

# Data Analysis Project Report

## Predicting Smoking Status using Bio-signals

BOUZID Atef

Supervised by: Mr. MESSAOUD Amor

2023-2024

### 1 Introduction

This report details a data analysis project aimed at predicting an individual's smoking status using bio-signals and other relevant factors. Smoking is a major public health concern, contributing significantly to preventable diseases and deaths globally. This project seeks to develop a machine learning model that can aid healthcare professionals and individuals in better understanding the likelihood of successful smoking cessation.

The dataset used in this analysis can be found at: [link](#)

### 2 Problem statement

Smoking is a well-established cause of various health issues and is a leading contributor to preventable diseases and deaths worldwide. It is projected that smoking-related deaths will reach 10 million by 2030. Efforts have been made to help people quit smoking, but success rates are relatively low, partly due to the complexity of factors influencing smoking cessation.

To enhance the effectiveness of smoking cessation strategies, we propose using machine learning to predict smoking status more accurately. Our objective is to develop a model that can predict an individual's likelihood of smoking using bio-signals. This model would consider a range of factors, including nicotine dependence, carbon monoxide levels, daily cigarette consumption, age of smoking initiation, previous quit attempts, emotional well-being, personality traits, and motivation to quit. By creating such a predictive model, healthcare professionals and individuals seeking to quit smoking can gain insights into the probability of successful smoking cessation. This approach shows promise in improving smoking cessation outcomes.

### 3 Data Description

The project utilizes a dataset containing information on 159,256 individuals, including bio-signals, demographic details, and smoking status. The dataset comprises 24 features, all of which are numerical. The target variable, "smoking," is binary, indicating whether an individual is a smoker (1) or not (0).

- Data Types:
  - Float64: 6 (waist, eyesight (left), eyesight (right), hemoglobin, serum creatinine)
  - Int64: 18 (All other variables including the target 'smoking')
- Missing Values: None (All variables have 0 missing values)

### 3.1 Features Overview

- **id**: Unique identifier for each data point
- **age**: Age of the individual, categorized in 5-year intervals
- **height(cm)**: Height of the individual in centimeters
- **weight(kg)**: Weight of the individual in kilograms
- **waist(cm)**: Waist circumference of the individual in centimeters
- **eyesight(left/right)**: Eyesight measurements for the left and right eyes
- **hearing(left/right)**: Hearing ability for the left and right ears, represented as binary
- **systolic**: Systolic blood pressure measurement
- **relaxation**: Diastolic blood pressure measurement
- **fasting blood sugar**: Fasting blood sugar level
- **Cholesterol**: Total cholesterol level
- **triglyceride**: Triglyceride level
- **HDL**: High-density lipoprotein cholesterol level
- **LDL**: Low-density lipoprotein cholesterol level
- **hemoglobin**: Hemoglobin level in the blood
- **Urine protein**: Level of protein in urine, categorized
- **serum creatinine**: Serum creatinine level
- **AST**: Level of aspartate aminotransferase enzyme
- **ALT**: Level of alanine aminotransferase enzyme
- **Gtp**: Level of gamma-glutamyl transferase enzyme
- **dental caries**: Presence (1) or absence (0) of dental cavities
- **smoking**: Target variable indicating if the individual is a smoker (1) or not (0)

All of the features in the dataset are numerical, meaning they consist of quantitative data that can be measured or counted. These numerical features provide valuable information for statistical analysis and modeling, allowing us to examine relationships, trends, and patterns in the data.

### 3.2 Summary Statistics for Numerical Features

The visual representation in Figure 1 illustrates the absence of missing data in our dataset.

	dtypes	missing count	missing percentage%	uniques	count
id	int64	0	0.000000	159256	159256
age	int64	0	0.000000	18	159256
height(cm)	int64	0	0.000000	14	159256
weight(kg)	int64	0	0.000000	28	159256
waist(cm)	float64	0	0.000000	531	159256
eyesight(left)	float64	0	0.000000	20	159256
eyesight(right)	float64	0	0.000000	17	159256
hearing(left)	int64	0	0.000000	2	159256
hearing(right)	int64	0	0.000000	2	159256
systolic	int64	0	0.000000	112	159256
relaxation	int64	0	0.000000	75	159256
fasting blood sugar	int64	0	0.000000	229	159256
Cholesterol	int64	0	0.000000	227	159256
triglyceride	int64	0	0.000000	392	159256
HDL	int64	0	0.000000	108	159256
LDL	int64	0	0.000000	222	159256
hemoglobin	float64	0	0.000000	134	159256
Urine protein	int64	0	0.000000	6	159256
serum creatinine	float64	0	0.000000	28	159256
AST	int64	0	0.000000	140	159256
ALT	int64	0	0.000000	188	159256
Gtp	int64	0	0.000000	362	159256
dental caries	int64	0	0.000000	2	159256
smoking	int64	0	0.000000	2	159256

Figure 1: Examination of Null Values, Uniqueness, and Missing Data

	count	mean	std	min	25%	50%	75%	max
id	159256.0	79627.500000	45973.391572	0.0	39813.75	79627.5	119441.25	159255.0
age	159256.0	44.306626	11.842286	20.0	40.00	40.0	55.00	85.0
height(cm)	159256.0	165.266929	8.818970	135.0	160.00	165.0	170.00	190.0
weight(kg)	159256.0	67.143662	12.586198	30.0	60.00	65.0	75.00	130.0
waist(cm)	159256.0	83.001990	8.957937	51.0	77.00	83.0	89.00	127.0
eyesight(left)	159256.0	1.005798	0.402113	0.1	0.80	1.0	1.20	9.9
eyesight(right)	159256.0	1.000989	0.392299	0.1	0.80	1.0	1.20	9.9
hearing(left)	159256.0	1.023974	0.152969	1.0	1.00	1.0	1.00	2.0
hearing(right)	159256.0	1.023421	0.151238	1.0	1.00	1.0	1.00	2.0
systolic	159256.0	122.503648	12.729315	77.0	114.00	121.0	130.00	213.0
relaxation	159256.0	76.874071	8.994642	44.0	70.00	78.0	82.00	133.0
fasting blood sugar	159256.0	98.352552	15.329740	46.0	90.00	96.0	103.00	375.0
Cholesterol	159256.0	195.796165	28.396959	77.0	175.00	196.0	217.00	393.0
triglyceride	159256.0	127.616046	66.188989	8.0	77.00	115.0	165.00	766.0
HDL	159256.0	55.852684	13.964141	9.0	45.00	54.0	64.00	136.0
LDL	159256.0	114.607682	28.158931	1.0	95.00	114.0	133.00	1860.0
hemoglobin	159256.0	14.796965	1.431213	4.9	13.80	15.0	15.80	21.0
Urine protein	159256.0	1.074233	0.347856	1.0	1.00	1.0	1.00	6.0
serum creatinine	159256.0	0.892764	0.179346	0.1	0.80	0.9	1.00	9.9
AST	159256.0	25.516853	9.464882	6.0	20.00	24.0	29.00	778.0
ALT	159256.0	26.550296	17.753070	1.0	16.00	22.0	32.00	2914.0
Gtp	159256.0	36.216004	31.204643	2.0	18.00	27.0	44.00	999.0
dental caries	159256.0	0.197996	0.398490	0.0	0.00	0.0	0.00	1.0
smoking	159256.0	0.437365	0.496063	0.0	0.00	0.0	1.00	1.0

Figure 2: Summary Statistics for Numerical Features

The dataset encompasses a diverse array of features spanning individuals' demographics, physical attributes, health metrics, and lifestyle habits. Here are notable statistics for these features:

- **id** ranges from 0 to 159,255, serving as a unique identifier for each individual.
- **Age** exhibits an average of approximately 44 years, with a standard deviation of around 11.8 years, showcasing a varied age distribution within the dataset.
- **Height (cm)** averages approximately 165.3 cm, with a standard deviation of about 8.8 cm, reflecting a typical range of heights.
- **Weight (kg)** has a mean of roughly 67.1 kg and a standard deviation of about 12.6 kg, indicating a considerable diversity in individuals' weights.
- **Waist circumference (cm)** averages approximately 83 cm, with a standard deviation of about 9 cm, suggesting variability in waist sizes among individuals.
- **Eyesight (left)** and **eyesight (right)** have mean values around 1.0, with standard deviations of about 0.4, indicating generally consistent eyesight across the dataset.
- **Hearing (left)** and **hearing (right)** represent binary values denoting hearing ability for each ear, with mean values close to 1.0, implying predominantly good hearing among individuals.
- **Systolic** and **diastolic** blood pressure measurements average around 122.5 and 76.9, respectively, underscoring the typical range of blood pressure readings in the dataset.
- Other columns encompass various blood sugar and cholesterol levels, blood substance concentrations, and dental health indicators, each exhibiting specific means and standard deviations indicative of diverse health profiles among individuals.
- The **smoking** column, serving as the target variable, indicates whether an individual is a smoker (1) or not (0). With a mean value of approximately 0.44, the data suggests that roughly 44

## 4 Data Exploration and Visualization

Extensive data exploration and visualization were conducted to understand the distribution of features, identify potential outliers, and explore relationships between variables. Key findings include:

- The dataset is imbalanced, with approximately 44% of individuals being smokers.
- Several features exhibit skewness and kurtosis, suggesting deviations from normality.
- Outlier analysis revealed the presence of extreme values in some features, which were subsequently handled using appropriate techniques.
- Bivariate analysis demonstrated significant associations between smoking status and various features, including age, BMI, blood pressure, cholesterol levels, dental caries, liver enzyme levels, and hemoglobin levels.
- Multivariate analysis using correlation matrices and hierarchical clustering provided insights into the interrelationships among features.

### 4.1 Univariate Analysis

#### 4.1.1 Smoking Status Distribution

A pie chart (Figure 3) illustrates that 56.3% of individuals are non-smokers (0), while 43.7% are smokers (1), indicating a moderate imbalance in the dataset.

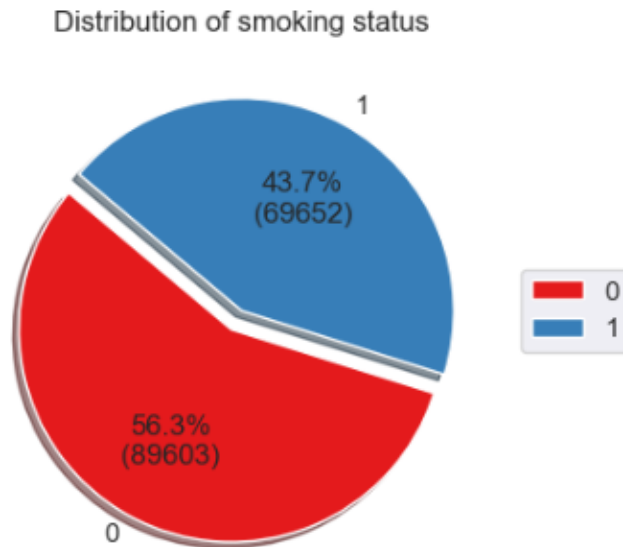


Figure 3: Distribution of Smoking Status

#### 4.1.2 Distributions of Individual Features

**Skewness:** Skewness measures the asymmetry of the distribution. A positive skewness indicates a right-skewed distribution (tail on the right), while a negative skewness indicates a left-skewed distribution (tail on the left).

**Kurtosis:** Kurtosis measures the peakedness of the distribution. Leptokurtic distributions have higher kurtosis and are more peaked, while platykurtic distributions have lower kurtosis and are flatter. Mesokurtic distributions have kurtosis similar to a normal distribution.

- **Age:** Exhibits a right-skewed distribution, suggesting a prevalence of younger individuals.

- **Height and Weight:** Approximately normal distributions.
- **Waist Circumference:** Right-skewed, indicating a higher proportion of individuals with smaller waist circumferences.
- **Eyesight (Left and Right):** Majority have good eyesight (values close to 1).
- **Hearing (Left and Right):** Majority have normal hearing (value of 1).
- **Blood Pressure (Systolic and Diastolic):** Right-skewed distributions, suggesting a higher proportion with lower blood pressure.
- **Fasting Blood Sugar, Cholesterol, Triglyceride, HDL, LDL, Hemoglobin, Urine Protein, Serum Creatinine, AST, ALT, Gtp:** Right-skewed, implying a higher proportion with lower values for these health indicators.
- **Dental Caries:** Predominantly absent among individuals (value of 0).

## 4.2 Outlier Analysis

Boxplots and Kernel Density Estimation (KDE) plots (Figure 4) reveal outliers in various features, notably in blood pressure, cholesterol levels, and liver enzyme levels. These outliers may necessitate careful consideration during data preprocessing to prevent undue influence on model performance.

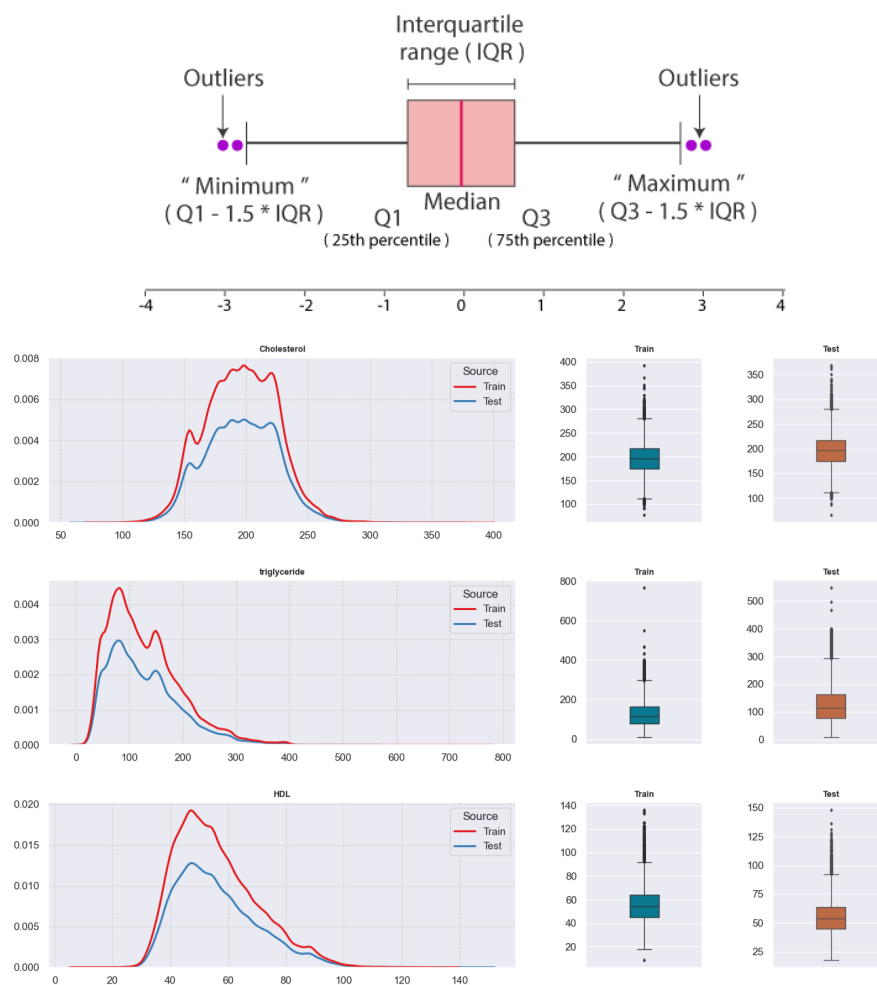


Figure 4: Outlier Analysis

## 4.3 Bivariate Analysis

### 4.3.1 Violin Plots

The distribution of features for smokers and non-smokers is depicted in violin plots (Figure 5), with noticeable differences observed in features like hemoglobin, weight, Gtp, and waist circumference between the two groups.

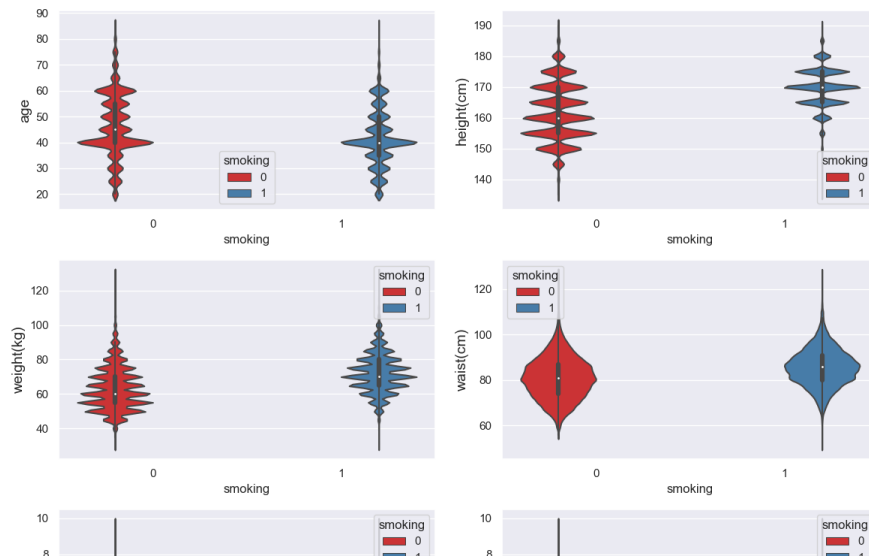


Figure 5: Violin Plots

### 4.3.2 Mean Blood Pressure

Bar plots (Figure 6) indicate that smokers tend to exhibit higher mean systolic and diastolic blood pressure compared to non-smokers.

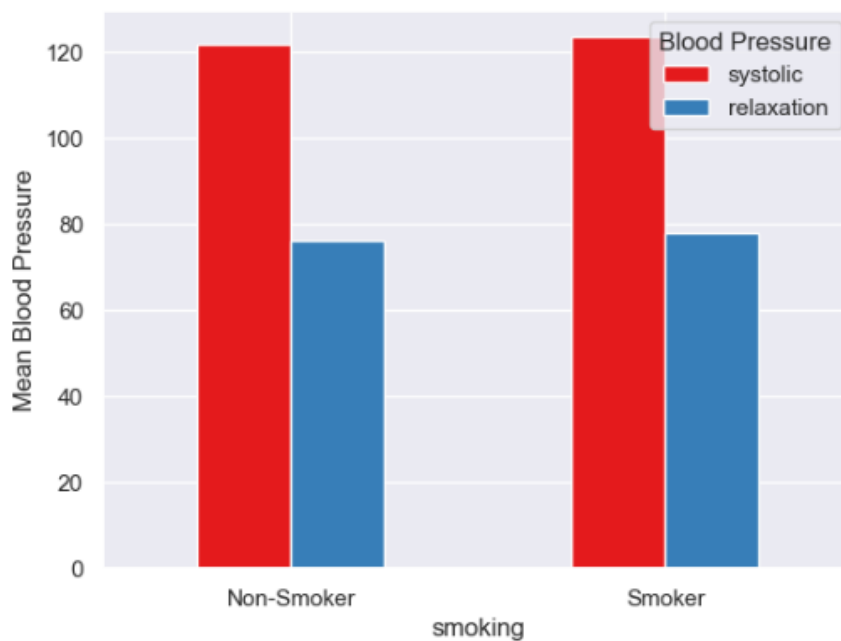


Figure 6: Mean Blood Pressure

### 4.3.3 Dental Caries

Stacked bar plots (Figure 7) demonstrate a higher prevalence of dental caries among smokers compared to non-smokers.

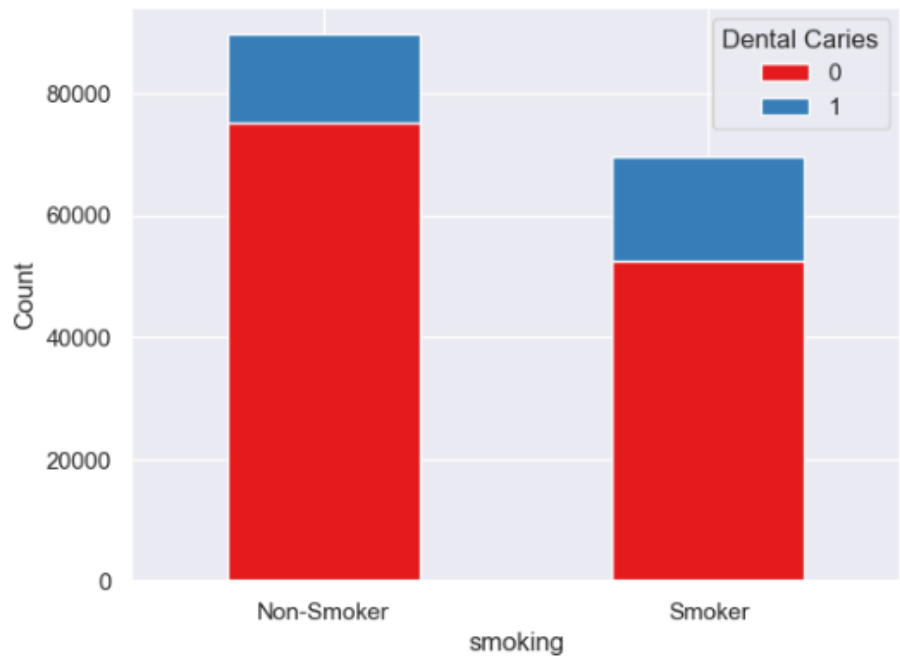


Figure 7: Dental Caries

### 4.3.4 Correlation Heatmap

Positive correlations are observed between **smoking** and features such as hemoglobin, weight, Gtp, and waist circumference. Conversely, negative correlations are noted with age and cholesterol levels (Figure 8).

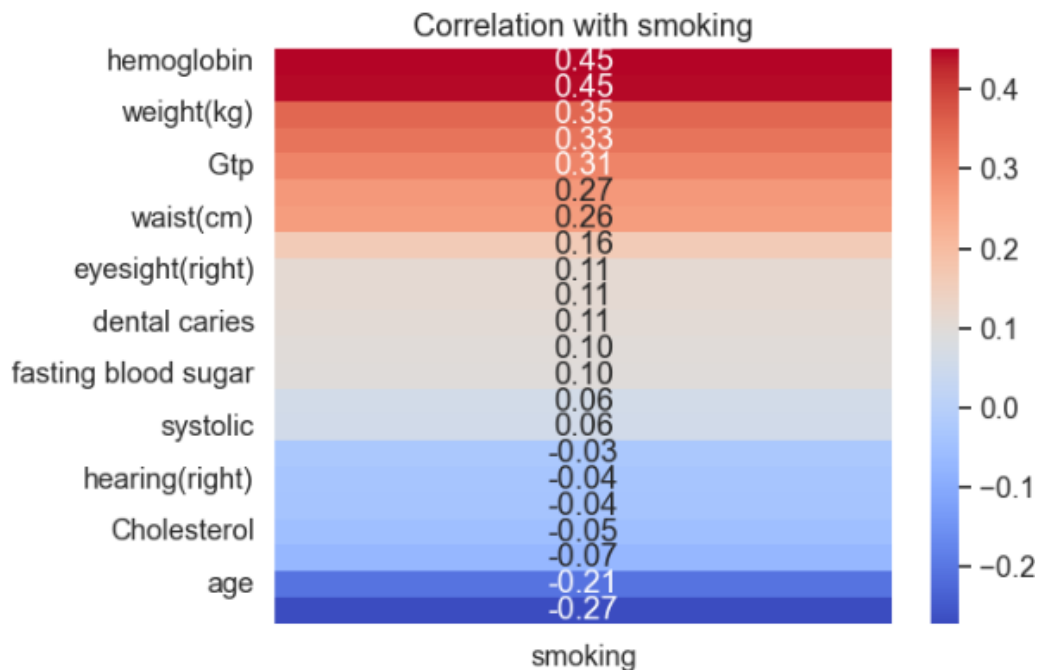


Figure 8: Correlation Heatmap

## 4.4 Multivariate Analysis

### 4.4.1 Correlation Matrix Heatmap

This heatmap (Figure 9) displays correlations between all feature pairs, revealing strong correlations that may indicate redundancy or multicollinearity, necessitating further investigation.

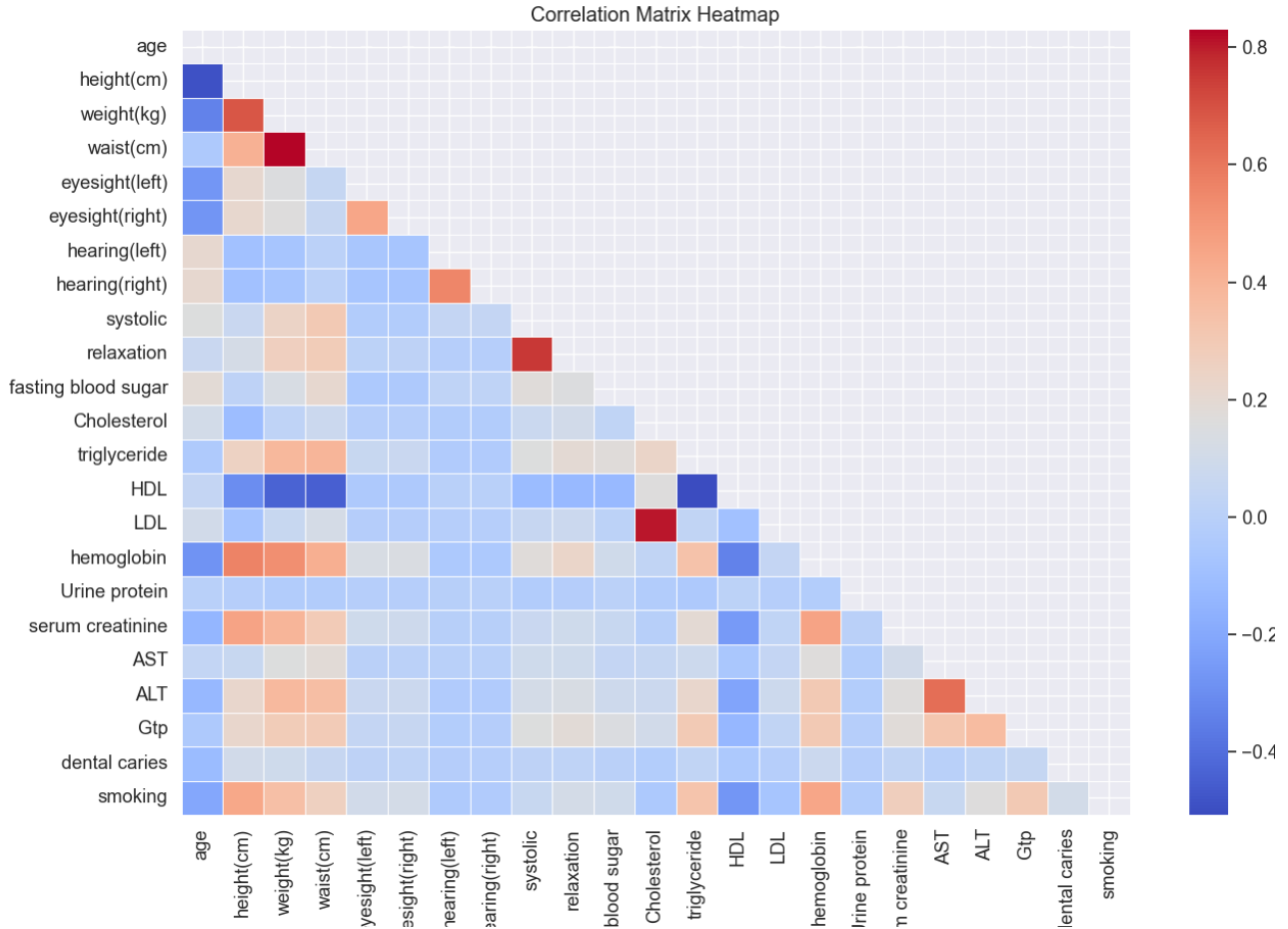


Figure 9: Correlation Matrix Heatmap

### 4.4.2 Hierarchical Clustering

The dendrogram (Figure 10) provided offers valuable insights into the relationships among features within the training dataset. By analyzing the main clusters and sub-clusters, significant insights can be gleaned to inform actionable strategies.

#### Physiological Health Features

*Age, Height (cm), Weight (kg), Waist Circumference (cm), Eyesight (left and right), Hearing (left and right), Hemoglobin, Urine Protein, Dental Caries*

#### Cardiovascular Health Features

*Systolic Blood Pressure, Resting Heart Rate, LDL Cholesterol, HDL Cholesterol, Total Cholesterol, Triglycerides*

#### Metabolic Health Features

*Fasting Blood Sugar, Serum Creatinine, AST (Aspartate Aminotransferase), ALT (Alanine Aminotransferase), GTP (Gamma-glutamyltransferase)*

In conclusion, a thorough interpretation of the dendrogram offers a holistic view of the training data. This analysis facilitates the identification of pertinent feature groups, establishes connections between



various health aspects, and supports the development of personalized intervention strategies.

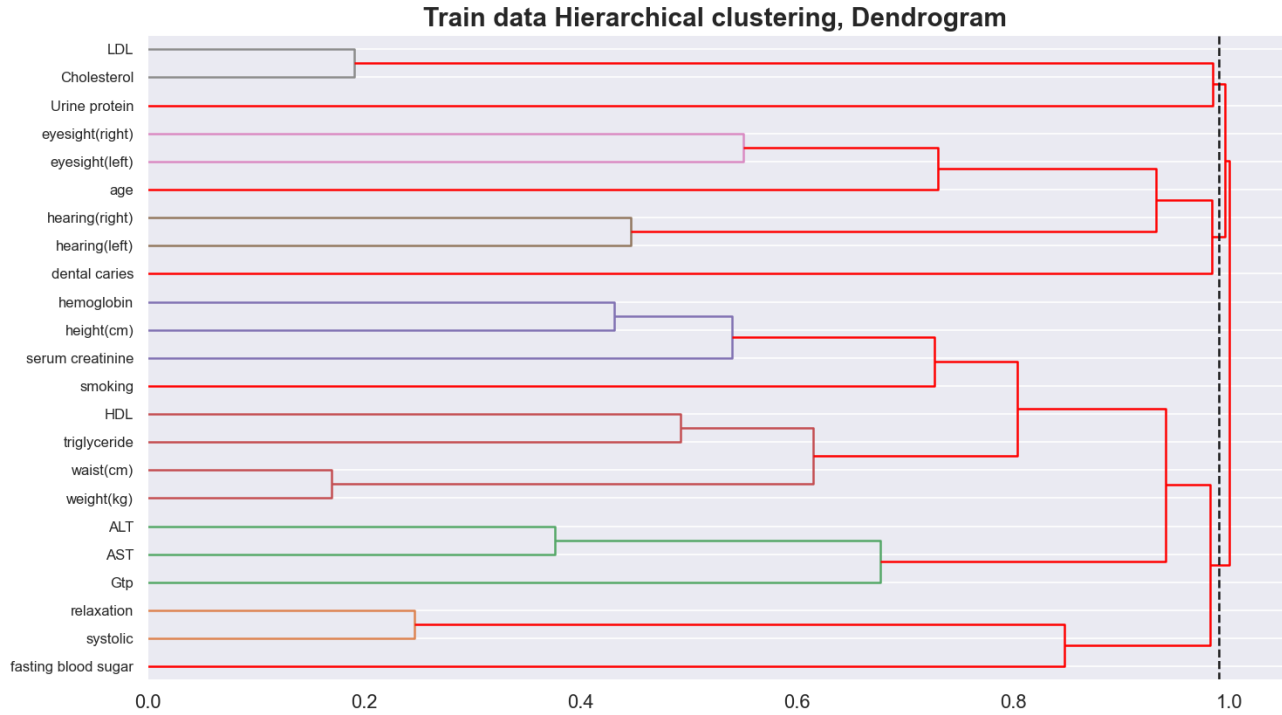


Figure 10: Hierarchical Clustering

## 5 Hypothesis Testing

Statistical hypothesis tests were employed to formally assess the relationships between smoking status and specific features. The tests confirmed significant associations between smoking and age, BMI, blood pressure, cholesterol levels, dental caries, liver enzyme levels, and hemoglobin levels. These findings align with existing medical knowledge about the health effects of smoking.

### 5.1 Statistical Test Selection

When analyzing the data, various statistical tests were chosen to evaluate different hypotheses. Here is a summary of the tests used along with their respective hypotheses:

Statistical Test	Use Case	Feature
Independent Samples t-test	Comparing means of a continuous variable between two independent groups	Age, BMI, Blood Pressure, Hemoglobin Level
ANOVA	Comparing means of a continuous variable across more than two groups	Cholesterol Levels, Liver Enzyme Levels
Chi-square Test	Analyzing the association between two categorical variables	Dental Caries

Table 1: Statistical Tests, Use Cases, and Features

These tests were chosen based on the nature of the data and the hypotheses being tested.

## 5.2 Hypotheses and Results

Features	Hypothesis	Results	Justification
Age	<b>H0:</b> There is no significant difference in life expectancy between smokers and non-smokers. <b>H1:</b> Non-smokers have a higher life expectancy compared to smokers.	T-statistic= $-84.023$ p-value= 0.0	we accept the alternative hypothesis (H1), which states that non-smokers have a higher life expectancy compared to smokers.
BMI	<b>H0:</b> There is no association between BMI and smoking status. <b>H1:</b> There is a significant association between BMI and smoking status.	T-statistic= 57.133 p-value= 0.0	we reject the null hypothesis (H0) because the extremely low p-value indicates strong evidence against it.so we accept the alternative hypothesis (H1)
Blood Pressure	<b>H0:</b> There is no significant difference in blood pressure measurements between smokers and non-smokers. <b>H1:</b> Smokers have significantly different blood pressure measurements compared to non-smokers.	T-statistic = 23.442 p-value = $2.54 \times 10^{-121}$	we reject the null hypothesis (H0) because the extremely low p-value indicates strong evidence against it.So,we accept the alternative hypothesis (H1).
Cholesterol levels	<b>H0:</b> There is no difference in cholesterol levels between smokers and non-smokers. <b>H1:</b> Cholesterol levels differ significantly between smokers and non-smokers.	F-statistic= 430.061 p-value= $2.108 \times 10^{-95}$	we accept the alternative hypothesis (H1), which states that cholesterol levels differ significantly between smokers and non-smokers.
Dental Caries	<b>H0:</b> There is no association between the presence of dental caries and smoking status. <b>H1:</b> Individuals who smoke are more likely to have dental caries compared to non-smokers.	$\chi^2$ statistic = 1810.406 p-value = 0.0	we accept the alternative hypothesis (H1), which states that individuals who smoke are more likely to have dental caries compared to non-smokers.
Liver Enzyme Levels	<b>H0:</b> There is no difference in liver enzyme levels between smokers and non-smokers. <b>H1:</b> Liver enzyme levels vary significantly between smokers and non-smokers.	<b>AST</b> F-statistic=563.7785 p-value= $2.0874 \times 10^{-124}$ <b>ALT</b> F-statistic=4347.5856 p-value=0.0 <b>GTP</b> F-statistic=16400.4220 p-value=0.0	We reject the null hypothesis (H0) because the extremely low p-values indicate strong evidence against it.So,we accept the alternative hypothesis (H1).
Hemoglobin Level	<b>H0:</b> There is no difference in hemoglobin levels between smokers and non-smokers. <b>H1:</b> Hemoglobin levels differ significantly between smokers and non-smokers.	T-statistic=201.4714 p-value=0.0	we reject the null hypothesis (H0) because the extremely low p-value indicates strong evidence against it.So,we accept the alternative hypothesis (H1).

Table 2: Hypotheses and Results

## 6 Model Selection

### 6.1 Models

Several machine learning models were trained and compared, including:

- Random Forest Classifier
- LightGBM Classifier
- CatBoost Classifier
- XGBoost Classifier

The choice of tree-based models (Random Forest, LightGBM, CatBoost, XGBoost) is well-suited for this problem because:

- They can handle mixed data types (numerical and categorical) effectively.
- They are robust to outliers and non-linear relationships.
- They often provide good performance without requiring extensive feature scaling or encoding.

### 6.2 Evaluation Metrics

Evaluation metrics are essential for assessing the performance of machine learning models, especially in classification tasks. In our project, we employ the following commonly used metrics to evaluate model performance:

- **ROC Curve and AUC:** The Receiver Operating Characteristic (ROC) curve served as a central tool for evaluation (Figure 11). It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. TPR is also known as sensitivity or recall, while FPR is the complement of specificity. Mathematically, TPR is defined as  $TPR = \frac{TP}{TP+FN}$  and FPR is defined as  $FPR = \frac{FP}{FP+TN}$ . The area under the curve (AUC) provided a quantitative measure of this ability, with a higher AUC indicating better overall performance.

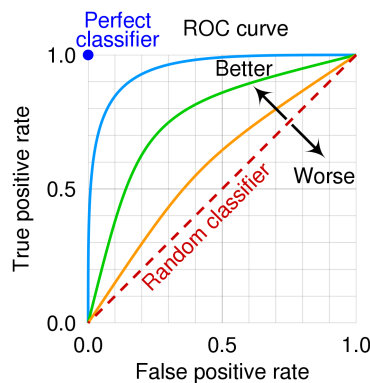


Figure 11: Illustration of ROC Curves with Varying Performance Levels.

- **Confusion Matrix:** This table summarized the model's predictions, detailing the counts of true positives, true negatives, false positives, and false negatives.
- **Classification Report:** Building upon the confusion matrix, the classification report provided a breakdown of key metrics like precision, recall, F1-score, and accuracy for each class. While accuracy offers a basic overview of performance, metrics like precision and recall provided more nuanced insights, especially when dealing with imbalanced class distributions.

## Key Finding: Predicting Probabilities vs. Binary Prediction

Through analysis of the ROC curve and associated metrics, we discovered a crucial insight: predicting probabilities yielded superior results compared to binary classification. The ROC curve's shape and the AUC value demonstrated that the model was more effective at assessing the likelihood of class membership than simply assigning a positive or negative label. This finding has significant implications for real-world applications, where knowing the probability of an event can be more valuable than a simple yes/no prediction.

### 6.3 Before Data Preprocessing

Before proceeding with data preprocessing, we utilized base models to gain insights into the data quality and assess feature importance. This initial exploration aimed to inform our feature selection process. The scores obtained from these base models indicated promising performance, as depicted in Figure 12. Additionally, during this analysis, we identified two common features, namely 'Urine protein' and 'hearing(left)', which had negligible importance across all models (Figure 13).

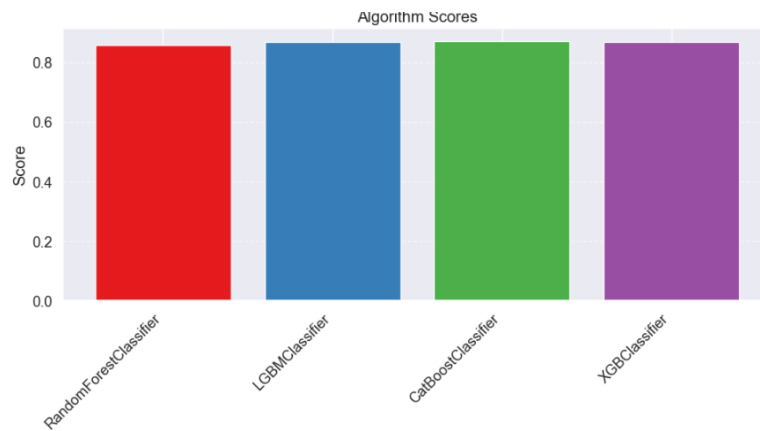


Figure 12: Base Model - ROC Curve Scores

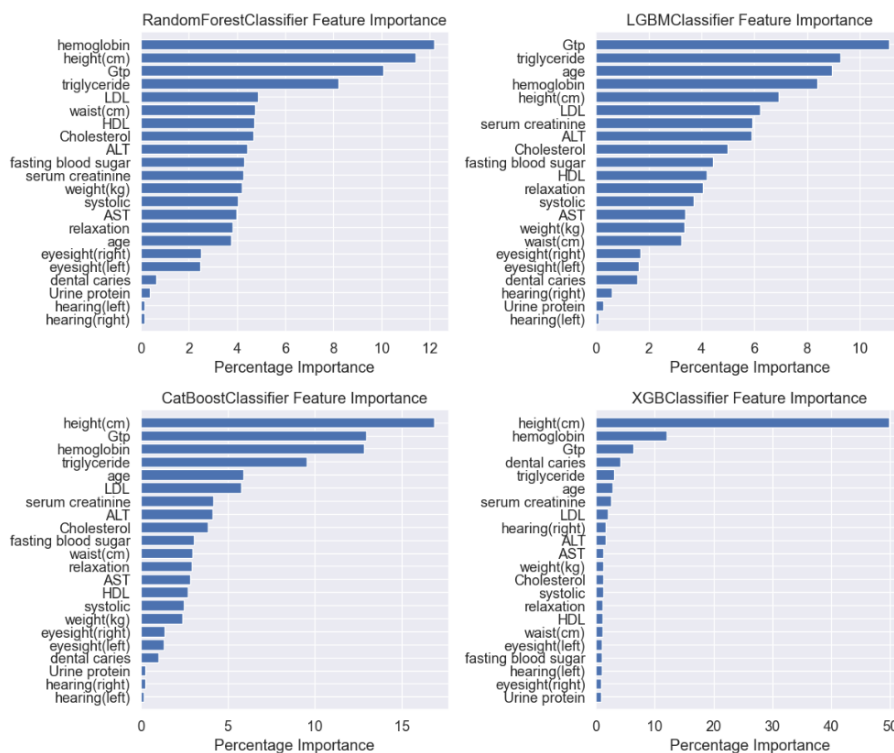


Figure 13: Feature Importance

## 7 Data Preprocessing and Feature Engineering

Data preprocessing steps were implemented to prepare the data for model training. It's noteworthy that *encoding* was not necessary as the dataset solely comprised numerical features, and tree-based models were employed.

### 7.1 Feature Engineering

Feature engineering played a crucial role in enhancing the dataset by creating new features to capture additional information and potentially improve model performance. The following engineered features were added:

- **BMI (Body Mass Index):** BMI is a widely used health indicator that provides a more comprehensive measure of healthy weight than weight alone. It is calculated as follows:

$$BMI = \frac{\text{weight (kg)}}{[\text{height (m)}]^2}$$

- **Clipped Serum Creatinine:** Serum creatinine values exceeding 3 were clipped to this upper limit to mitigate the influence of potential outliers or extreme medical conditions.
- **Waist-to-Height Ratio:** This ratio serves as a strong indicator of abdominal obesity, a significant health risk factor. It is calculated as:

$$\text{Waist-to-Height Ratio} = \frac{\text{waist circumference (cm)}}{\text{height (cm)}}$$

- **Average Eyesight:** The average of left and right eye eyesight measurements provides a single metric representing overall visual acuity.
- **Average Hearing:** Similar to eyesight, averaging left and right ear hearing abilities provides a single metric for overall auditory function.
- **Blood Pressure Category:** Systolic blood pressure was categorized into clinically relevant ranges (normal, prehypertension, hypertension) to provide more informative input for the model.
- **Cholesterol Ratio:** The ratio of HDL cholesterol to LDL cholesterol serves as a better indicator of cardiovascular health risk compared to individual cholesterol levels. It is calculated as:

$$\text{Cholesterol Ratio} = \frac{\text{HDL cholesterol}}{\text{LDL cholesterol}}$$

These engineered features aimed to capture relevant health information and improve the model's ability to identify relationships between these factors and smoking habits.

### 7.2 Handling outliers

Addressing outliers within the dataset was crucial for ensuring data quality and model reliability. Initially, we explored the Interquartile Range (**IQR**) method for outlier detection; however, this approach did not effectively identify outliers relevant to our specific dataset and domain. Therefore, we implemented a targeted outlier handling strategy based on domain knowledge and research-informed thresholds for each numerical feature. This involved establishing upper and lower bounds for values considered plausible within the context of human physiology and medical understanding. For instance, **systolic blood pressure** values were constrained to a range of **80 to 220 mmHg**, while **fasting blood sugar levels** were capped at **350 mg/dL**, both based on established medical guidelines and typical ranges. This meticulous outlier management process resulted in the removal of **44 rows** (out of **159,256**) identified as containing extreme values, ensuring the remaining data reflects realistic and reliable measurements.

### 7.3 Dropping unnecessary columns

Following a thorough review of the feature set and an assessment of feature importance from the initial base models, we proceeded to remove columns deemed **irrelevant** or **redundant** for predicting smoking habits. This step aimed to streamline the dataset, enhance model efficiency, and potentially improve predictive performance by *reducing noise* and *dimensionality*. The following columns were dropped:

- **'id'**: As a *unique identifier* with no inherent predictive value, the 'id' column was removed.
- **'height (cm)'**: Information regarding height is already captured and utilized within the engineered 'BMI' feature, rendering the original height column redundant.
- **'eyesight (left)', 'eyesight (right)', 'hearing (left)', 'hearing (right)'**: These individual eye and ear measurements were replaced by the engineered features '*average\_eyesight*' and '*average\_hearing*', respectively, which provide a more concise representation of overall sensory abilities while *reducing* data dimensionality.

This feature selection process ensured that the remaining columns retained valuable information directly relevant to the prediction task, leading to a more focused and efficient model development process.

### 7.4 Feature scaling

Initially, we chose not to employ feature scaling techniques due to our reliance on tree-based algorithms for modeling. These algorithms inherently manage variations in feature magnitudes without the need for feature scaling. However, as our work progressed, we found it beneficial to implement feature scaling techniques, particularly in the context of PCA application.

### 7.5 PCA: Principal Component Analysis

Given the high correlations among some features (**Figure 14**), PCA (Principal Component Analysis) seemed like a potential solution.

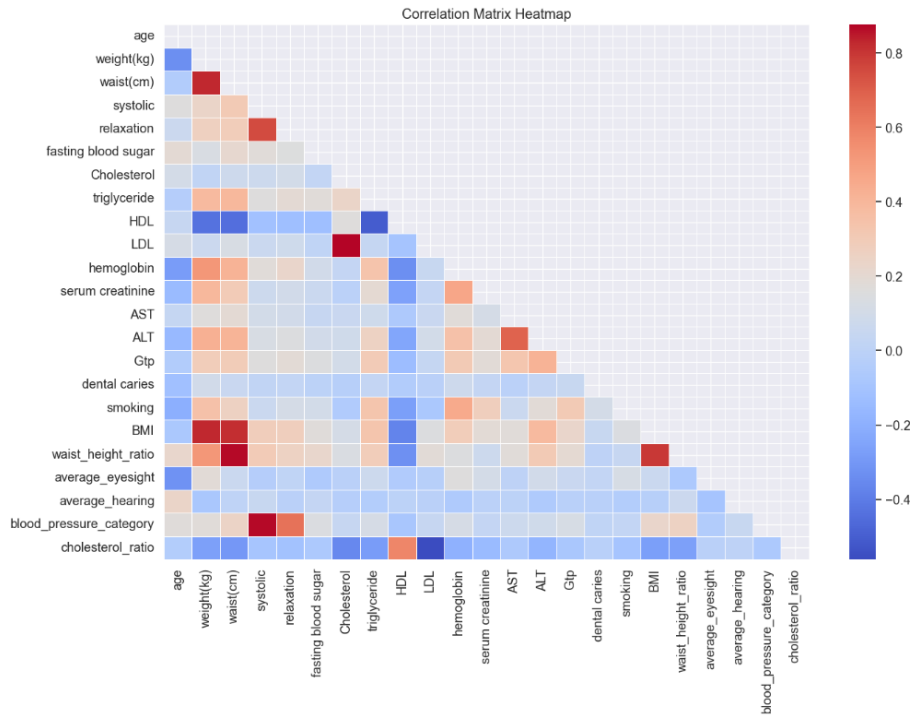


Figure 14: Correlation matrix - After preprocessing

PCA reduces dimensionality by focusing on the most significant data variances.

Steps:

- 1. Standardize the data.
- 2. Apply PCA on the training data.
- 3. Transform the test data using the same PCA transformation.

Figure 19 illustrates that **17 components** capture around **98%** of variance.

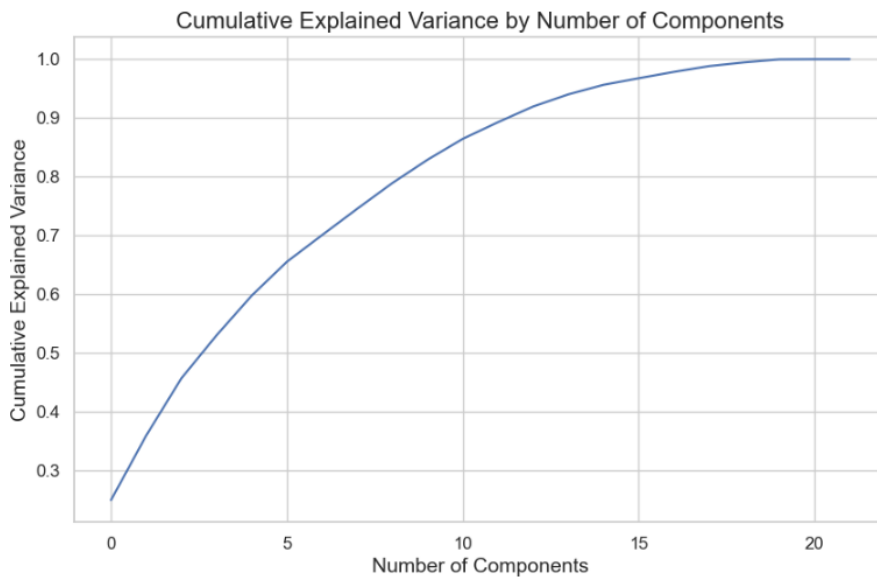


Figure 15: Cumulative explained variance ratio by number of components in PCA.

***Unexpected Outcome and Adaptation:** Despite our initial optimism about PCA's potential, its failure to improve results caught us off guard. However, recognizing this outcome allowed us to adapt our*

approach, focusing on alternative strategies to enhance model performance. Hence, we didn't proceed with PCA.

## 8 Model Training and Evaluation

In this section, we discuss the training and evaluation of various machine learning models for predicting smoking habits. We start by splitting the train data into training and validation sets. The target variable for prediction is "smoking".

### 8.1 Data Splitting

We split the data into training and validation sets with a 90-10 split ratio (approximately 143290 samples for training and 15921 for validation). This choice was made to ensure a sufficiently large training set for complex models like XGBoost and CatBoost while maintaining a reasonable validation set size for performance evaluation. The splitting process was performed using a fixed random seed (42) to guarantee reproducibility of the results.

### 8.2 Model Parameters

Although our main focus wasn't on fine-tuning model performance, we experimented with a selection of basic parameters chosen randomly. We iteratively trained the models, comparing their performance each time.

- **Random Forest Classifier:** We initialized a Random Forest Classifier model with default parameters as a baseline for comparison.
- **LightGBM (LGBM) Classifier:** Based on initial experimentation, we set the parameters for the LGBM Classifier as follows: `n_estimators = 1700`, `max_depth = 7`, `learning_rate = 0.09`, `colsample_bytree = 0.8`. These values were chosen to balance model complexity and learning speed while preventing overfitting.
- **XGBoost (XGB) Classifier:** Similar to LightGBM, we tuned the XGBoost Classifier parameters: `n_estimators = 1700`, `max_depth = 7`, `learning_rate = 0.09`, `colsample_bytree = 0.8`, `reg_lambda = 0.5`. The additional `reg_lambda` parameter was used to control regularization and further prevent overfitting.
- **CatBoost Classifier:** For the CatBoost Classifier, we used the following parameters: `iterations = 1700`, `depth = 7`, `learning_rate = 0.09`. CatBoost automatically handles categorical features and offers robust performance with relatively fewer parameter adjustments.

### 8.3 Evaluation

We trained each model using the training data and evaluated its performance on the validation set using various metrics: precision, recall, F1-score, and AUC-ROC. The results are presented in the table below.

Model	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.77	0.78	0.77	0.858
LightGBM	0.79	0.79	0.79	0.873
CatBoost	0.79	0.79	0.79	0.874
XGBoost	0.78	0.79	0.78	0.869

Table 3: Model Performance Comparison



## LightGBM :

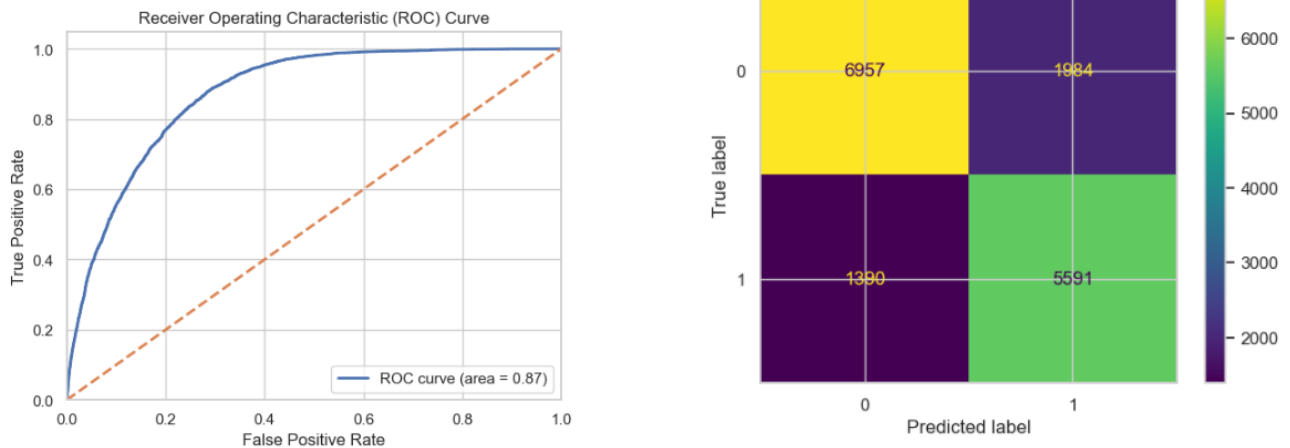


Figure 16: LightGBM: ROC Curve (left) and Confusion Matrix (right)

## CatBoost :

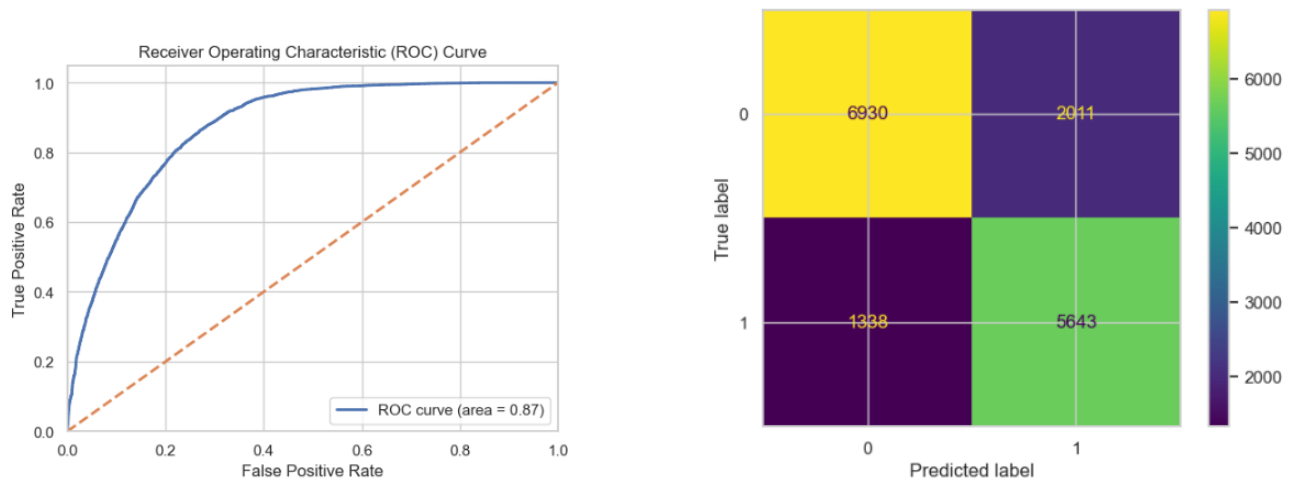


Figure 17: CatBoost: ROC Curve (left) and Confusion Matrix (right)

### 8.3.1 Analysis

LightGBM and CatBoost **clearly outperformed** Random Forest and XGBoost in predicting smoking status, achieving impressive AUC-ROC scores above 0.87, indicating their strong ability to differentiate between smokers and non-smokers. While both models excel, a closer look reveals subtle differences: **LightGBM** demonstrates a more balanced performance across both classes, as evidenced by its confusion matrix, while **CatBoost**, despite achieving the highest AUC-ROC, exhibits a slight tendency towards false negatives, potentially overlooking some smokers. **XGBoost** performs comparably in terms of precision and recall but falls slightly behind in overall discrimination. **Random Forest**, with the lowest performance across all metrics, underscores the effectiveness of the other three models in this task.

## 8.4 Model Comparison (Before and After Preprocessing)

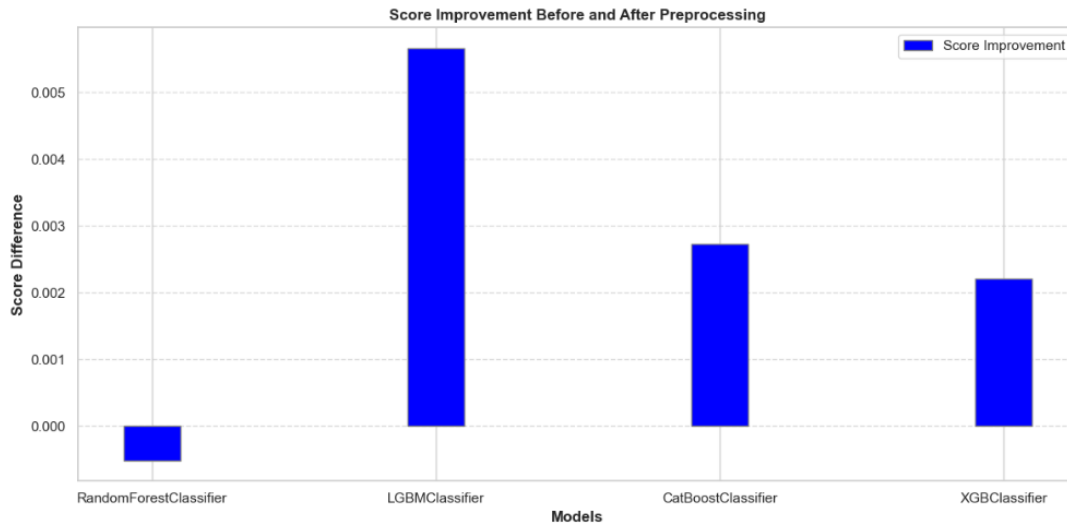


Figure 18: Score Improvement.

The plot above demonstrates the improvement in AUC-ROC scores for each model after data preprocessing and feature engineering. All models benefitted from these steps, with **LightGBM** showing the most significant improvement.

## 9 Ensemble method and final results

To leverage the strengths of both **LightGBM** and **CatBoost**, we created an ensemble model by averaging their predictions. This approach aims to achieve better generalization and robustness compared to individual models.

### 9.1 Cross-Validation

To further evaluate the ensemble model and individual models, we employed k-fold cross-validation with both standard and stratified approaches. This provided more reliable performance estimates and helped assess potential overfitting.

#### Standard Cross-Validation:

CatBoostClassifier: [0.8661, 0.8655, 0.8668, 0.8689, 0.8684]

LGBMClassifier: [0.8642, 0.8648, 0.8657, 0.8675, 0.8667]

#### Stratified K-Fold Cross-Validation:

CatBoostClassifier: [0.8689, 0.8694, 0.8648, 0.8648, 0.8687]

LGBMClassifier: [0.8676, 0.8681, 0.8642, 0.8630, 0.8673]

The cross-validation results confirm the consistent performance of both CatBoost and LightGBM across different data subsets. Stratified k-fold ensured that class distributions were preserved in each fold, further validating the models' ability to generalize to the overall population.

### 9.2 Evaluation on Unseen Data Using the Ensemble Approach

The final ensemble model, consisting of averaged predictions from CatBoost and LightGBM, was deployed to predict smoking status on the unseen test data for the Kaggle competition.

Competition link: <https://www.kaggle.com/competitions/playground-series-s3e24>




Submission and Description		Private Score 	Public Score 	Selected
	submission.csv Complete (after deadline) · 18s ago	0.87087	0.87236	<input type="checkbox"/>

Figure 19: Final results.

The final Kaggle score of **0.87** demonstrates the effectiveness of our ensemble model in predicting smoking status based on health indicators.

### 9.3 Further Considerations

While our models achieved promising results, there's room for further improvement and exploration:

- **Hyperparameter tuning:** Fine-tuning hyperparameters for each individual model and the ensemble could potentially lead to further performance gains.
- **Additional data sources:** Incorporating more diverse data, such as socioeconomic factors or behavioral data, could provide additional insights and improve model accuracy.
- **Model interpretability:** Employing techniques like feature importance analysis or LIME would allow us to understand the factors driving predictions and potentially uncover new relationships between health indicators and smoking behavior.

## 10 Conclusion and Future Work

This project highlighted the synergy between data analysis and machine learning in addressing the public health challenge of smoking. Through meticulous exploration and visualization of health data, we uncovered key insights into the factors influencing smoking behavior, which guided our model development process. Employing powerful algorithms like LightGBM and CatBoost, we built accurate predictive models, particularly the ensemble model, achieving a competitive Kaggle score. This success underscores the importance of a data-driven approach in understanding and tackling public health issues. Future efforts will focus on further data exploration, model refinement, and ultimately, translating these findings into impactful interventions for smoking cessation and improved public health outcomes.