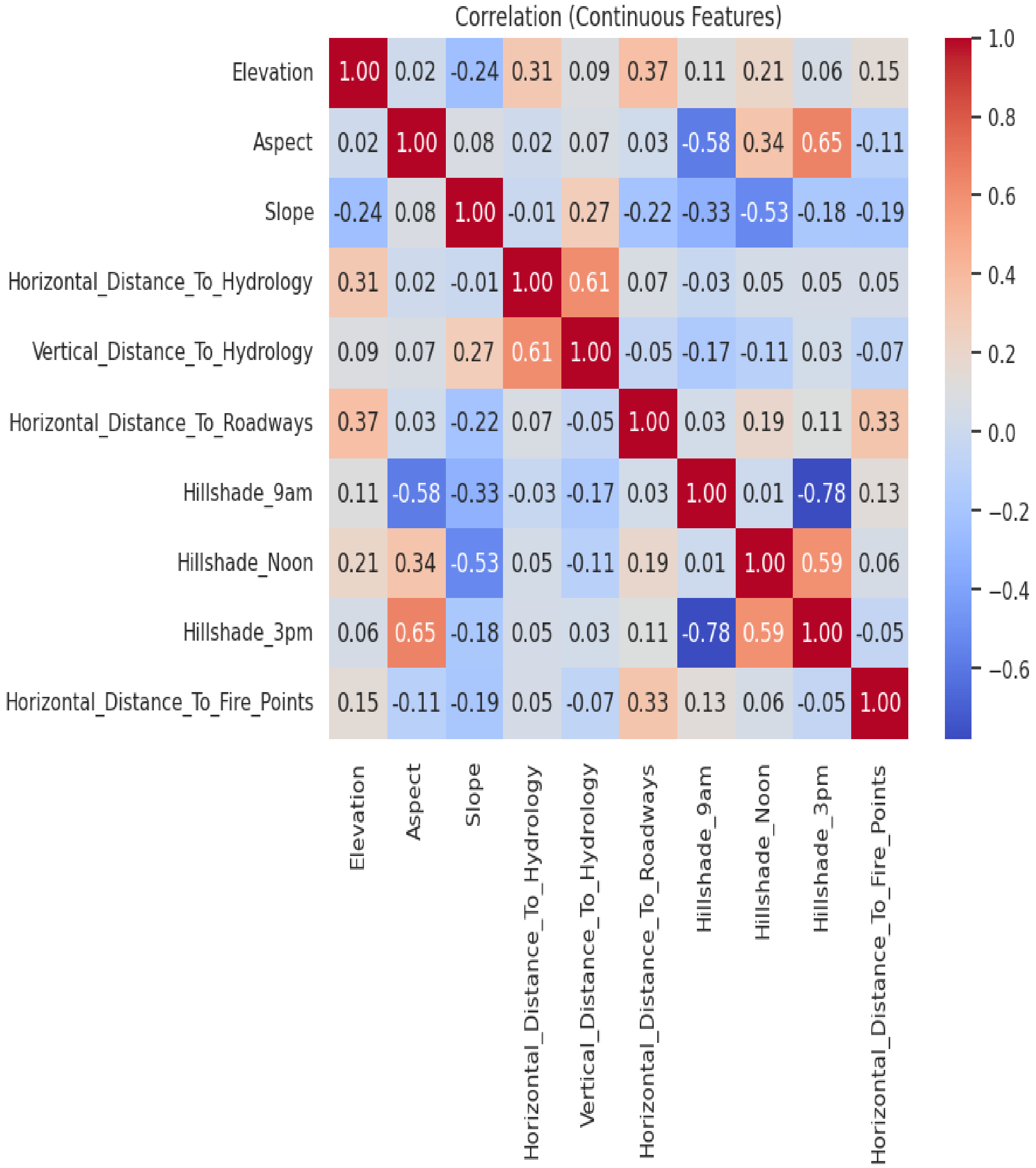

Figure 7: Correlation Matrix of Forest Dataset Features.

2.2.4 Boxplots

The boxplots illustrate distribution, skew, and outliers for continuous features; these informed IQR capping decisions. See forest notebook.

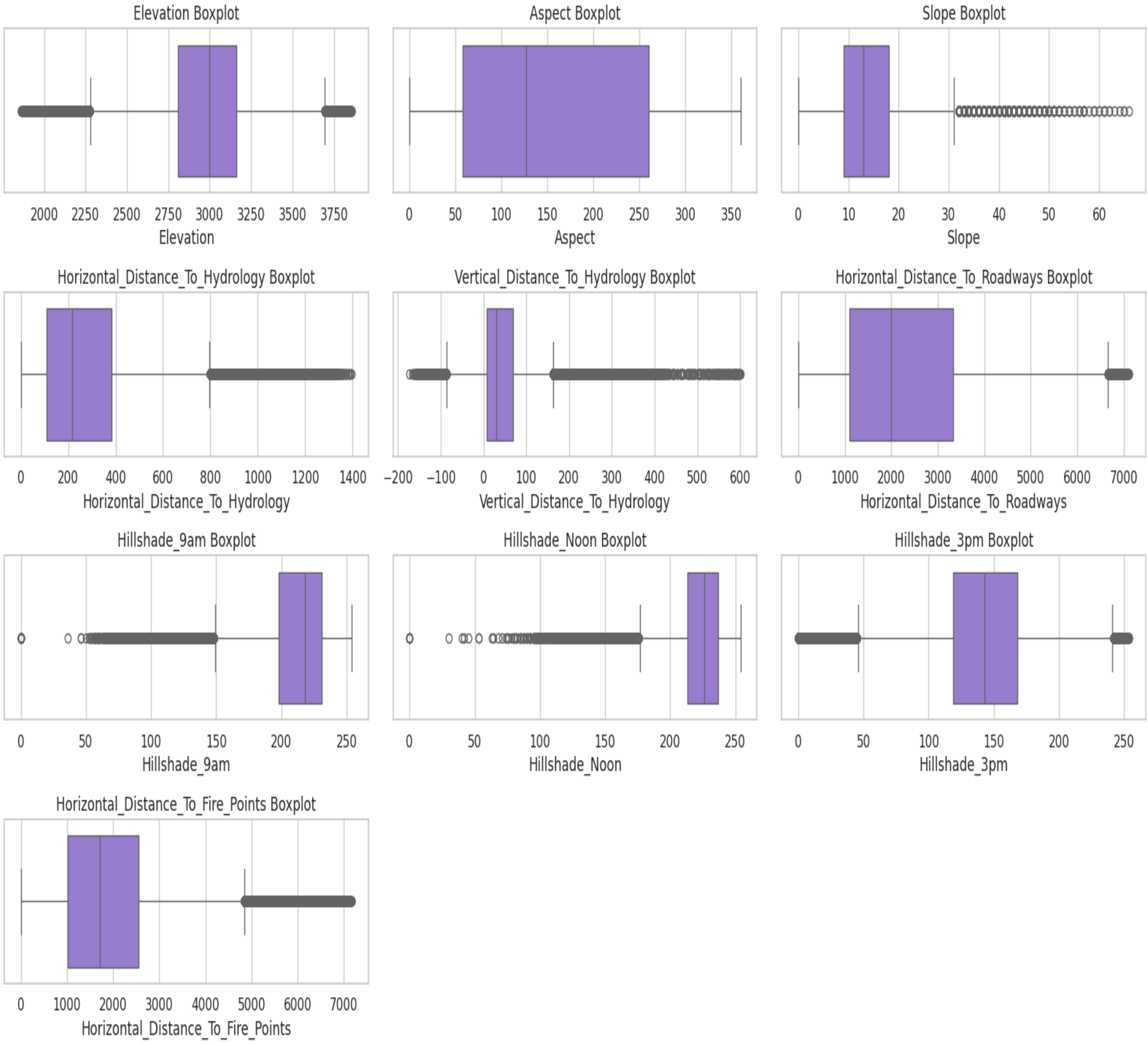

Figure 8: Boxplots of Continuous Features (Forest Dataset).

3 Data Preprocessing

3.1 Overview

We applied dataset-specific but consistent preprocessing to ensure models receive clean, comparable inputs. Major steps: deduplication, feature selection, outlier handling, scaling/encoding, and class imbalance handling where required (see notebooks for code).

3.2 Dataset 1: Smoker Status

- **Deduplication & Missing Values:** Removed 5,517 duplicate rows; median imputation where necessary.
- **Feature Selection:** Dropped `Cholesterol` and `waist(cm)` due to high correlation with LDL and weight respectively.
- **Outlier Handling:** IQR-based capping (1.5*IQR) applied to physiological features.
- **Scaling & Encoding:** StandardScaler for continuous features; one-hot encoding for categorical variables.

3.3 Dataset 2: Forest Cover Type (Condensed)

- **Scaling:** StandardScaler for continuous features; binary indicators left unchanged.
- **Subsampling & Efficiency:** Random subsampling used during hyperparameter tuning (SVM) to speed experiments; final models trained on appropriate data.
- **Class Imbalance:** Class weights computed and applied in NN training; monitored per-class recall/F1.

4 Models Used & Hyperparameters

We evaluated baseline and advanced models. Hyperparameters below match the notebook runs.

4.1 Smoker Status (Binary) — Hyperparameters

Table 1: Model Hyperparameters — Smoker Dataset

Model	**Hyperparameter**	**Value**
Logistic Regression	penalty	L2
	C	1.0
	solver	`lbfgs`
SVM (Linear)	C	1.0
SVM (RBF)	C	1.0
	gamma	`scale`
Neural Network (MLP)	Layers	[128, 32]
	Activation	ReLU
	Dropout	0.3
	Optimizer	Adam (lr = 1e-3)
	Batch size	32
	Epochs	50

4.2 Forest Cover Type (Multiclass) — Hyperparameters

Table 2: Model Hyperparameters — Forest Dataset

Model	**Hyperparameter**	**Value**
Logistic Regression	multi_class	multinomial
	solver	lbfgs
	max_iter	2000
Linear SVM	C	10.0
	dual	False
	max_iter	2000
Neural Network (MLP)	Layers	[1024, 512, 256]
	Activation	ReLU
	Dropout	0.15
	L2 reg	1e-5
	BatchNorm	Yes
	Optimizer	Adam (lr=3e-4)
	Batch size	1024
	Epochs	30
	Class Weights	balanced

5 Evaluation Metrics & Experiment Setup

5.1 Evaluation Metrics

To ensure a comprehensive and fair comparison across models, we employed multiple evaluation metrics suited for both binary and multiclass classification tasks:

- **Accuracy:** Proportion of correctly predicted labels over the total samples. Used as a primary metric for both datasets.
- **Precision:** Measures the correctness of positive predictions; important for assessing false-positive tendencies.
- **Recall (Sensitivity):** Measures the model's ability to identify all positive instances, especially relevant for imbalanced classes.
- **F1-Score:** Harmonic mean of precision and recall; useful when class distribution is skewed.
- **Confusion Matrix:** Provides detailed insight into class-wise misclassifications, enabling diagnostic error analysis.

All metrics were computed using `scikit-learn` and evaluated on the held-out test set to avoid overfitting.

5.2 Experiment Setup

The experimental workflow was carefully structured to ensure reproducibility, fairness, and consistency across both datasets:

- **Train/Test Split:** An 80/20 stratified split was used to preserve the class distribution across partitions for both datasets.
- **Preprocessing Protocol:** All preprocessing steps—including scaling, outlier handling, feature selection, and encoding—were fit exclusively on the training set and later applied to the test set. This ensured **no data leakage** into the model training phase.
- **Class Imbalance Handling:** For the Forest Cover dataset, **class weights** were computed from the training data and incorporated into the neural network's loss function to counteract imbalance among the seven cover types.
- **Efficiency Considerations:** Certain hyperparameter tuning steps for SVMs on the Forest dataset used **random subsampling** to reduce computational cost while preserving model behavior.
- **Reproducibility:** Random seeds were fixed across NumPy, scikit-learn, and TensorFlow (where applicable) to maintain consistent results across runs.

This standardized and rigorous experimental setup ensured that model comparisons were meaningful, reproducible, and directly reflective of the underlying dataset characteristics.

6 Methodology

We implemented and compared multiple algorithms following a consistent pipeline:

1. Load and inspect data (duplicates, missing values).
2. Deduplicate and cap outliers using IQR (1.5*IQR).
3. Drop highly correlated features (e.g., Cholesterol, waist).
4. Train/test split (stratified), scale continuous features with StandardScaler.
5. Train models (Logistic, SVM variants, NN), tune using CV where feasible.
6. Evaluate on held-out test set and record accuracy / per-class metrics. Notebook code shows all steps.

6.1 Models implemented

- **Smoker dataset:** Logistic Regression, Linear SVM, RBF SVM (K-Fold), Neural Network (Keras).
- **Forest dataset:** Logistic Regression, Linear SVM, Neural Network (MLP with class weights).

7 Results, Comparative Analysis & Discussion

7.1 Smoker Status Prediction Results

7.1.1 Performance Summary (Selected)

Table 3: Selected Test Accuracy — Smoker Dataset

Model	Accuracy
Logistic Regression	72.69%
Linear SVM	73.08%
SVM (RBF, K-Fold)	74.79%
Neural Network (K-Fold)	**75.41%**

7.1.2 Comparative Discussion

The MLP performed best (75.4%), followed by RBF-SVM — both indicate non-linear feature interactions in the bio-signal space. Logistic and linear SVMs served as solid baselines with lower compute cost. Detailed confusion matrices and classification reports are available in the notebook.

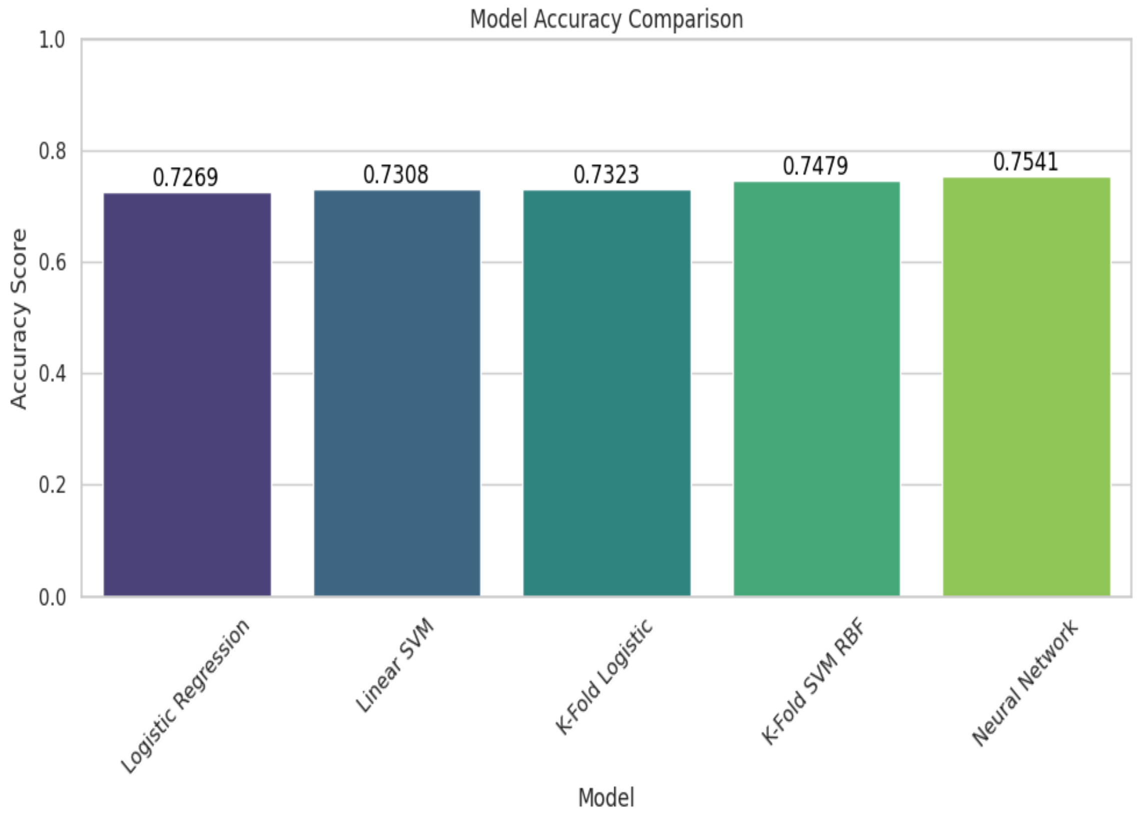

Figure 9: Model Accuracy Comparison

7.1.3 Confusion Matrices

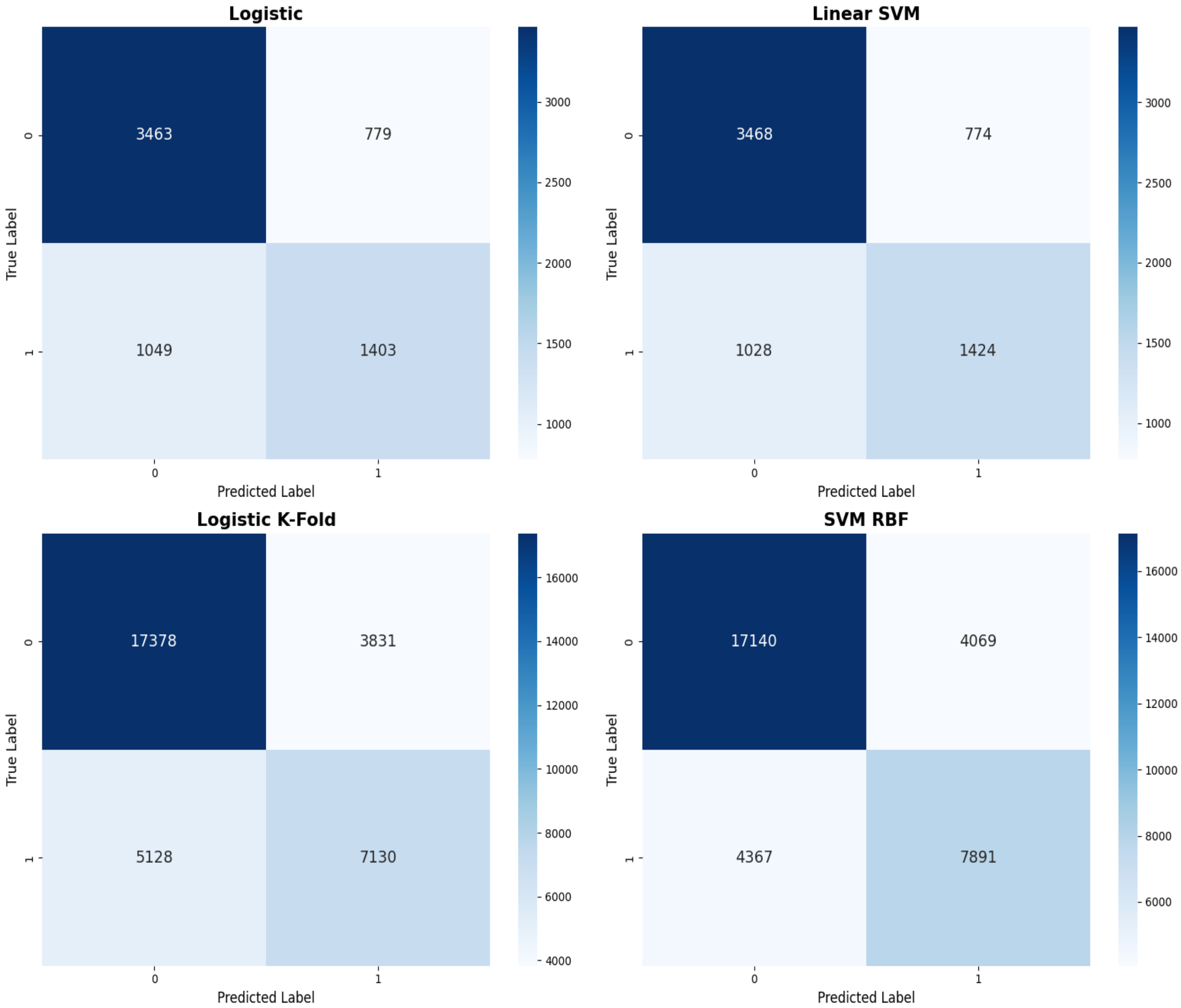

Figure 10: Confusion Matrices

7.2 Forest Cover Type Results

7.2.1 Performance Summary (Selected)

Table 4: Selected Test Accuracy — Forest Dataset

Model	Accuracy
Logistic Regression	72.37%
Linear SVM	71.14%
Neural Network (MLP)	**91.03%**

(Logistic/SVM results and NN accuracy are from your two forest notebooks).

7.2.2 Comparative Discussion

The MLP (deep MLP with BN/Dropout/L2) produced a major improvement over linear methods — demonstrating strong non-linear structure and feature interactions in the forest data. Class weighting was important to improve minority-class recall; confusion matrices in the NN notebook show remaining confusions between similar cover types.

7.2.3 Confusion Matrices

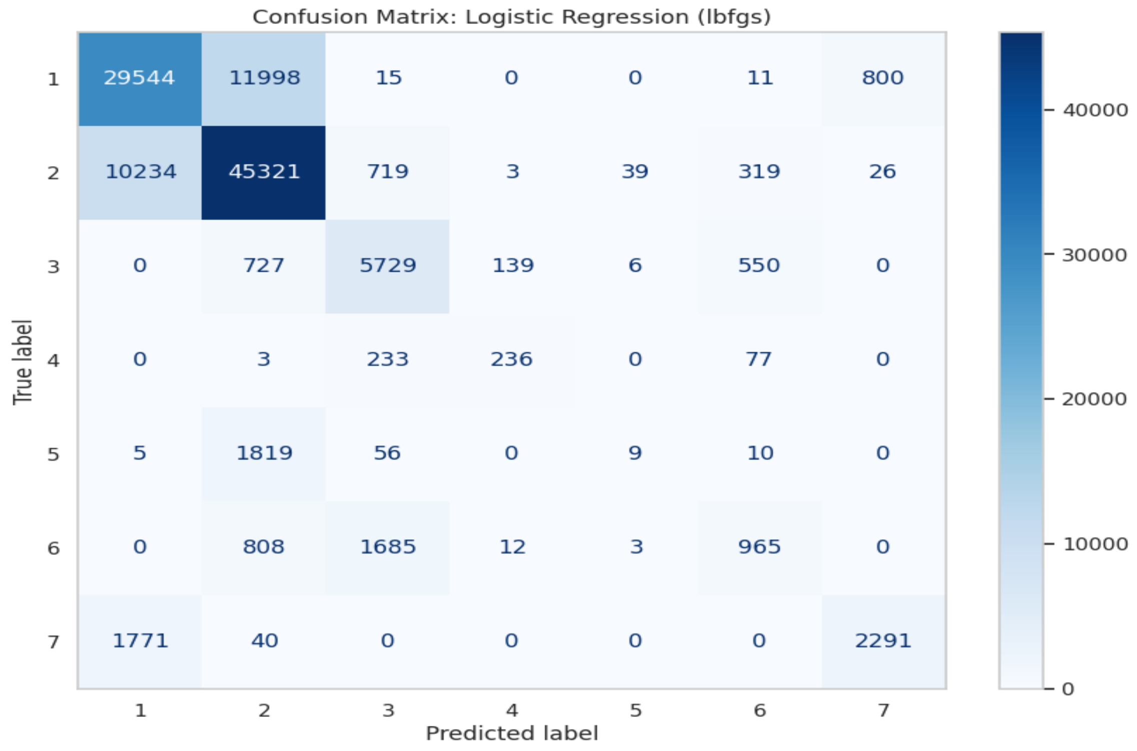

Figure 11: Logistic Regression

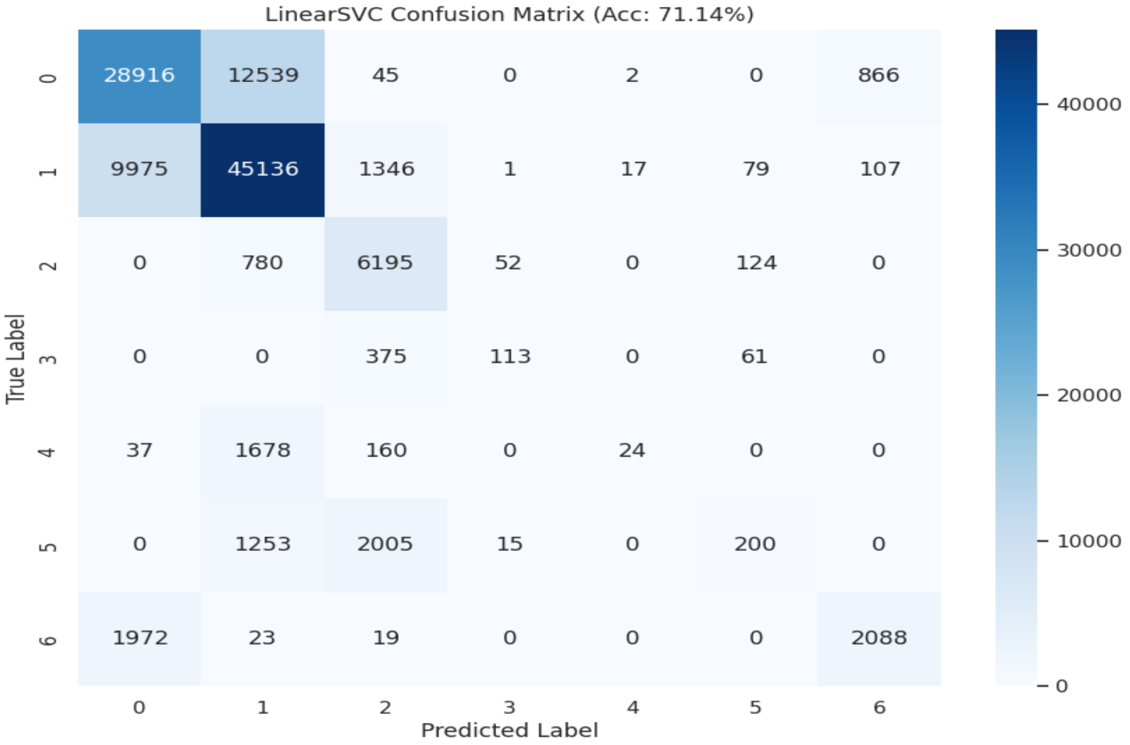

Figure 12: LINEAR SUPPORT VECTOR MACHINE

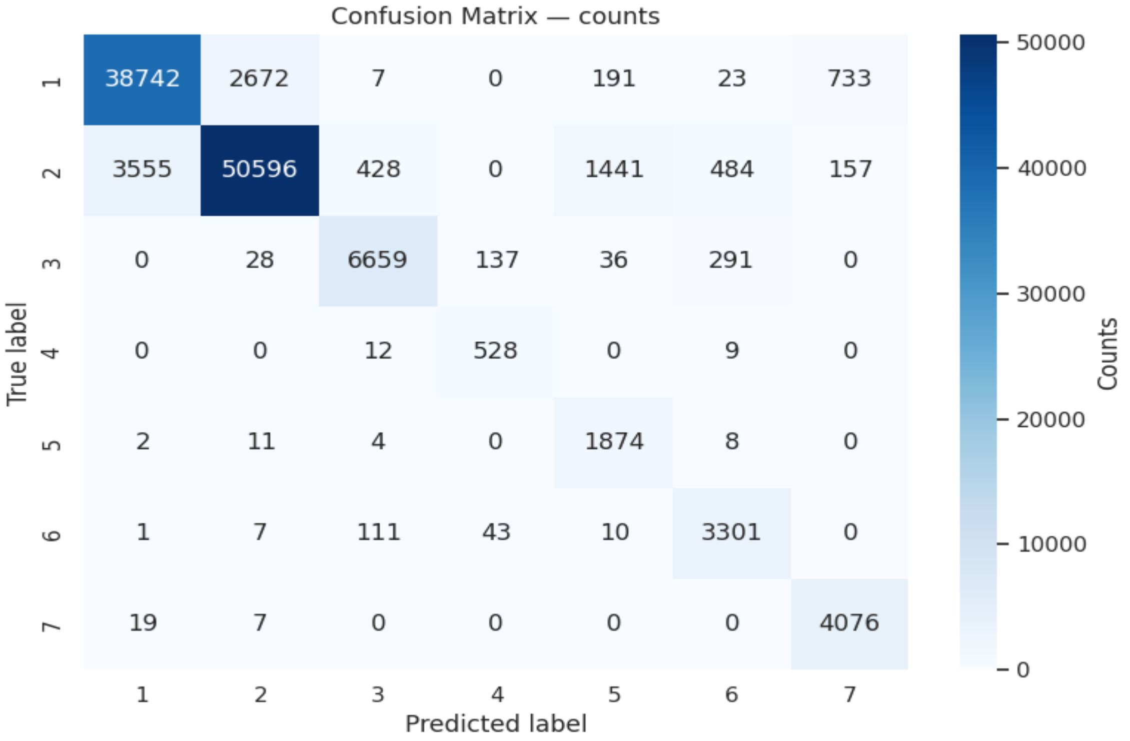

Figure 13: Neural Network

8 Conclusion and Future Work

8.1 Conclusion

This comparative study successfully evaluated multiple classification algorithms across diverse domains.

- For **Smoker Status Prediction**, we achieved a peak accuracy of **75.41%** using a **Neural Network**. The study highlighted the importance of cleaning duplicates and handling multicollinearity in bio-signal data.
- For **Forest Cover Classification**, the **Neural Network** proved dominant, achieving **91.03%** accuracy. The project underscored the necessity of non-linear models and class-weighting strategies when dealing with large, imbalanced, and geologically complex datasets.

8.2 Future Work

- **Feature Engineering:** For the Smoker dataset, creating interaction terms (e.g., BMI from height/weight) could likely improve linear model performance.
- **Ensemble Methods:** While we tested Random Forest, stacking it with the Neural Network could potentially push the Forest Cover accuracy even higher by combining tree-based logic with deep learning representations.
- **Hyperparameter Tuning:** More extensive Bayesian Optimization (e.g., using Optuna) could further refine the Neural Network architectures.

9 References

1. Dataset 1: Smoker Status Prediction using Bio-Signals. Kaggle. https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals

2. Dataset 2: Forest Cover Type Dataset. UCI Machine Learning Repository. https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset

3. Scikit-learn Documentation. https://scikit-learn.org/stable/

4. Tensorflow Documentation. https://www.tensorflow.org/