Machine Learning Project 2 Dataset 1

Smoker Status Prediction Report

Team Name: MT2025065
Member: Keyur Sanjaykumar Padiya
Course: Machine Learning
Date: December 12, 2025

Contents

List of Figures

List of Tables

1 Introduction

The objective of this project is to analyze a bio-signal dataset to predict the smoker status of individuals. By performing Exploratory Data Analysis (EDA) and preprocessing the data, we aim to build a foundation for machine learning modeling. The dataset includes various physiological and biochemical markers, such as height, weight, eyesight, and serum levels (e.g., cholesterol, hemoglobin), which are critical for understanding the health impact of smoking.

2 Data Processing

2.1 Exploratory Data Analysis (EDA)

An extensive initial analysis was conducted to understand the data structure, quality, and distributions. The analysis confirmed that the dataset is robust, with a substantial number of samples and no missing values.

2.1.1 Initial Analysis

The dataset is divided into training and testing sets. The training set comprises 38,984 samples with 23 columns, while the test set contains 16,708 samples. The features are predominantly numerical, consisting of 22 numerical attributes and 0 categorical attributes (excluding the target).

Attribute	**Count**
Total Samples (Train)	38,984
Total Features	22
Missing (Null) Values	0
Duplicate Rows (Removed)	5,517

Table 1: High-Level Data Summary

A check for missing values confirmed that there are zero null values across both training and test datasets, indicating high data quality. However, 5,517 duplicate rows were identified and removed to prevent model bias and ensure unbiased evaluation.

2.1.2 Target Variable Analysis

The target variable, `smoking`, is a binary classification target (0 for non-smoker, 1 for smoker). Analysis of the class distribution reveals a class imbalance.

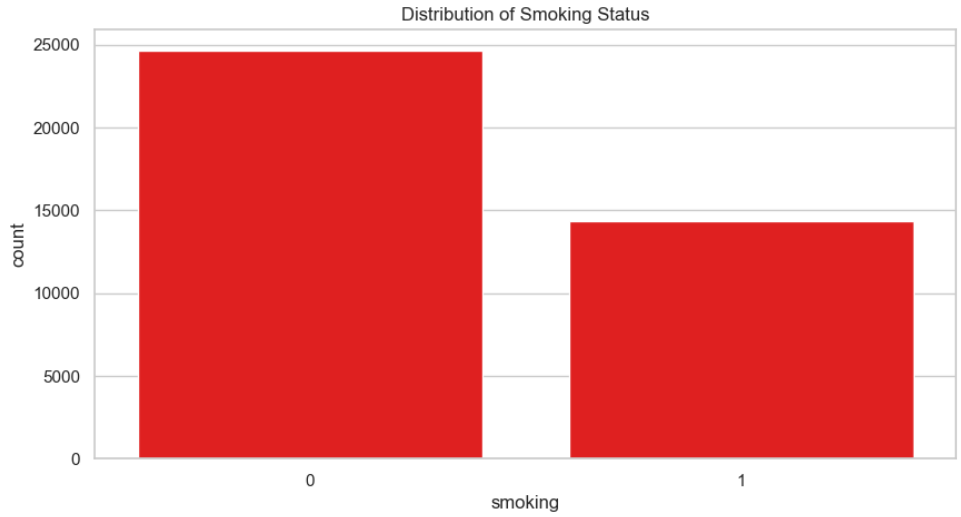

Figure 1: Distribution of the Target Variable (Smoking Status)

As shown in Figure 1, Non-smokers constitute approximately 63.3% of the data, while smokers make up 36.7%. The minority-to-majority class ratio is 0.580. While this imbalance is not extreme, it suggests that stratified splitting techniques should be employed during model training to maintain representative class proportions.

2.1.3 Numerical Feature Analysis

We analyzed the distributions of all 22 numerical features. To ensure clarity, the features are visualized across four separate figures, following their order in the dataset.

Group 1: Demographics and Physical Attributes Figure 2 displays the distributions for `age`, `height`, `weight`, `waist`, and `eyesight` (left/right).

- **Physical Stats:** `height`, `weight`, and `waist` exhibit roughly normal distributions. `Waist` shows a slight skew to the right, indicating a subset of individuals with higher abdominal measurements.
- **Eyesight:** Both left and right eyesight features are heavily right-skewed. The distribution peaks sharply at standard vision values (around 1.0-1.2) with a long tail extending towards higher values, representing individuals with poorer vision.

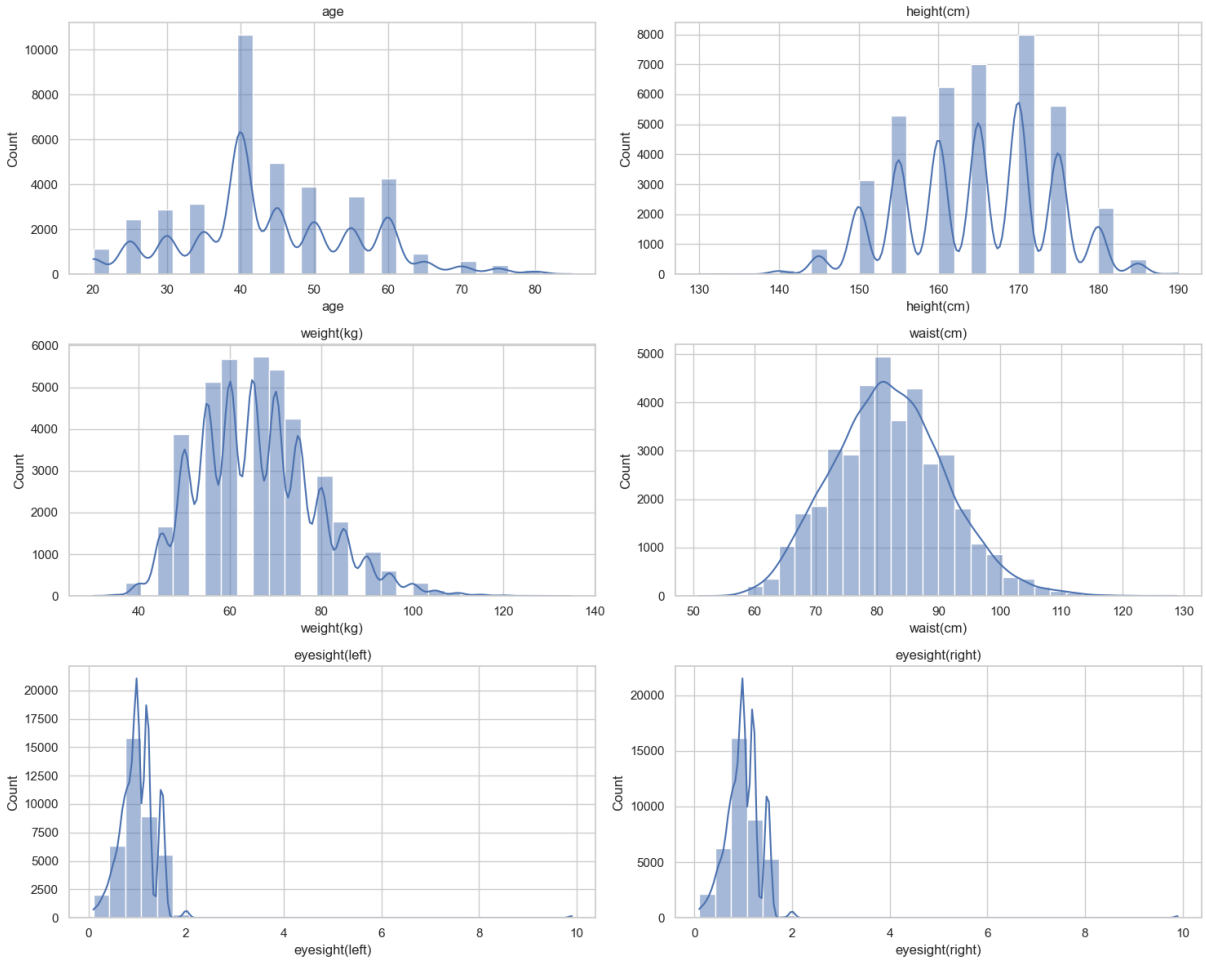

Figure 2: Distribution of Group 1 (Age, Height, Weight, Waist, Eyesight)

Group 2: Sensory and Blood Pressure Figure 3 covers `hearing` (left/right), blood pressure (`systolic`, `relaxation`), `fasting blood sugar`, and `cholesterol`.

- **Hearing:** Both hearing features appear as discrete bars rather than continuous curves, indicating they are likely binary or categorical variables (e.g., 1 for normal, 2 for impaired).
- **Blood Pressure:** `systolic` and `relaxation` follow a classic normal distribution (bell curve).
- **Metabolic:** `fasting blood sugar` is extremely right-skewed, while `Cholesterol` follows a relatively normal distribution centered around the population mean.

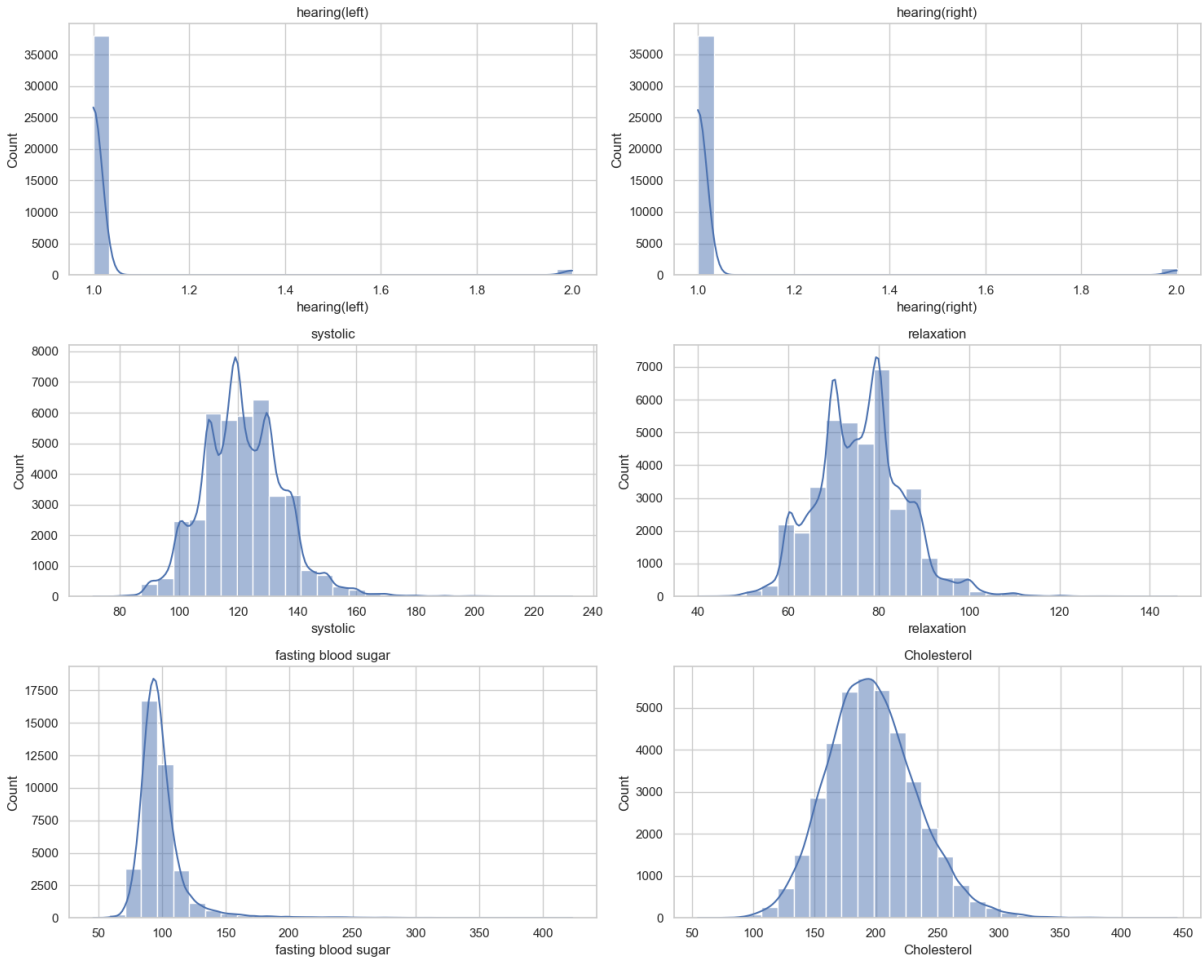

Figure 3: Distribution of Group 2 (Hearing, Blood Pressure, Sugar, Cholesterol)

Group 3: Lipids and Kidney Function Figure 4 analyzes `triglyceride`, `HDL`, `LDL`, `hemoglobin`, `urine protein`, and `serum creatinine`.

- **Lipids & Hemoglobin:** `HDL`, `LDL`, and `hemoglobin` display normal distributions. `Triglyceride` is notably right-skewed, suggesting the presence of high-value outliers.
- **Kidney Markers:** `Urine protein` is highly skewed, with the vast majority of samples clustered at the lowest value. `Serum creatinine` also shows a right skew, typical for this biological marker in a general population.

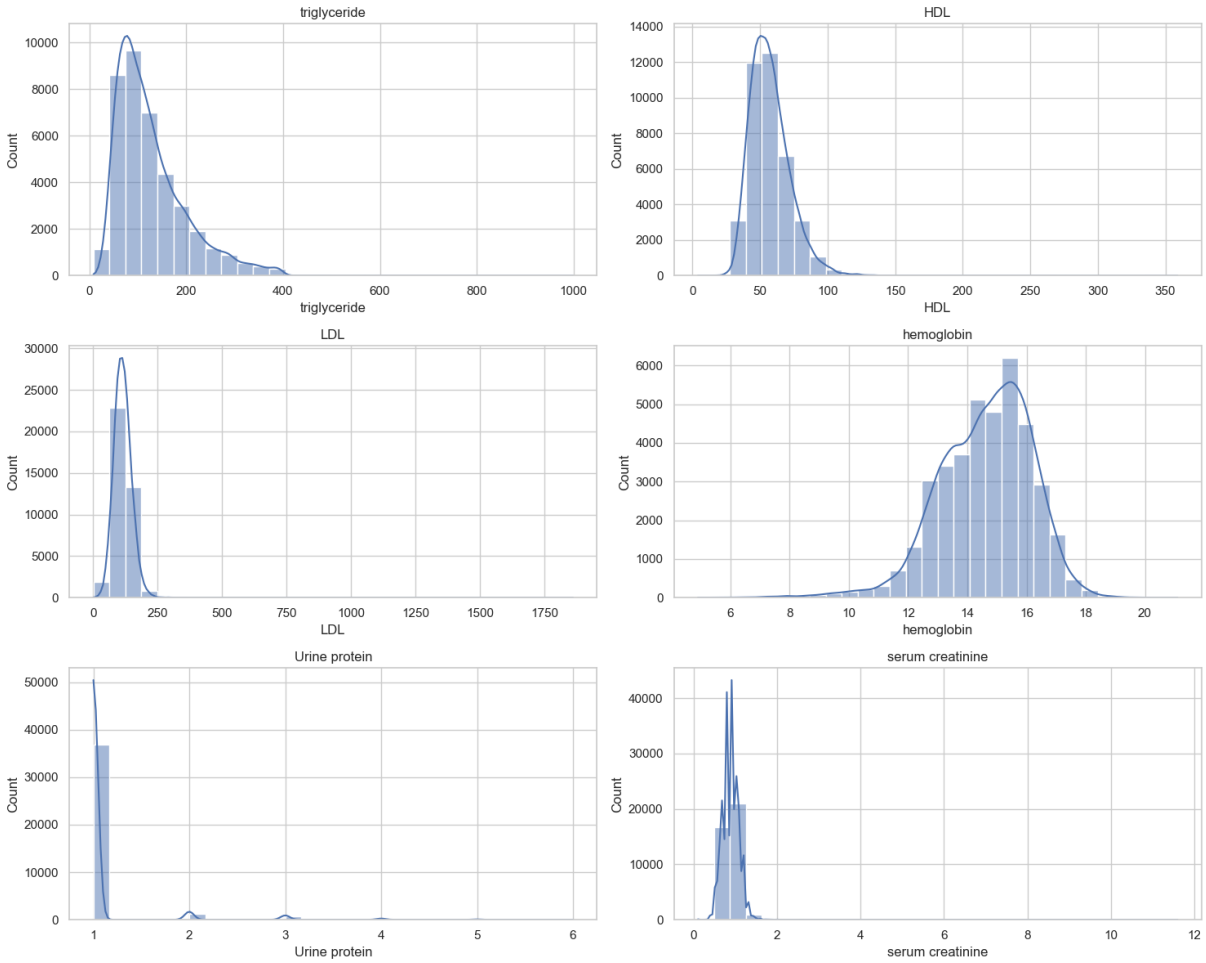

Figure 4: Distribution of Group 3 (Lipids, Hemoglobin, Kidney Markers)

Group 4: Liver Enzymes and Dental Health Figure 5 shows the remaining features: `AST`, `ALT`, `Gtp`, and `dental caries`.

- **Liver Enzymes:** All three liver function markers (`AST`, `ALT`, `Gtp`) are heavily right-skewed. This indicates that while most individuals have low enzyme levels, there is a significant tail of individuals with elevated levels, which is often correlated with smoking or alcohol consumption.
- **Dental Caries:** This feature displays a binary distribution (0 or 1), representing the presence or absence of cavities.

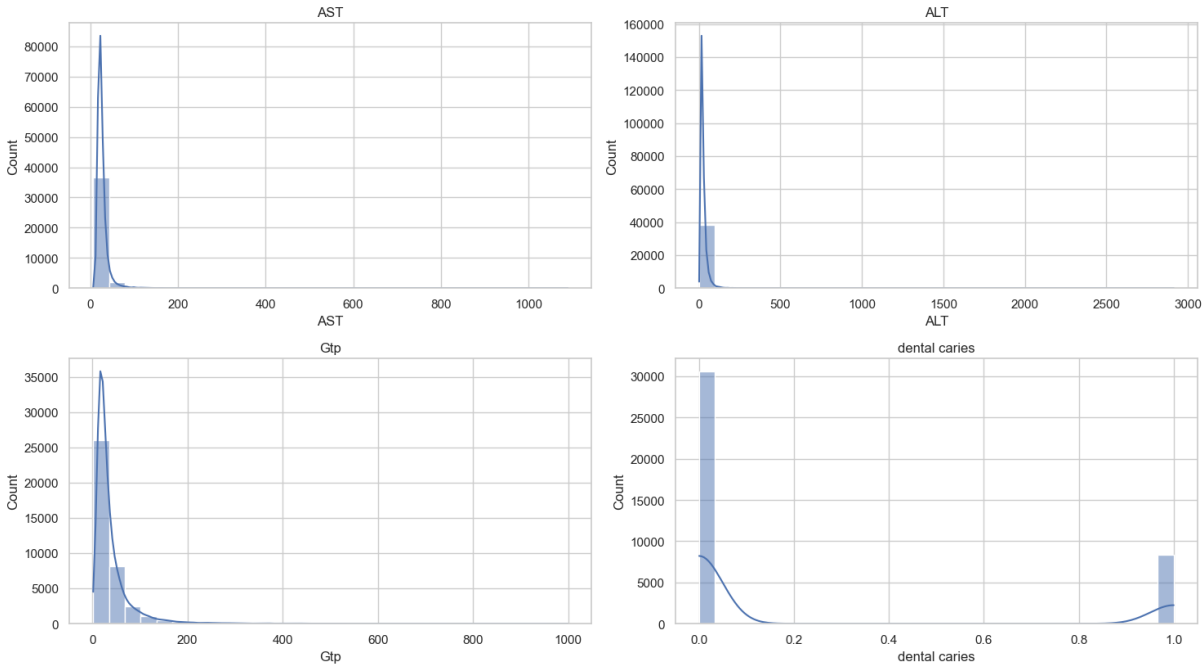

Figure 5: Distribution of Group 4 (Liver Enzymes, Dental Caries)

2.1.4 Outlier Analysis

Specific attention was given to the `triglyceride` feature to evaluate outlier detection methods. We compared Z-score and Interquartile Range (IQR) methods.

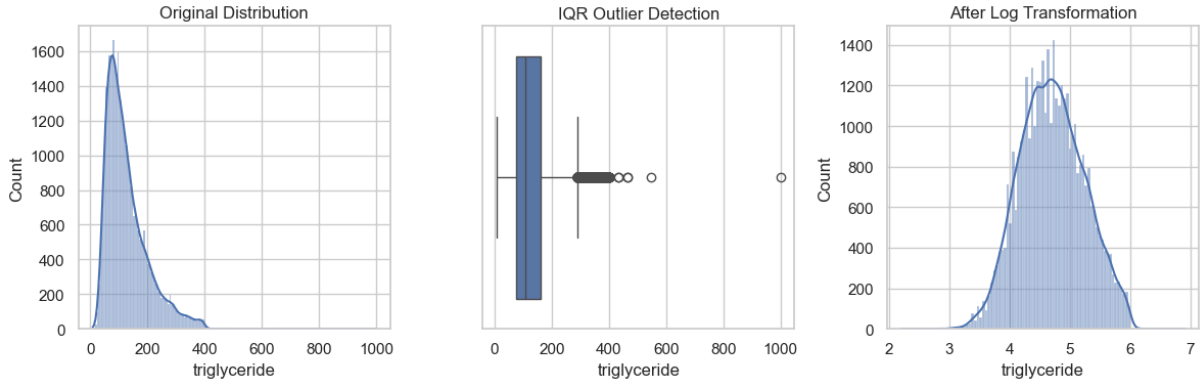

Figure 6: Outlier Analysis on Triglyceride Feature

As illustrated in Figure 6, the Z-score method identified 641 outliers, while the IQR method identified 1,607. Given the biological nature of the data, "extreme" values might be genuine health indicators rather than errors. Therefore, Robust Scaler (which uses median and IQR) was selected for preprocessing to handle these outliers effectively without discarding valuable data.

2.1.5 Feature Importance and Correlation

A correlation heatmap was generated to identify linear relationships between features and the target variable.

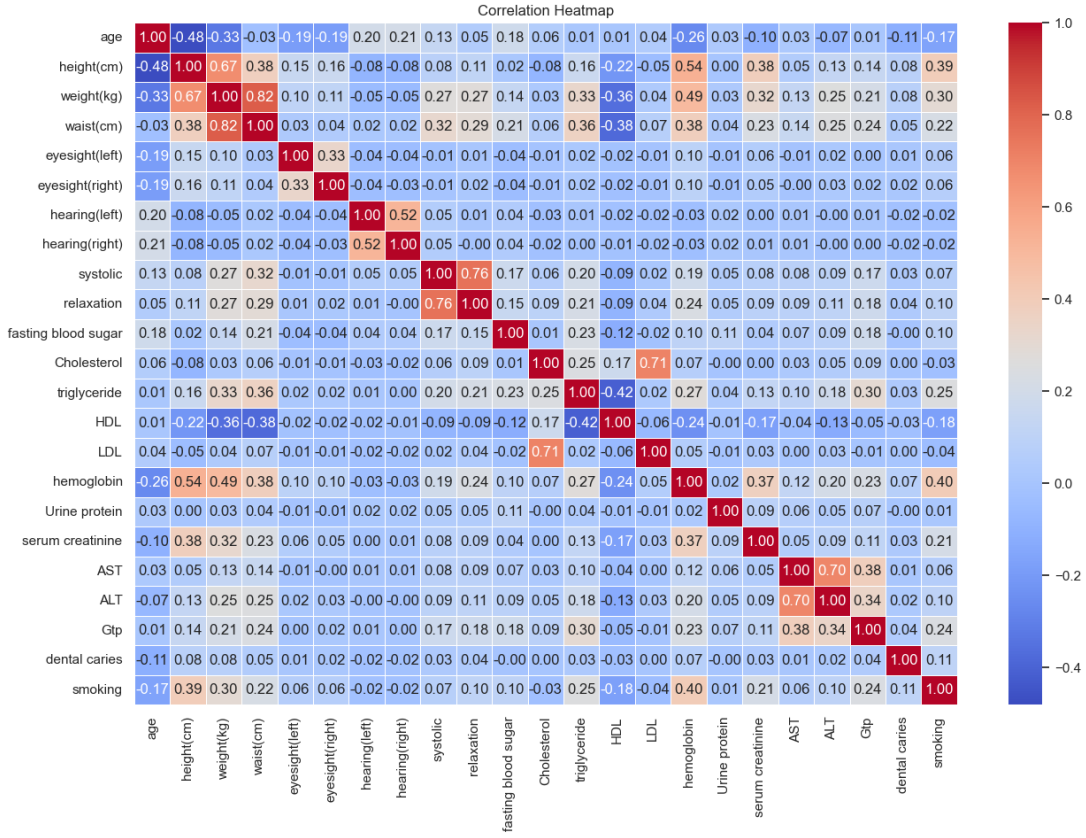

Figure 7: Correlation Heatmap

Figure 7 highlights several key insights:

- **Top Predictors:** `hemoglobin` (0.40), `height` (0.39), and `weight` (0.30) show the strongest positive correlation with smoking status.
- **Multicollinearity:** Strong correlations exist between `weight` and `waist(cm)`, as well as `systolic` and `relaxation` (blood pressure metrics). Tree-based ensembles are generally capable of handling this multicollinearity.
- `Gtp` and `triglyceride` also show notable correlations, reinforcing their potential predictive power.

2.1.6 Target Value Analysis

To validate the predictive power of specific bio-signals, boxplots were created to visualize the separation between smokers and non-smokers.

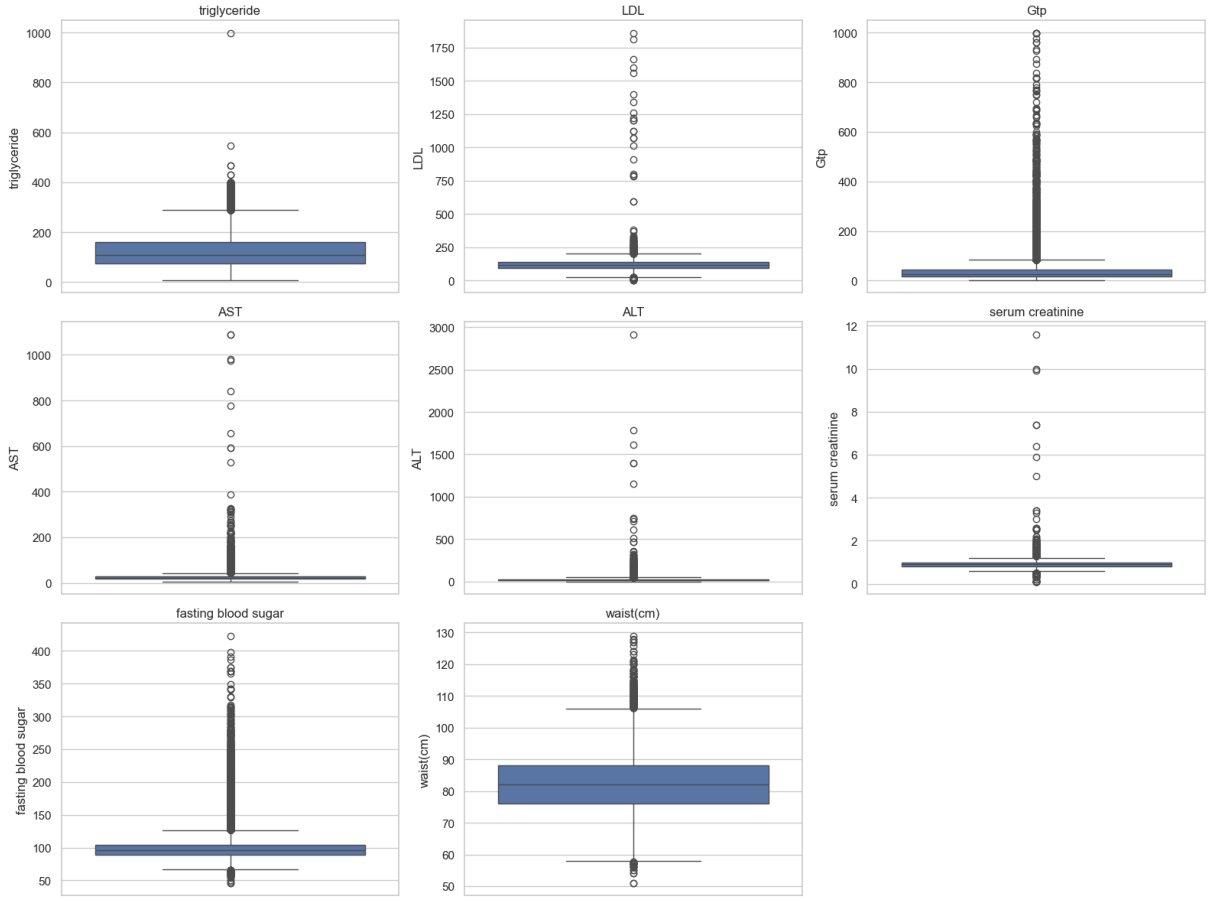

Figure 8: Bio-signal Distributions by Smoking Status

Figure 8 demonstrates clear distinctions in the central tendency for features like `hemoglobin` and `Gtp` between the two classes. Smokers tend to have higher median values for these indicators, confirming their relevance as strong predictors for the machine learning model.

2.2 Data Preprocessing Pipeline

Data preprocessing is a critical step to ensure the model receives clean, normalized, and structured data. Based on the findings from the EDA, we implemented a pipeline involving data cleaning, feature type handling, robust scaling, and stratified splitting.

2.2.1 Data Loading and Exploration

The initial step involved loading the raw training and test datasets. We inspected the data for integrity issues such as missing values and duplicates.

Metric	**Value**
Original Training Samples	38,984
Test Samples	16,708
Missing Values Detected	0
Duplicate Rows Identified	5,517
Action Taken	**Duplicates Removed**
Final Training Samples	33,467

Table 2: Data Cleaning Summary

As detailed in Table 2, while the dataset contained no missing values, a significant number of duplicate rows (5,517) were found. These were removed to ensure the model learns generalized patterns rather than memorizing repeated samples.

2.2.2 Handle Categorical Features

We explicitly checked for categorical variables (data type 'object') to determine if One-Hot Encoding or Label Encoding was necessary.

- **Identified Categorical Features:** 0 (None)
- **Action:** No encoding required.

The dataset consists entirely of numerical bio-signals, simplifying the pipeline as no complex text processing or cardinality reduction was needed.

2.2.3 Handle Numerical Features

Given the presence of significant outliers identified in Section 2.1.4 (specifically in `triglyceride` and `Gtp`), standard mean-variance scaling (`StandardScaler`) is less suitable, as it can be heavily influenced by extreme values. Instead, we employed `RobustScaler`.

This scaling technique utilizes the median and the Interquartile Range (IQR) to scale the data, rather than the mean and standard deviation. By focusing on the central 50% of the data, `RobustScaler` ensures that the outliers found in the bio-signal data do not distort the scaling of the majority of the samples. This preserves the integrity of the feature distributions, allowing the model to learn effective patterns without being biased by extreme physiological measurements.

2.2.4 Train-Validation Split

To evaluate the model effectively, the cleaned training data was split into a training set and a validation set. We used Stratified Sampling to preserve the class imbalance ratio (approx. 63:37) in both subsets.

Dataset	**Samples**	**Features**	**Role**
Training Set (X_{train})	26,773	22	Model Fitting
Validation Set (X_{val})	6,694	22	Hyperparameter Tuning
Test Set (Unseen)	16,708	22	Final Evaluation

Table 3: Final Dataset Split Dimensions

Table 3 summarizes the final shapes of the data fed into the model. This 80/20 split provides a substantial amount of data (over 26k samples) for training while reserving a statistically significant portion (over 6.6k samples) for validation.

3 Models Used

To address the binary classification task of predicting smoker status, we selected three distinct machine learning algorithms. Each model was chosen for its unique properties and capability to handle the tabular bio-signal data.

3.1 Logistic Regression

Logistic Regression was selected as the baseline model due to its simplicity, interpretability, and efficiency. It models the probability of the target variable as a function of the features using the logistic function. This model serves as a benchmark to determine if complex non-linear models offer significant improvements over a linear decision boundary.

3.2 Neural Network (MLP Classifier)

We implemented a Multi-Layer Perceptron (MLP), a type of Artificial Neural Network. Neural networks are capable of capturing complex, non-linear interactions between features through multiple layers of neurons and non-linear activation functions. Given the biological nature of the data, where factors like liver enzymes and blood sugar may interact in complex ways, an MLP is well-suited to learn these latent patterns.

3.3 Support Vector Machine (SVM)

The Support Vector Machine was chosen for its effectiveness in high-dimensional spaces and its ability to create complex decision boundaries using kernel functions. By mapping the input data into a higher-dimensional feature space using the Radial Basis Function (RBF) kernel, the SVM can handle non-linear separations that Logistic Regression might miss.

4 Hyperparameter Tuning

Extensive hyperparameter tuning was conducted for each model to maximize performance. We utilized both `GridSearchCV` for exhaustive search and `Optuna` for efficient Bayesian optimization.

4.1 Logistic Regression Tuning

The initial baseline accuracy was approximately 71.8%. We performed multiple rounds of tuning:

1. **Grid Search:** We exhaustively tested solvers (`liblinear`, `lbfgs`) and penalties (`l1`, `l2`).
2. **Optuna Optimization:** We expanded the search space for the regularization parameter C (using a log-uniform distribution) and tolerance.
3. **Data Transformation:** A significant breakthrough occurred when we applied Log Transformation to skewed features (`triglyceride`, `Gtp`, etc.) combined with `RobustScaler`.

Parameter	Optimal Value
Solver	`liblinear`
Penalty	`l1`
C (Inverse Regularization)	0.0650
Max Iterations	200
Tolerance	0.0022

Table 4: Best Hyperparameters for Logistic Regression

This final configuration significantly improved the accuracy to **73.50%**.

4.2 Support Vector Machine (SVM) Tuning

SVM training is computationally expensive, so we employed a multi-stage Optuna search strategy.

1. **Initial Search:** Tuned C and γ on the full dataset.
2. **Refined Search:** To speed up convergence, we trained on subsets (40% and 60%) of the data to narrow down the parameter space for C (range 0.5 to 50) and γ (range 1e-4 to 0.05).
3. **Final Optimization:** Included additional parameters like `shrinking` and `decision_function_sh`

Parameter	Optimal Value
Kernel	`rbf`
C	8.8986
Gamma	0.0144
Tolerance	0.0037
Shrinking	`True`

Table 5: Best Hyperparameters for SVM

This rigorous tuning process resulted in a validation accuracy of **74.69%**.

4.3 Neural Network Tuning

The Neural Network tuning focused on the architecture (number of hidden layers and neurons) and the solver.

- **Architecture Search:** We tested various configurations, including (128), (128, 64), and (128, 64, 32). Deeper networks did not necessarily yield better results, suggesting the dataset complexity could be captured by fewer layers.
- **Solver & Activation:** The `adam` solver with `logistic` (sigmoid) activation function provided the most stable convergence.

Parameter	Optimal Value
Hidden Layer Sizes	(128, 64)
Activation	`logistic`
Solver	`adam`
Alpha (L2 Penalty)	0.0001
Learning Rate	`constant`

Table 6: Best Hyperparameters for Neural Network

The optimal configuration achieved a validation accuracy of **75.71%**, outperforming the baseline.

5 Performance Evaluation

After extensive preprocessing and hyperparameter tuning, we compared the best-performing configurations for each of the three models. The Neural Network achieved the highest validation accuracy, demonstrating its ability to effectively model the non-linear relationships in the bio-signal data.

Model	**Best Validation Accuracy (%)**
Logistic Regression	73.50
Support Vector Machine (SVM)	74.69
Neural Network (MLP)	**75.71**

Table 7: Performance Comparison Across All Models

While Logistic Regression provided a solid baseline, the non-linear models (SVM and NN) showed a clear advantage. The Neural Network's superior performance suggests that the interactions between physiological features like GTP, Hemoglobin, and Triglycerides are best captured by a multi-layer perceptron architecture.

6 Conclusion

This study successfully demonstrated the efficacy of machine learning in predicting smoker status from bio-signal data, emphasizing the critical role of domain-aware preprocessing. By addressing significant outliers and skewed distributions through *RobustScaler* and log transformations, we established a high-quality dataset for modeling. Among the algorithms evaluated, the **Neural Network** emerged as the superior model, marginally outperforming the Support Vector Machine and the baseline Logistic Regression. These results confirm that while linear baselines are effective, capturing non-linear biological interactions via advanced architectures like neural networks yields the most predictive power for complex medical datasets.

Overall, this project highlights the critical role of domain-aware preprocessing and model selection in analyzing medical datasets.

7 GitHub Repository Link

The complete source code, including data preprocessing scripts, model training notebooks, and analysis reports, is available at:

`https://github.com/KeyPad717/smoke_status_prediction`