Machine Learning-MT-AIT 511 Project Report

Aditya Lahkar
MT2025014

December 12, 2025

Contents

1 Introduction

This report details the implementation, analysis, and results of applying machine learning techniques to two distinct datasets: **Smoker Status Prediction** (Binary Classification) and **Forest Cover Type** (Multiclass Classification). The primary objective was to compare the performance of Logistic Regression, Support Vector Machines (SVM), and Neural Networks (MLP) after rigorous preprocessing and hyperparameter tuning.

2 Part 1: Smoker Status Prediction

2.1 Dataset Details

The Smoker Status Prediction dataset consists of bio-signal data aimed at classifying individuals as smokers or non-smokers.

- **Type**: Binary Classification.
- **Samples**: Approx. 39,000 (after cleaning).
- **Features**: Age, Height, Weight, Waist, Blood Pressure (systolic/relaxation), Cholesterol (HDL/LDL), Triglycerides, Hemoglobin, Urine Protein, Serum Creatinine, AST, ALT, Gtp, Dental Caries.
- **Target**: `smoking` (0 = Non-smoker, 1 = Smoker).

2.2 Preprocessing EDA

Exploratory Data Analysis revealed significant correlations:

- **Hemoglobin**: Strong positive correlation with smoking status.
- **Gtp Triglycerides**: Showed heavy right-skewed distributions with many outliers.

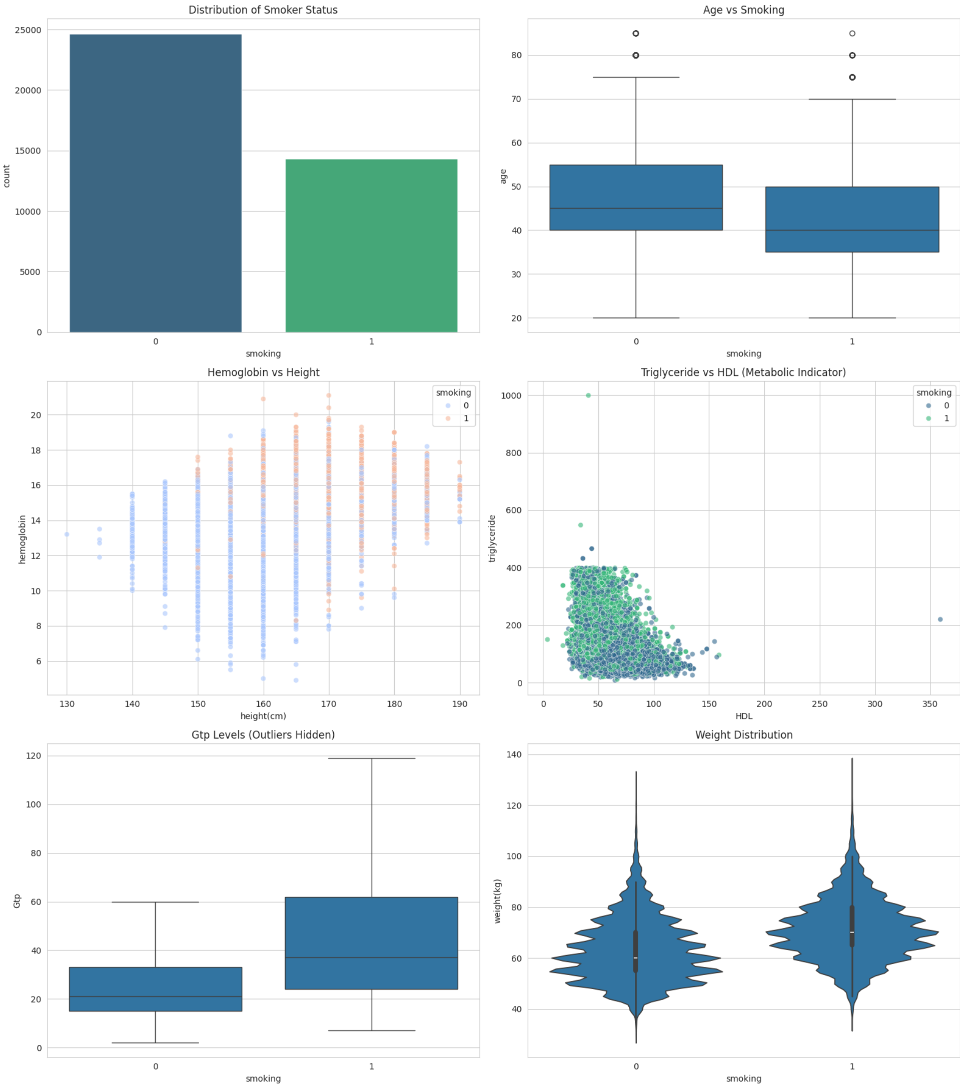

Figure 1: EDA: Bio-signal distributions and correlations with Smoking Status.

Steps Taken:

1. **Feature Engineering**: Created `BMI` (Weight/Height^2) and `Waist-to-Height Ratio (WHtr)` to capture body shape health indicators.
2. **Scaling**: Used `RobustScaler` instead of StandardScaler to mitigate the impact of extreme outliers in Gtp and Triglycerides.
3. **Encoding**: One-hot encoding was not required as all input features were numerical.

2.3 Model Implementation Hyperparameters

Models were tuned using `RandomizedSearchCV`.

1. **Logistic Regression**: Tuned `C` and `solver`.
2. **SVM**: Tuned `C`, `gamma`, and `kernel` (RBF vs Linear). *Note: Subsampling (5000 samples) was used for tuning to optimize runtime.*
3. **Neural Network**: Tuned `hidden_layer_sizes` and `learning_rate`.

2.4 Evaluation Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.718	0.627	0.572	0.598
SVM (Champion)	**0.753**	**0.667**	**0.658**	**0.662**
Neural Network	0.745	0.658	0.637	0.647

Table 1: Smoker Dataset Results

Conclusion: The SVM (with RBF kernel) outperformed others, likely due to its ability to capture non-linear decision boundaries in the bio-signal space.

3 Part 2: Forest Cover Type

3.1 Dataset Details

The Forest Cover Type dataset contains cartographic variables to predict forest cover type.

- **Type**: Multiclass Classification (7 Classes).
- **Samples**: 581,012 (Large Data).
- **Features**: Elevation, Aspect, Slope, Distances to Hydrology/Roadways/Firepoints, Wilderness Areas (Binary), Soil Types (Binary).

3.2 Preprocessing EDA

Key Insights: `Elevation` is the single most discriminative feature. `Distance to Hydrology` also showed strong class separation.

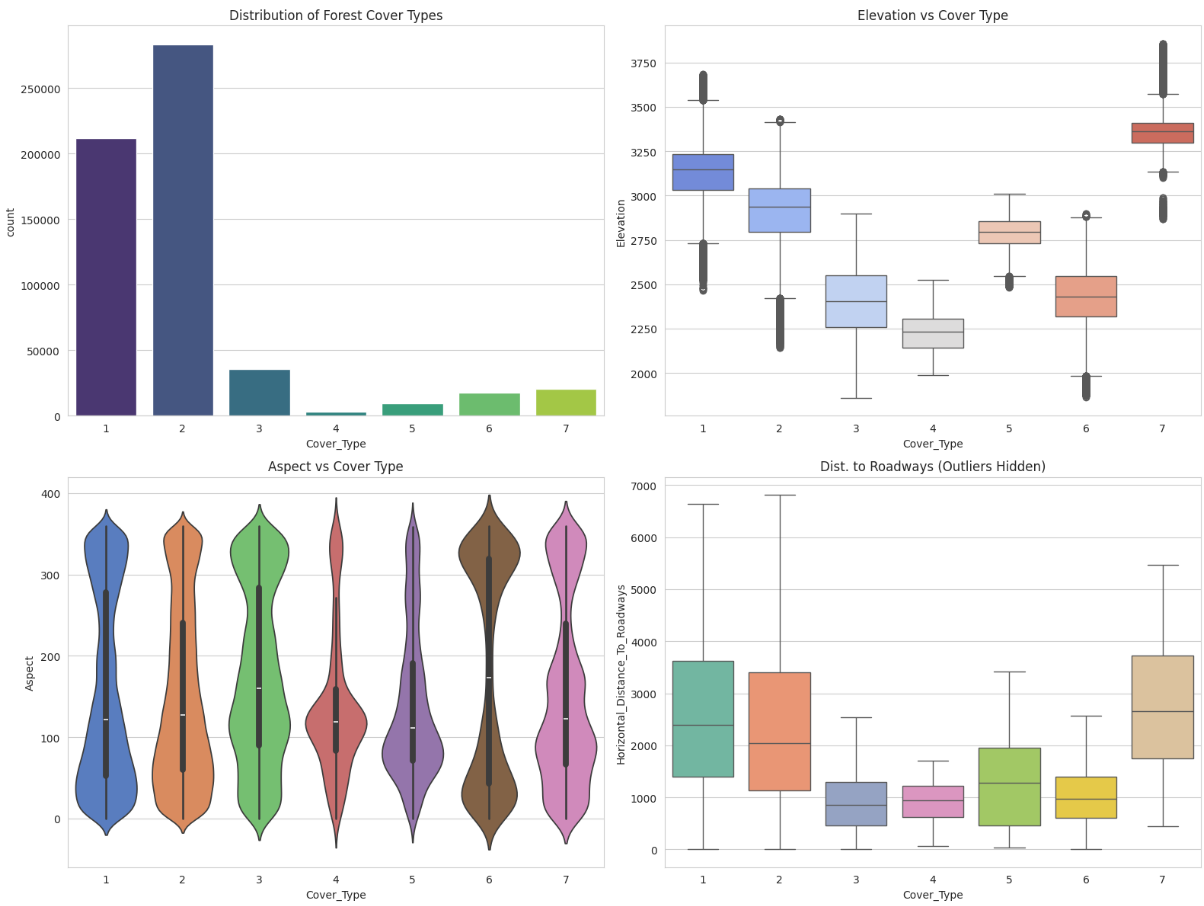

Figure 2: EDA: Elevation and Hydrology Distance impact on Forest Cover Type.

Steps Taken:

1. **Feature Engineering**:
   - **Euclidean Distance to Hydrology**: Combined Horizontal and Vertical distances ($\sqrt{H^2 + V^2}$).

- **Mean Distance to Amenities**: Average distance to Roads, Firepoints, and Water.
- **Water Elevation**: `Elevation - Vertical_Distance_To_Hydrology`.

2. **Subsampling Strategy**:
   - Logistics Regression Neural Network: Trained on **Full Dataset** (581k rows).
   - SVM: Capped at **20,000 samples** due to $O(n^3)$ complexity making full training infeasible.
3. **Scaling**: `StandardScaler` applied to continuous features.

3.3 Model Implementation

1. **Logistic Regression**: Multinomial solver (`lbfgs`).
2. **SVM**: RBF Kernel, effectively capped sample size.
3. **Neural Network**: MLP with sizes like (100, 50), ReLu activation.

3.4 Evaluation Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.724	0.713	0.723	0.714
SVM (Subset)	0.793	0.789	0.793	0.787
Neural Network (Champion)	**0.876**	**0.875**	**0.876**	**0.875**

Table 2: Forest Dataset Results (Weighted Averages)

Conclusion: The Neural Network dominated this task (87.6% Accuracy). The dataset's complexity and large sample size naturally favor deep learning approaches over linear models or sample-constrained SVMs.

4 Final Conclusion

- For the **Smoker dataset** (mid-sized, biological data), **SVM** proved most effective at finding the refined decision boundary.
- For the **Forest dataset** (large-scale, high-dimensional cartographic data), the **Neural Network** significantly outperformed traditional methods.
- **Preprocessing Impact**: Feature engineering (hydrology distance, BMI) and robust scaling were critical in achieving these scores.

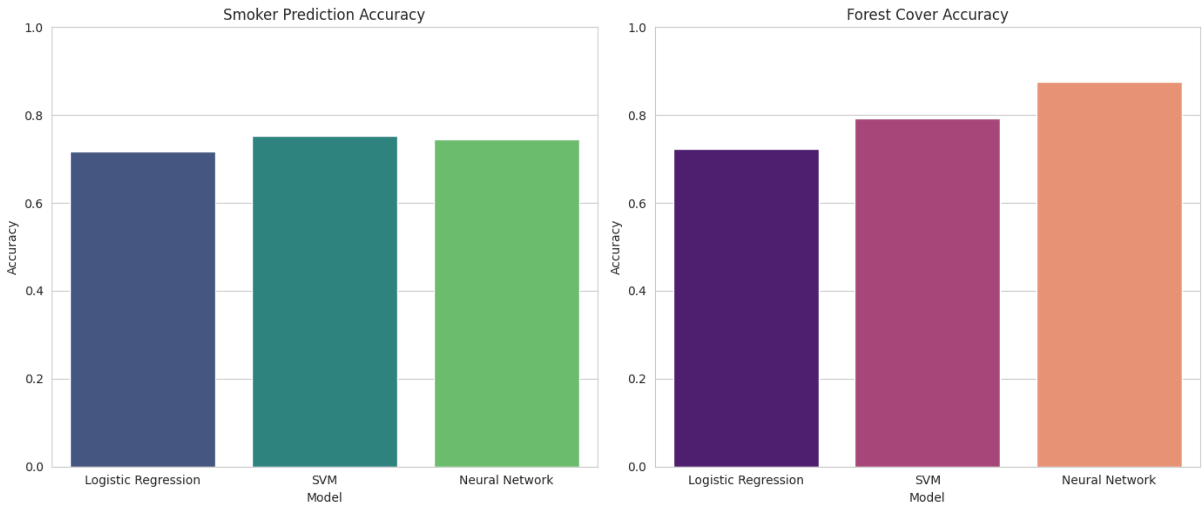

Figure 3: Final Accuracy Comparison: SVM wins Smoker task, Neural Net wins Forest task.

5 Reproducibility

The complete source code is available on GitHub