Report submitted on

Predictive Analytics for Binary and Multiclass Classification

Submitted by

Suchir Okram (MT2025124)

Course Instructor: Prof. Sushree Behera

Course: Machine Learning

Institute: International Institute of Information Technology, Bangalore

Contents

Abstract

This report presents a comparative study of predictive modeling techniques for two supervised learning problems: (i) binary classification of Smoker Status using biometric and lifestyle attributes, and (ii) multiclass classification of Forest Cover Type using cartographic variables. The models used in this work include Logistic Regression, Support Vector Machines (Linear and RBF kernels), and Neural Networks (MLP). All models comply with the syllabus restrictions.

The report includes Exploratory Data Analysis (EDA), duplicate handling strategies, data preprocessing, model training, hyperparameter tuning, and performance evaluation. Results show that while simpler linear models perform reasonably well on both datasets, the RBF-SVM and Neural Network models provide stronger nonlinear decision boundaries, especially in the forest cover task.

1 Overview

The objective of this project is to evaluate classical machine learning models on two datasets with distinct characteristics:

- **Smoker Status Prediction (Binary Classification):** Predict whether a person is a smoker based on biometric features.
- **Forest Cover Type Prediction (Multiclass Classification):** Predict one of seven forest cover types using cartographic data.

Each dataset requires separate EDA, model training, tuning, and evaluation. The project compares the performance of three models across both tasks.

2 Dataset Descriptions

2.1 Smoker Status Prediction Dataset

- Source: Kaggle — Smoker Status Prediction Using Biosignals.
- Task: Binary classification (`smoking` = 0 or 1).
- Features: Biometric measurements (blood pressure, cholesterol, triglycerides, etc.).
- Target: Whether the individual is a smoker.

2.2 Forest Cover Type Dataset

- Source: UCI Machine Learning Repository.
- Task: Multiclass classification with 7 classes.
- Features: 54 cartographic attributes including elevation, slope, soil type.
- Target: Forest cover type (1–7).

3 Exploratory Data Analysis (EDA)

3.1 Smoker Status Dataset

3.1.1 Dataset Overview

The training dataset contains `38984 rows` and `23 columns`, and the test dataset contains `16708 rows` and `22 columns`. The target variable is `smoking`.

3.1.2 Basic Information

- All features correspond to biological measurements (e.g., height, weight, hemoglobin, cholesterol).
- The target variable is balanced to a reasonable extent, allowing stratified splitting for training and validation.
- The absence of string or categorical fields simplifies downstream modeling.

3.1.3 Missing Values

- **No missing values** were present in any feature.
- All columns showed complete data, suggesting prior preprocessing by dataset creators.

3.1.4 Duplicate Detection and Resolution

Duplicate analysis was performed at two levels:

1. **Exact duplicates** — rows where both features and target label were identical.
2. **Feature-duplicate groups** — rows sharing the same feature values but having potentially different labels.

The following observations were made:

- Several groups of feature-identical samples were detected, indicating redundancies likely arising from repeated measurements.
- In some groups, conflicting labels existed (identical biometric data assigned to both smoker and non-smoker categories), indicating label noise.

To address these inconsistencies, a systematic resolution strategy was applied:

- Feature-identical groups with consistent labels were reduced to a single representative instance.
- Groups with conflicting labels were resolved by **majority voting** to choose the most probable label.
- Groups where a clear majority did not exist were removed entirely to avoid injecting noise into the training process.

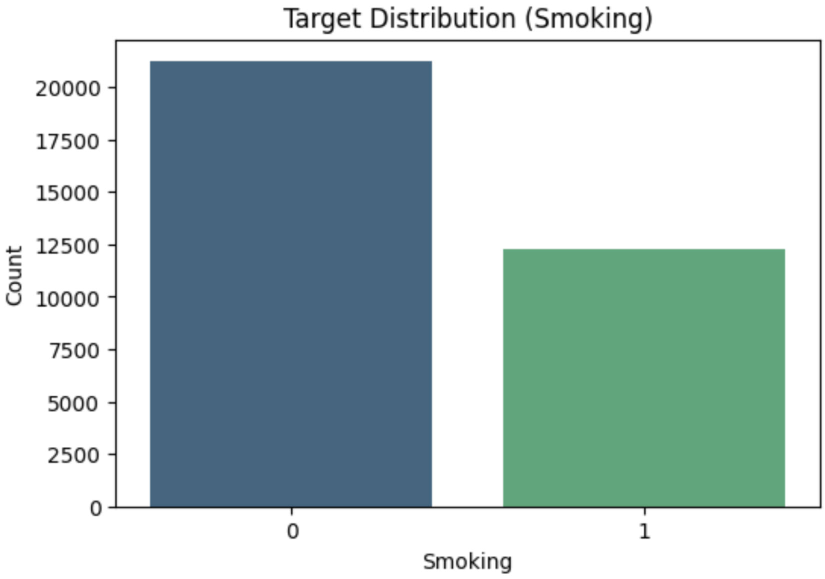

Figure 1: Distribution of Smoker and Nonsmoker

3.1.5 Target Variable Distribution

The target variable `smoking` is distributed as shown in Figure 1.

3.1.6 Numerical Features and Correlation

Correlation heatmap for features are shown in Figure 2.

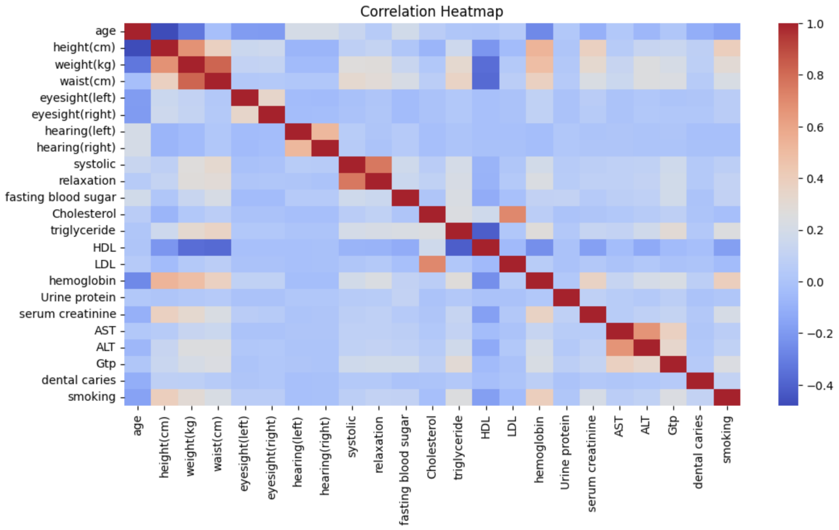

Figure 2: Correlation Matrix of all Features

3.1.7 Feature–Target Relationship

The relationship between all features and the target variable was visualized using scatter plots (e.g. Figure 3, Figure 4, Figure 5, Figure 6, Figure 7).

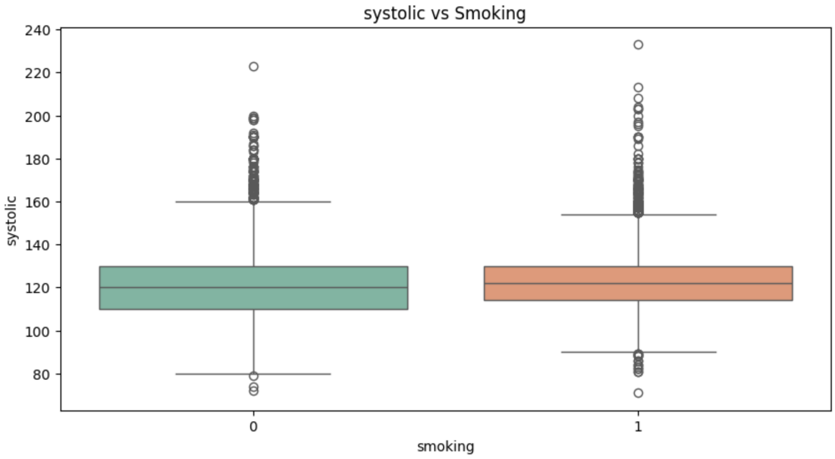

Figure 3: systolic Vs Smoking

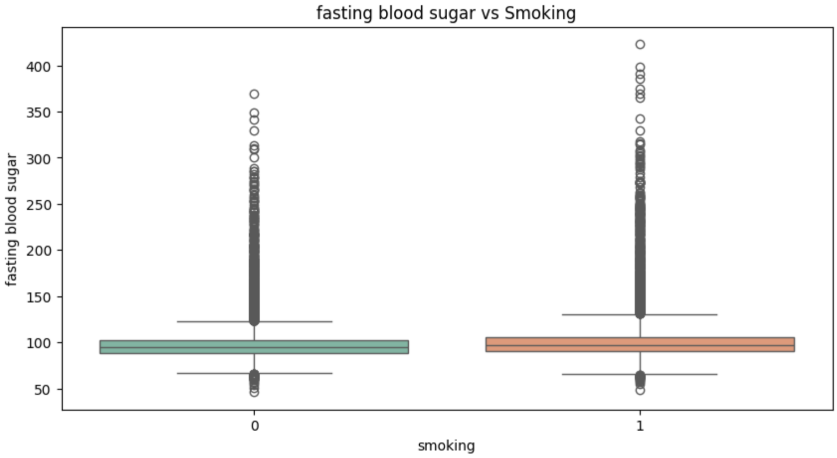

Figure 4: fasting blood sugar Vs Smoking

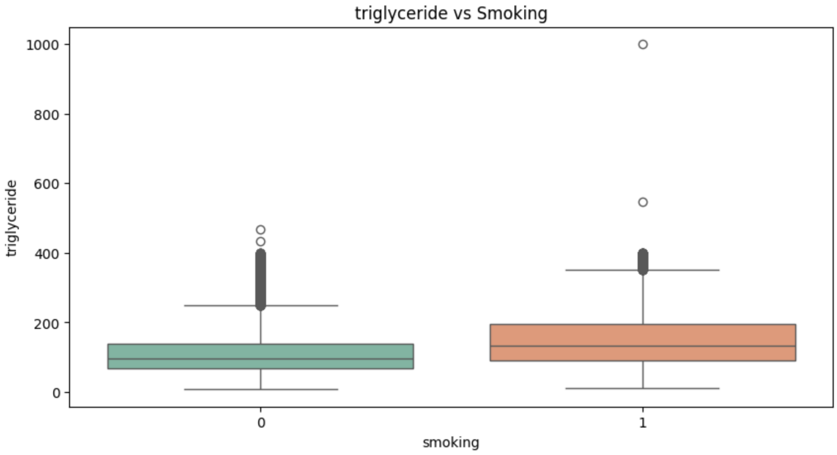

Figure 5: triglyceride Vs Smoking

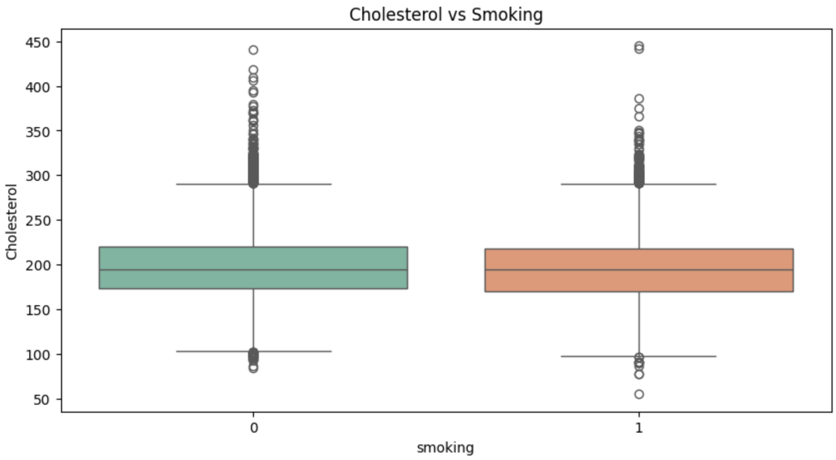

Figure 6: Cholesterol Vs Smoking

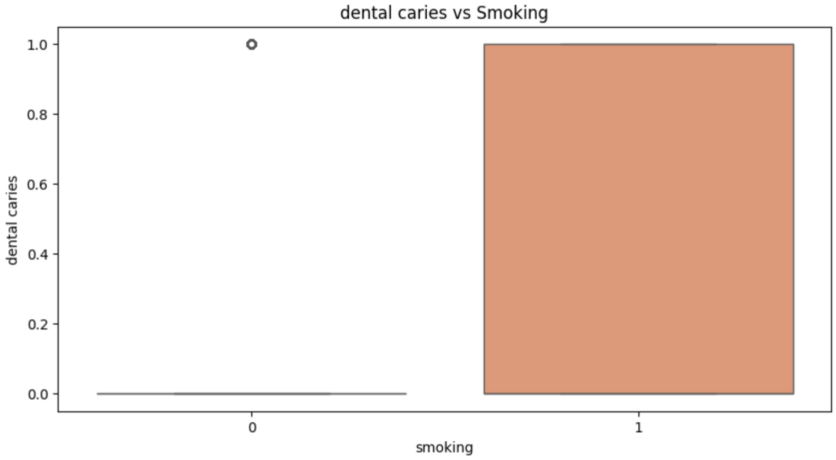

Figure 7: dental caries Vs Smoking

3.2 Forest Cover Dataset

3.2.1 Dataset Overview

The dataset comprises **581,012 rows** and **54 columns**, making it a large-scale and high-dimensional dataset suitable for evaluating model performance in multiclass classification settings. The target variable `Cover_Type` spans seven discrete classes corresponding to forest cover categories.

3.2.2 Basic Information

- The dataset consists of **only numerical features**, with the exception of one-hot encoded binary variables for soil type and wilderness area.
- A total of **40 binary soil indicators** and **4 wilderness area flags** contribute to a sparse but structured feature space.
- The remaining numeric features (elevation, aspect, slope, hydrology distances) are continuous measurements.

- The target label distribution is moderately imbalanced, with Cover Type 2 and 1 having the highest frequencies.

3.2.3 Missing Values

- **No missing entries** in any column.
- All binary indicator features are strictly 0 or 1, as expected.
- All continuous variables contain valid numerical entries with no anomalies or null markers.

3.2.4 Target Variable Distribution

The target variable Cover_Type is distributed as shown in Figure 8.

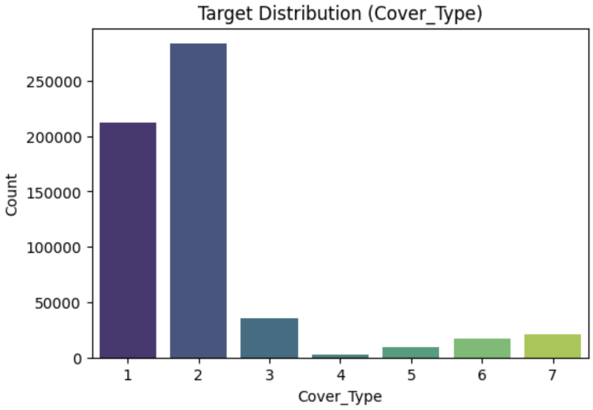

Figure 8: Distribution of Cover_Type

3.2.5 Numerical Features and Correlation

Correlation heatmap for features are shown in Figure 9.

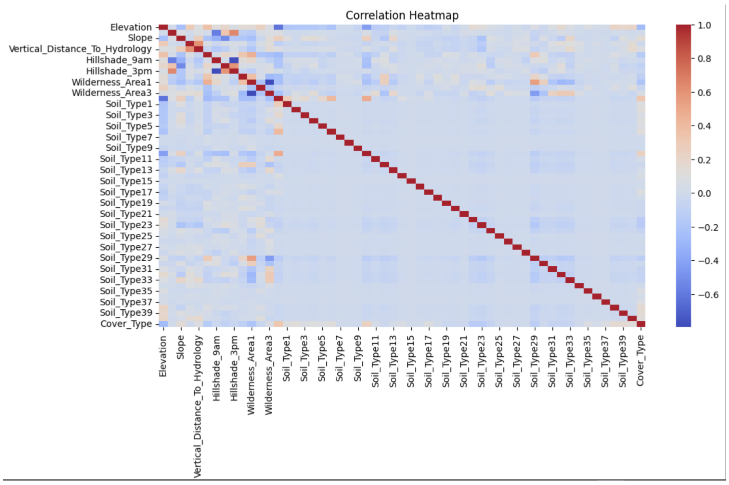

Figure 9: Correlation matrix of all Features

3.2.6 Feature–Target Relationship

The relationship between all features and the target variable was visualized using scatter plots (e.g. Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, Figure 15).

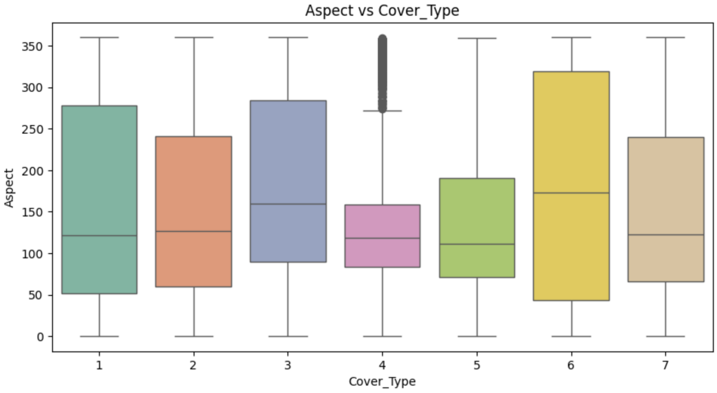

Figure 10: Aspect Vs Cover_Type

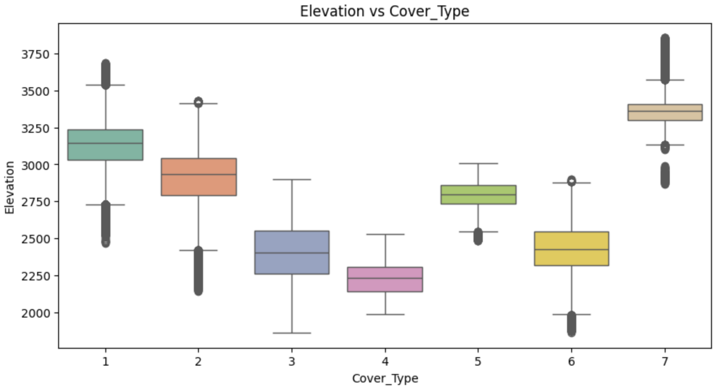

Figure 11: Elevation Vs Cover_Type

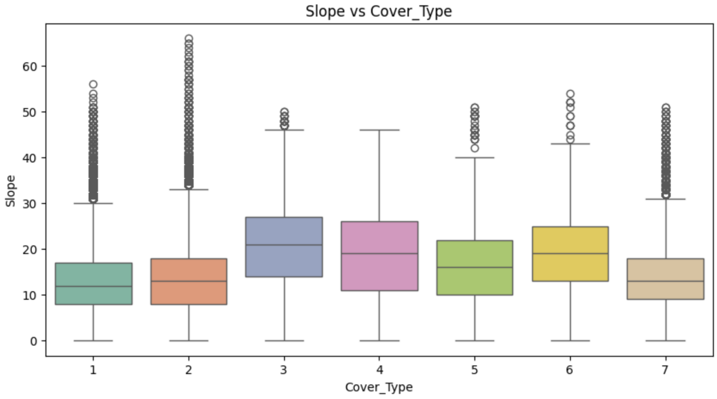

Figure 12: Slope Vs Cover_Type

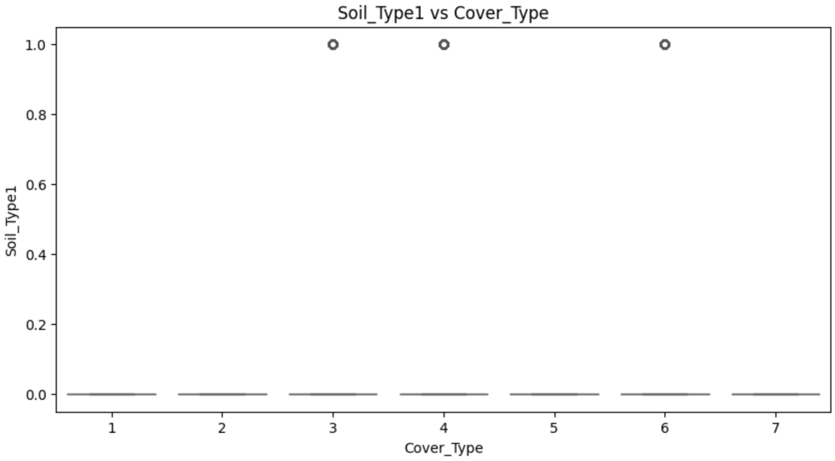

Figure 13: Soil_Type1 Vs Cover_Type

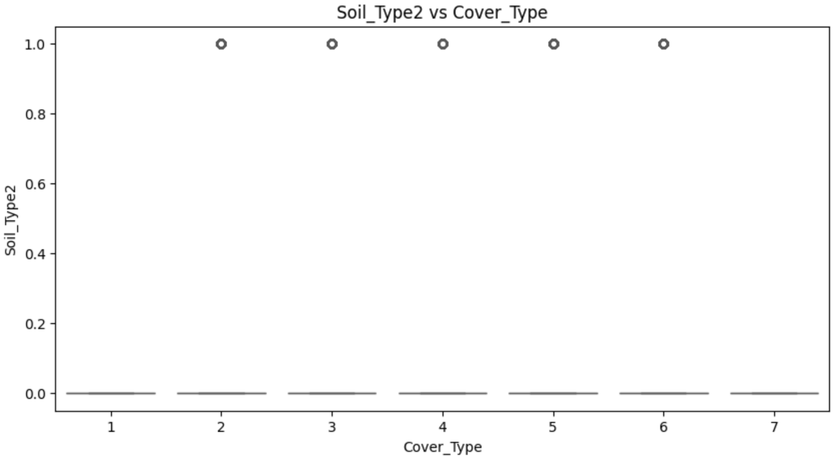

Figure 14: Soil_Type2 Vs Cover_Type

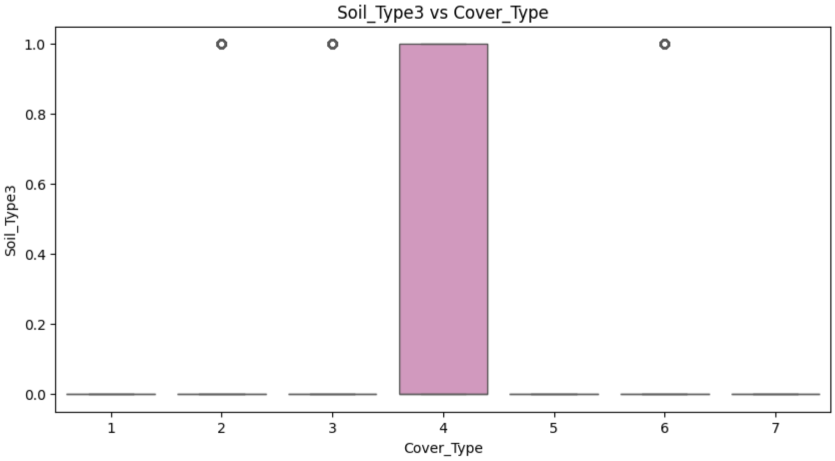

Figure 15: Soil_Type3 Vs Cover_Type

4 Data Processing Steps

4.1 Duplicate Handling (Smoker Dataset)

Two types of duplicates were examined:

- exact duplicates where both features and labels were identical, and
- feature-level duplicates where samples shared identical predictors but had differing labels.

Exact duplicates were reduced to a single representative instance to avoid redundant training examples. For feature-level duplicates with conflicting labels, a majority-vote strategy was applied to infer the most likely class. In cases where no clear majority existed, the entire group was removed to prevent label noise from degrading model performance.

4.2 Feature Scaling

All numerical features were standardized using the `StandardScaler`, which transforms each variable to have zero mean and unit variance. Feature scaling is essential for models such as Support Vector Machines, Logistic Regression, and Neural Networks, which are sensitive to variable magnitude.

4.3 Train–Validation Split

An stratified train–validation 80–20 split ratio was used throughout the experiments, providing sufficient data for both effective training and reliable evaluation while preventing overfitting to a small validation set.

4.4 Pipeline Summary

The data preprocessing workflow consists of four key steps executed in sequence:

1. **Input Cleaning:** Detection and removal or correction of duplicates (only applicable to the Smoker dataset).
2. **Feature Scaling:** Standardization of all numerical predictors using a fitted `StandardScaler`.
3. **Dataset Splitting:** Stratified partitioning into training and validation sets.
4. **Model Training:** Applying Logistic Regression, SVM, or Neural Network models to the processed data, followed by hyperparameter tuning.

5 Models Used

The models used for are Logistic Regression, Support Vector Machines (Linear and RBF kernels) and Neural Networks (MLP), Among which, **Neural Networks** achieved the highest accuracy of **0.75** for smoker dataset and **0.90** for forest cover dataset.

6 Hyperparameter Tuning

Table 1: Best Hyperparameters for All Models Across Both Datasets

Dataset	Model	Best Hyperparameters
Smoker Status (Binary)	Logistic Regression	C = 10.0, Solver = `liblinear`, Max Iter = 5000
	SVM (RBF)	C = 1.0, γ = `scale`, Kernel = RBF, Class Weight = Balanced
	Neural Network (MLP)	Hidden Layer Sizes = (32, 16), Activation = ReLU, α = 0.001, Learning Rate = Adaptive, Early Stopping = True
Forest Cover (Multiclass)	Logistic Regression	C = 3.5, Solver = `lbfgs`, Multi-class = OvR
	SVM (Linear)	C = 1.0, Loss = `squared_hinge`, Max Iter = 2000
	Neural Network (MLP)	Hidden Layer Sizes = (128, 64), Activation = ReLU, α = 0.0005, Learning Rate = Adaptive, Early Stopping = True

7 Performance and Evaluation

This section presents the performance of all trained models on the validation sets for both the Smoker Status (binary) and Forest Cover Type (multiclass) classification tasks.

7.1 Smoker Dataset Results

The model was trained using Adaptive Learning and achieved a validation accuracy of **0.747386**.

Classification Report

```
              precision    recall  f1-score   support

           0       0.79      0.82      0.81      4242
           1       0.67      0.62      0.64      2452

    accuracy                           0.75      6694
```

Insights

- The MLP model performs strongly for non-smokers (0), achieving high precision and recall, but struggles with the smoker class(1), where both precision and recall are noticeably lower.
- Overall accuracy is good, but class-wise performance highlights the need for improved detection of smokers.

Model comparison table

Table 2: Model comparison for smoker dataset

Model	Validation Accuracy
Logistic Regression	0.726023
SVM	0.723484
Neural Networks	0.747386

7.2 Forest Cover Dataset Results

The model was trained using Adaptive Learning and achieved a validation accuracy of **0.904598**.

Classification Report

```
              precision    recall  f1-score   support

           1       0.90      0.91      0.90     42368
           2       0.92      0.92      0.92     56661
           3       0.86      0.94      0.90      7151
           4       0.81      0.81      0.81       549
           5       0.80      0.69      0.74      1899
           6       0.83      0.76      0.80      3473
           7       0.96      0.87      0.91      4102

    accuracy                           0.90    116203
```

Insights

- The MLP model achieves strong overall accuracy with consistently high precision and recall for the major classes (1–3 and 7), indicating excellent generalization on dominant forest types.
- Performance drops for minority classes, especially classes 4, 5, and 6, where lower recall suggests difficulty in detecting rare cover types.

Model comparison table

Table 3: Model comparison for smoker dataset

Model	Validation Accuracy
Logistic Regression	0.714181
SVM	0.711419
Neural Networks	0.904598

8 Discussion

- Across both datasets, the models exhibited different generalization behaviors. Mild overfitting was observed in the Smoker dataset, particularly for nonlinear models, while the large size of the Forest Cover dataset naturally reduced overfitting and enabled more stable learning.
- Duplicate handling in the Smoker dataset improved label consistency and reduced noise, leading to more reliable training outcomes.
- Model performance varied notably across tasks: linear models worked reasonably well for the binary problem but were insufficient for the complex multiclass forest dataset, where MLP clearly outperformed.
- From a practical standpoint, MLPs provided higher accuracy but required significantly more computation.

9 GitHub Repository

All code, notebooks, and submission files are available at:
`https://github.com/Su-ok/MT2025124_ML_Project2`

10 Conclusion

This project successfully demonstrates preprocessing, EDA, model tuning, and evaluation.
Future improvements include exploring advanced regularization strategies, adjusting decision thresholds for better minority-class detection, and experimenting with deeper neural architectures.