San Jose State University

Department of Applied Data Science

Dr. Guannan Liu

The Impact of Smoking and BMI on Health: A Comprehensive Analysis

Team 5: Cheng-Huan Yu, Chun-Chieh Kuo, Khac Minh Dai Vo, Lam Tran

December 4th, 2024

Abstract

This study investigates the complex relationships between smoking, BMI, and critical health metrics such as liver function, cardiovascular health, and metabolic indicators. Using a dataset of over 55,000 individuals, key correlations were identified between BMI and health outcomes, emphasizing the compounded risks associated with smoking and obesity. Advanced data visualization techniques, including radar charts and scatter plots, reveal gender- and age-specific trends, highlighting the heightened liver stress and metabolic challenges faced by smokers, particularly in obese populations. Predictive modeling, such as Random Forest classifiers, achieved high accuracy (88%) in identifying individuals at risk based on key health metrics, validating the importance of BMI and related factors as reliable predictors. These findings underscore the need for targeted health interventions and lifestyle changes to mitigate the long-term impacts of smoking and obesity on public health.

Table of Contents

1. Introduction

1.1- Problem Statement

This study digs into the relationship between smoking behavior and body composition metrics such as body mass index (BMI) to evaluate its impact on health indicators. The goal is to determine whether these metrics can predict smoking-related health risks and suggest changes in individual body states through the data.

1.2 - Motivation

Addressing the impact of smoking on health requires a comprehensive understanding of its effects on key health indicators such as cardiovascular health, liver function, and metabolic metrics. The combination of smoking and other factors such as body mass index (BMI) may increase health risks, so it is important to systematically study their relationships. Using clear and powerful visualization tools such as bar charts, box plots, and histogram plots to catch the trends can highlight the expansion of these influences and discover potential insights.

The objective of this study is to provide meaningful insights by exploring the correlation and patterns between smoking behavior and health indicators. Identifying these patterns will help design more targeted health programs and preventions from diseases. Moreover, this study helps bridge the knowledge gap on how lifestyle choices impact health risks and how these effects vary across population groups. By transforming data into persuasive and actionable recommendations, the analysis of this study supports the development of strategies aimed at reducing the adverse effects of smoking and abnormal physical states on the health of individuals.

2 - Data Exploration

2.1 - Data Description

The dataset, titled *Body Signal of Smoking*, was obtained from Kaggle [1] and originally sourced from the official e-Government website of the Republic of Korea [2]. It contains 55,692 observations and 26 features, of which 23 are numerical and the remaining are categorical. These features are categorized into two main groups: personal information and blood test results. Key categorical variables include gender (Male-M, Female-F) and oral health indicators (‘oral’ and

'tartar,' where Y denotes yes and N denotes no), alongside numerical variables such as age (in years), height (in centimeters), weight (in kilograms), and smoking status (0 for non-smoking, 1 for smoking). Each participant is uniquely identified by an ID variable.

The dataset also provides a detailed set of health metrics, including systolic and diastolic blood pressure (mmHg), fasting blood sugar (mg/dL), cholesterol-related variables (total cholesterol, triglycerides, HDL, and LDL, all in mg/dL), and hemoglobin concentration (g/dL). Additional biochemical markers include serum creatinine (mg/dL), urine protein, and enzyme levels such as AST, ALT, and Gtp (IU/L). These features are stored as either integers (ID, age, height(cm), weight(kg), smoking, dental caries) or floating-point numbers (e.g., waist(cm), systolic, relaxation, fasting blood sugar, and various enzyme levels). While the dataset does not include BMI as a pre-computed feature, it was calculated separately using the standard formula:

$$BM = \frac{weight\ (kg)}{height\ (m)^2}$$

This calculated feature was used in the analysis to explore trends and associations. The clear distinction in data types ensures numerical precision and facilitates statistical analyses.

These features collectively offer a solid foundation for analyzing BMI trends and their associations with smoking behavior. The dataset is provided in CSV format, with no missing values reported, ensuring data completeness and reliability for analysis. Participants were selected randomly from a pool of 1 million individuals, including employed subscribers, dependents aged 20 or older, regional subscribers who are heads of households, and regional subscribers aged 20 or older with a history of general health checkups.[2]

2.2 - Exploratory Analysis

2.2.1 - Overview of Data Distribution

	Cholesterol	triglyceride	HDL	LDL	hemoglobin	Urine protein	serum creatinine	AST	ALT	Gtp	dental caries
count	55692	55692	55692	55692	55692	55692	55692	55692	55692	55692	55692
mean	196.9014221	126.665697	57.29034691	114.9645012	14.62259211	1.087211808	0.8857376284	26.18293471	27.03603749	39.95220139	0.2133340516
std	36.29794042	71.63981726	14.73896278	40.92647644	1.564498437	0.4048824024	0.2215241416	19.35545978	30.94785288	50.29053882	0.4096652871
min	55	8	4	1	4.9	1	0.1	6	1	1	0
25%	172	74	47	92	13.6	1	0.8	19	15	17	0
50%	195	108	55	113	14.8	1	0.9	23	21	25	0
75%	220	160	66	136	15.8	1	1	28	31	43	0
max	445	999	618	1860	21.1	6	11.6	1311	2914	999	1

	Cholesterol	triglyceride	HDL	LDL	hemoglobin	Urine protein	serum creatinine	AST	ALT	Gtp	dental caries	smoking
count	55692	55692	55692	55692	55692	55692	55692	55692	55692	55692	55692	55692
mean	196.9014221	126.665697	57.29034691	114.9645012	14.62259211	1.087211808	0.8857376284	26.18293471	27.03603749	39.95220139	0.2133340516	0.3672879408
std	36.29794042	71.63981726	14.73896278	40.92647644	1.564498437	0.4048824024	0.2215241416	19.35545978	30.94785288	50.29053882	0.4096652871	0.4820702046
min	55	8	4	1	4.9	1	0.1	6	1	1	0	0
25%	172	74	47	92	13.6	1	0.8	19	15	17	0	0
50%	195	108	55	113	14.8	1	0.9	23	21	25	0	0
75%	220	160	66	136	15.8	1	1	28	31	43	0	1
max	445	999	618	1860	21.1	6	11.6	1311	2914	999	1	1

Table 1: Statistical Overview of Health Variables

2.2.3 - Comparison of Smoking and Non-Smoking Groups

This bar chart provides a comparison of the average Body Mass Index (BMI) between smokers and non-smokers. The data reveals that smokers have a slightly higher average BMI of 24.65 compared to non-smokers, whose average BMI is 23.88. This difference suggests that smoking may be associated with variations in BMI. The chart emphasizes how lifestyle choices, such as smoking, could potentially influence body composition metrics like BMI.

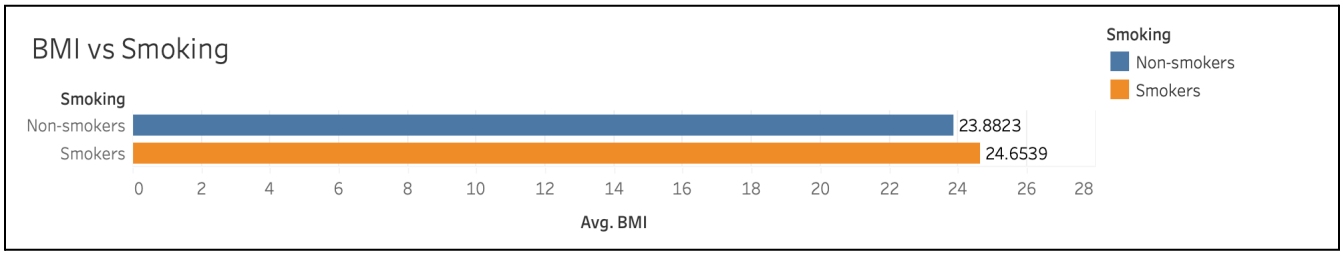


Figure 1: Bar Chart of BMI vs Smoking

2.2.4 - Gender-Based Analysis

This bart chart shows the connection between BMI, smoking, and gender. It shows that both male and female smokers tend to have a slightly higher average BMI compared to non-smokers. Female smokers have a higher BMI than female non-smokers, and the same pattern is seen in males. Additionally, male smokers generally have a slightly higher BMI than female smokers, while BMI for non-smokers is more similar between genders. This highlights that smoking is linked to higher BMI in both men and women, suggesting it may impact weight differently based on gender.

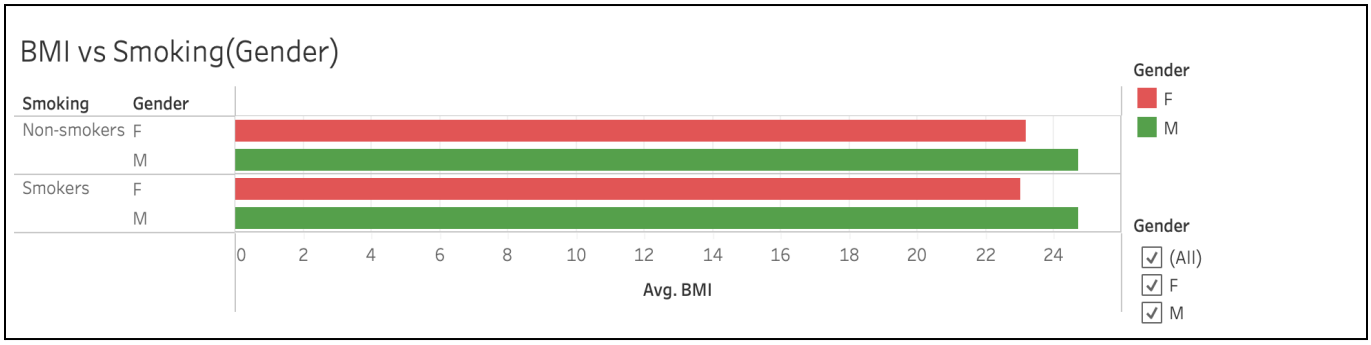


Figure 2: Bar Chart of BMI vs Smoking based on Gender

2.2.5 - Correlation Analysis

Correlation of BMI vs Gender and Smoking Status

	Gender / Smoking			
	F		M	
	N	Y	N	Y
Systolic	0.3087	0.3245	0.2563	0.2487
Triglyceride	0.2945	0.3018	0.2840	0.2782
Relaxation	0.2674	0.2590	0.2520	0.2380
ALT	0.1205	0.1853	0.2335	0.2628
Fasting Blood Sugar	0.2266	0.1795	0.1460	0.1157
Hemoglobin	0.1073	0.1315	0.1485	0.2017
Cholesterol	0.0895	0.1236	0.1235	0.1346
LDL	0.0951	0.1222	0.0905	0.1113
GTP	0.1712	0.0868	0.1865	0.1056
Urine protein	0.0212	0.0611	0.0398	0.0461
AST	0.0714	0.0370	0.1622	0.1098
AGE	0.1673	0.0072	-0.0494	-0.0856
Serum Creatinine	0.0328	-0.0082	0.0349	0.0671
HDL	-0.2650	-0.2660	-0.2598	-0.2992

Table 2 : Correlation of BMI with health metrics by gender and smoking status

The correlation analysis reveals significant relationships between BMI and various health metrics, with distinct differences across gender and smoking status. Systolic blood pressure and triglycerides demonstrate the strongest positive correlations with BMI, particularly among female smokers, emphasizing the link between BMI and cardiovascular health. Conversely, HDL (good cholesterol) shows a consistent negative correlation with BMI, indicating that higher BMI levels are associated with lower HDL levels, with the strongest negative correlation observed in male smokers. Gender-based differences are evident, as females show stronger correlations between BMI and systolic blood pressure or triglycerides, while males display higher correlations with liver-related metrics such as ALT and hemoglobin. Smoking further strengthens these relationships, with smokers showing stronger correlations between BMI and systolic blood pressure or ALT compared to non-smokers, suggesting that smoking worsens BMI-related impacts on cardiovascular and liver health. Overall, BMI is closely associated with both cardiovascular and liver health metrics, with smoking enhancing these effects, particularly in males. To further investigate the relationship between BMI, gender, and smoking status, the analysis focuses on the top 11 features with the strongest correlations with BMI: systolic blood pressure, triglycerides, relaxation, ALT, fasting blood sugar, hemoglobin, cholesterol, LDL, Gtp, urine protein, and AST. These features provide a comprehensive view of BMI trends across gender and smoking behavior.

3 - Data Visualization

This section provides a detailed analysis of the relationships between smoking habits, BMI, and key health metrics, offering insights into how smoking influences both external

anthropometric measurements and internal health markers. The visualizations will emphasize the critical patterns and provide the trends observed in the data.

3.1 - Health Metrics and Data Overview

3.1.1 - BMI and Smoking Trends

Box Plot: Health Metrics by BMI and Smoking

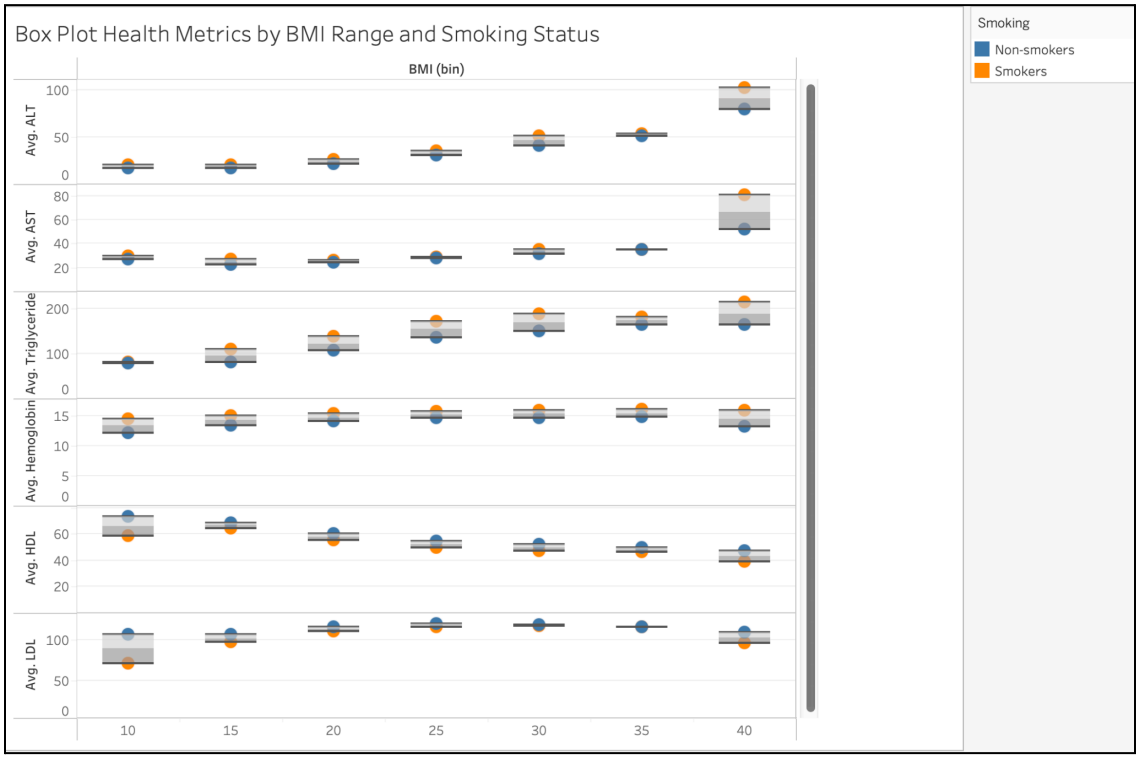


Figure 3: Box Plot of Health Metrics by BMI and Smoking.

The box plot provides a detailed comparison of health metrics, including ALT, AST, HDL, LDL, triglycerides, and hemoglobin, across different BMI categories, grouped by smoking status. This visualization highlights the distribution, median values, and outliers within each BMI bin. Smokers generally exhibit greater variability and higher median values for ALT and triglycerides, especially in the higher BMI categories, underscoring the compounded health risks

associated with smoking and obesity. The presence of outliers within the smoker group further emphasizes the potential for extreme health effects. In contrast, non-smokers tend to show more consistent distributions with fewer extreme values. Box plots are particularly effective for this analysis, as they not only illustrate the average effects of smoking on health metrics but also provide insights into the range, spread, and variability within BMI categories, offering a comprehensive understanding of health outcomes.

Bar Chart: Smoking Trends Across Age Groups

The bar chart displays the average health metric values for serum creatinine, ALT, and triglycerides across age groups for smokers and non-smokers. This visualization highlights that average values for certain health metrics are consistently higher among smokers, with the variations becoming more noticeable in older age groups. For instance, triglyceride levels in the 60-70 age range are significantly higher in smokers compared to non-smokers, illustrating the cumulative negative effects of smoking over time. The bar graphs effectively present categorical comparisons, offering a clear side-by-side view of health outcomes across age groups, making trends easy to interpret.

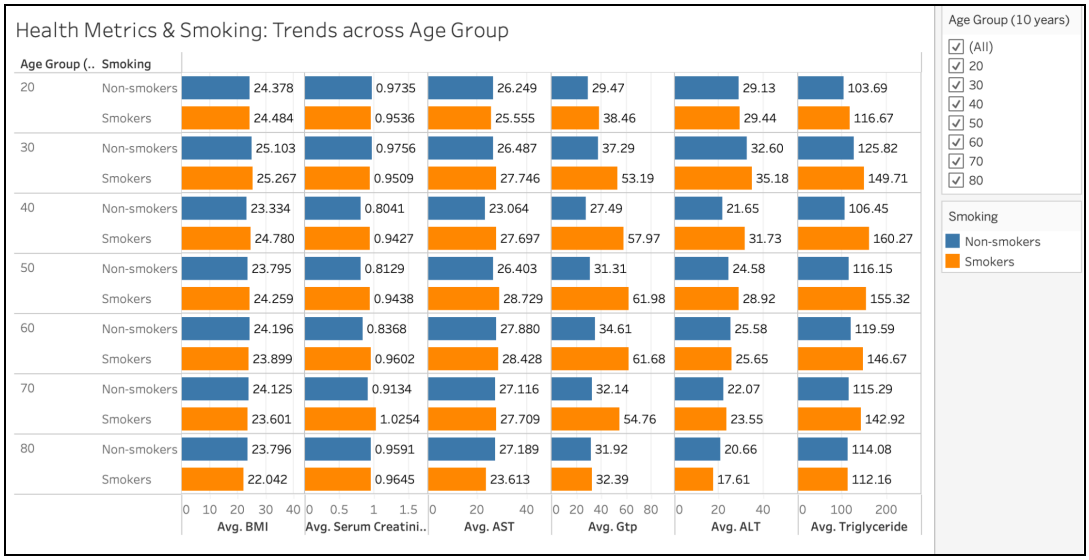


Figure 4: Bar Chart of Smoking and Health Metrics Trends by Age Group

The combination box plots and bar charts for different health metrics offer a more complete portrayal of the data. While the box plots capture the variability and distribution of health metrics within BMI categories, the bar charts focus on group averages across age ranges. Together, these visual tools offer both depth and clarity, effectively illustrating the complex relationships between smoking, BMI, and key health metrics.

3.1.2 - Anthropometric Measurement

Anthropometric measurements include various metrics used to assess physical characteristics, such as weight, height, and waist circumference. While weight and height are fundamental indicators, they are not analyzed separately in this report since Body Mass Index (BMI) already incorporates these values into a standardized calculation. Instead, the analysis focuses on waist circumference, a critical measure of abdominal fat distribution, which provides additional insights into health risks that BMI alone may not fully capture.

This section examines the relationship between BMI, waist circumference, and smoking status using a grouped bar chart and a scatter plot. These visualizations complement each other to provide a comprehensive understanding of how smoking impacts anthropometric measurements

Bar Chart: BMI and Waist Circumference by Smoking Status

The grouped bar chart compares the average waist circumference of smokers and non-smokers across different BMI categories, revealing that smokers generally have larger waist circumferences than non-smokers within the same BMI range. Notably, in the highest BMI category (BMI = 40), smokers have an average waist circumference of 123.83 cm, compared to 121.70 cm for non-smokers. This pattern persists across most BMI categories, indicating a

significant association between smoking and increased waist circumference. The bar chart effectively presents this information by providing a clear, side-by-side comparison across BMI groups, making the differences between smokers and non-smokers easy to interpret.

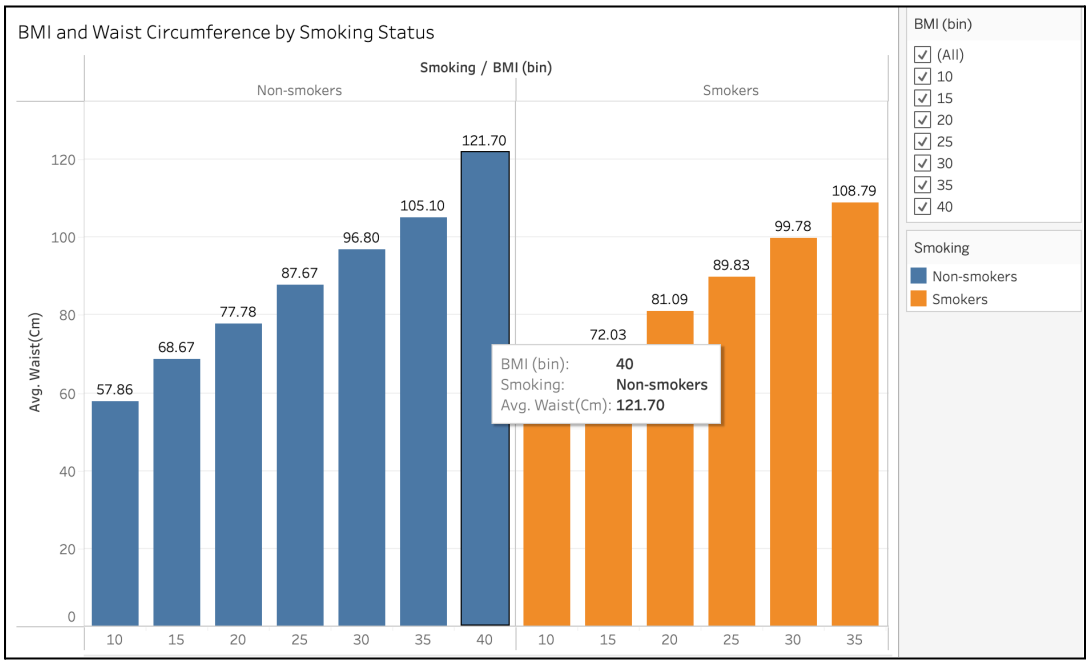


Figure 5: average waist circumference by smoking status and BMI bin

Scatter Plot: BMI vs. Waist Circumference

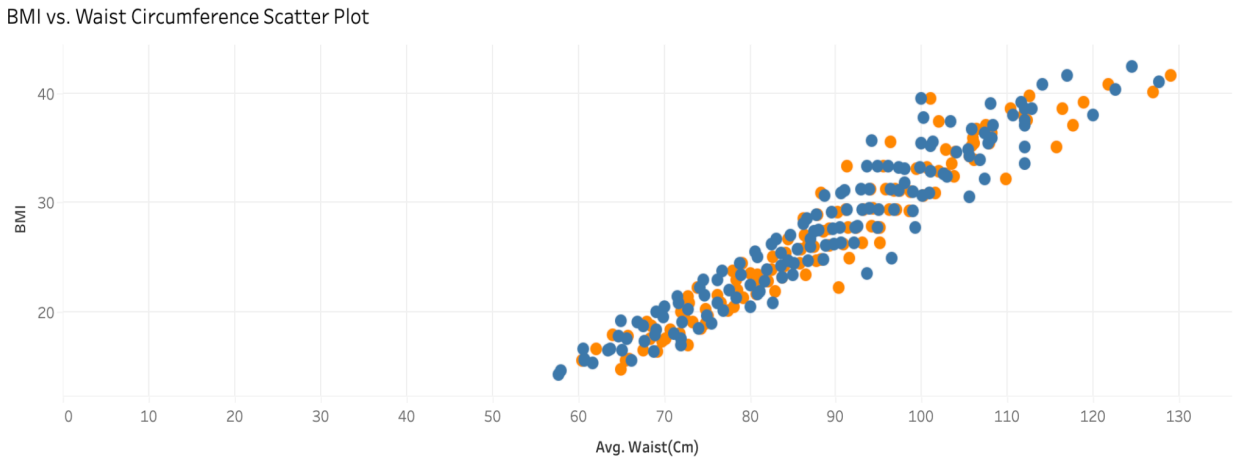


Figure 6: Scatter Plot of BMI vs. Waist Circumference by Smoking Status

The scatter plot provides a detailed representation of the relationship between BMI, and waist circumference for smokers, and non-smokers. Individual data points are plotted, with orange representing smokers and blue representing non-smokers. The plot illustrates a strong positive correlation between BMI and waist circumference. This demonstrates that as BMI increases, waist circumference also increases for both groups.

Furthermore, the scatter plot supports the bar chart by capturing variability within each group. While the bar chart summarizes group-level trends, the scatter plot visualizes individual data points, showcasing a broader distribution of waist circumferences among smokers compared to non-smokers.

The bar chart, and scatter plot together provide a comprehensive perspective. The bar chart emphasizes average trends, while the scatter plot provides a detailed view of individual data points which support better understanding the overall data. This combination highlights the influence of smoking on anthropometric measurements, with waist circumference serving as a critical indicator of abdominal obesity.

3.1.3 Liver Function Metrics

This section examines the relationship between liver health, BMI, and smoking habits using two scatter plots. These visualizations highlight how smoking influences liver health and its connection to BMI.

Scatter Plot 1: Liver Metrics, Smoking, and BMI Analysis

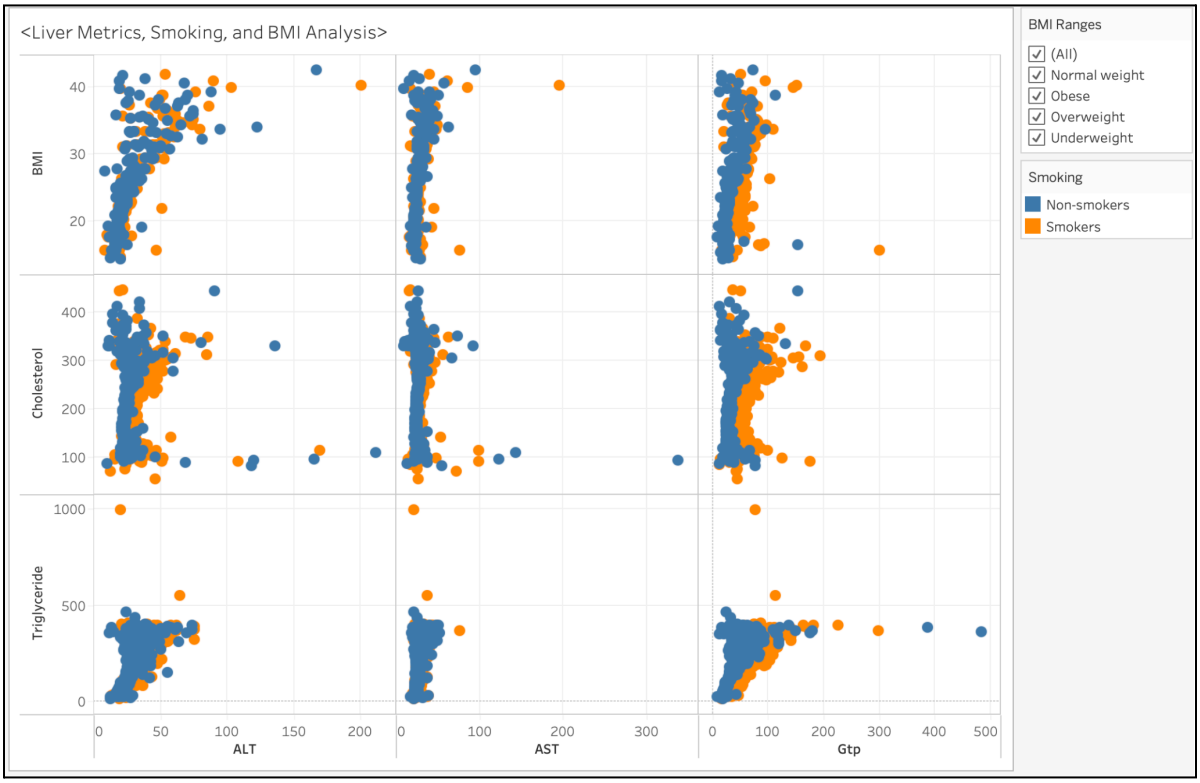


Figure 7: Scatter Plot Liver Metrics, Smoking, and BMI Analysis

The first scatter plot illustrates the relationship between liver enzymes—such as ALT (Alanine Transaminase), AST (Aspartate Transaminase), and GTP (Gamma-Glutamyl Transferase)—and BMI, categorized by smoking status. Each data point represents an individual observation, clearly differentiating between smokers and non-smokers. Clusters within the plot reveal distinct groupings, with smokers forming tighter clusters at higher enzyme levels, particularly for GTP, indicating significant liver stress. Non-smokers, in contrast, exhibit more widespread patterns, generally concentrated at lower enzyme levels, reflecting better liver health.

Understand the axis

The axes of the plot work together to reveal critical insights into liver health. The x-axis displays independent variables, representing liver enzyme levels (ALT, AST, and GTP), which are key markers of liver function. The y-axis includes metrics such as BMI, cholesterol, and triglycerides, capturing broader health outcomes. By aligning liver enzyme data (x-axis) with systemic health outcomes (y-axis), the plot effectively visualizes how liver stress, BMI, and lipid profiles are interconnected. For instance, increased GTP levels (x-axis) correspond to higher BMI (y-axis), particularly among smokers, underscoring the compounded health risks of smoking and obesity.

Data Insights from the Plot

The scatter plot also highlights differences in overall health between smokers and non-smokers. Non-smokers tend to cluster at healthier ranges of liver enzyme levels, cholesterol, and triglycerides, which are critical indicators of better metabolic and liver health. Smokers, on the other hand, exhibit tighter clusters at elevated enzyme levels, indicating greater liver stress. This pattern is particularly evident for GTP, where smokers consistently show higher levels compared to non-smokers, reinforcing the notion that smoking significantly impacts liver function.

In summary, this scatter plot emphasizes the systemic impact of smoking and high BMI on liver health. The distinct clustering patterns and the relationship between the x-axes and y-axes provide a clear narrative: smokers face compounded health risks, particularly when obesity is also a factor. Non-smokers consistently display healthier outcomes, underscoring the protective benefits of avoiding smoking on liver and metabolic health. By presenting these

metrics in a cohesive visualization, the scatter plot offers valuable insights into the detrimental effects of smoking and obesity on liver function and systemic health.

Scatter Plot 2: Liver Metrics and Metabolic Health Insights

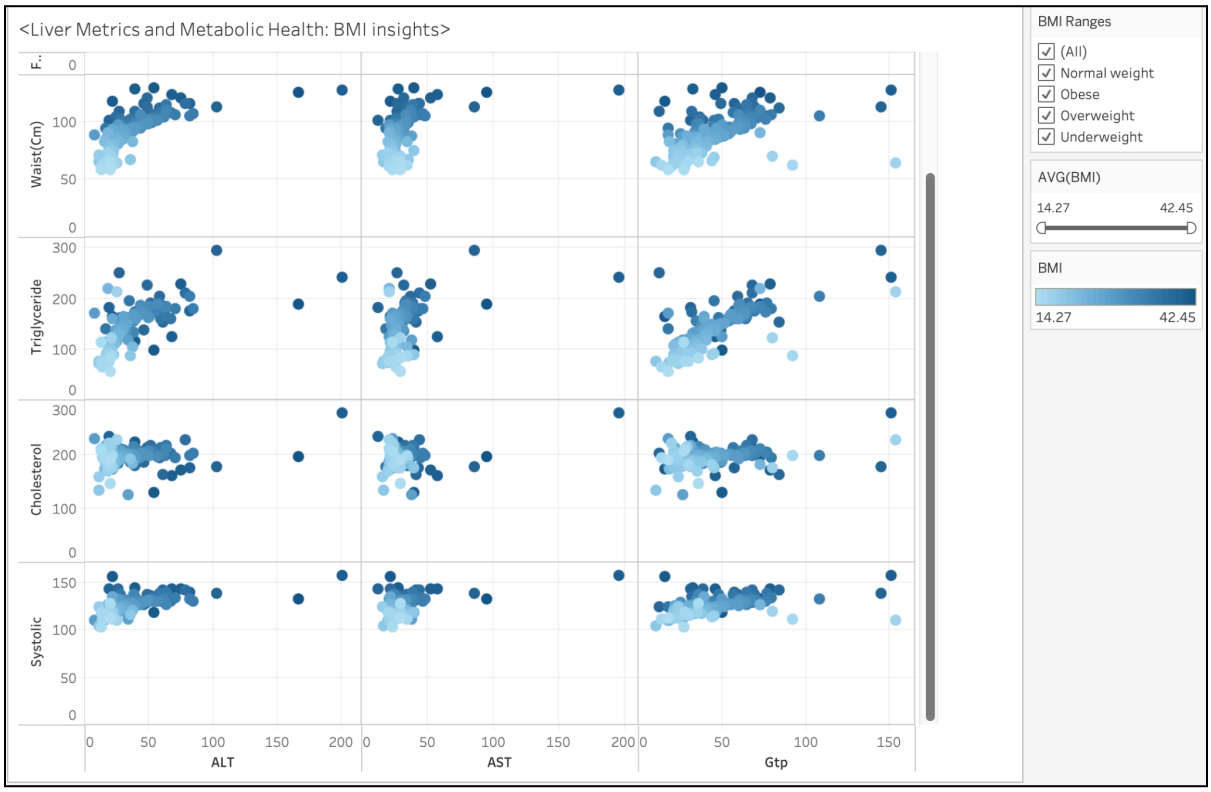


Figure 8: Scatter Plot Liver Metrics and Metabolic Health with BMI insights

The second scatter plot broadens the analysis by examining the relationship between liver function metrics—such as ALT, AST, and GTP and key metabolic health indicators like triglycerides, cholesterol, fasting blood sugar, and systolic blood pressure. This visualization provides a complete view of how these health factors interact, particularly in the context of BMI.

Understanding the axis

The x-axis in this plot represents liver enzyme levels (ALT, AST, and GTP), which are critical markers of liver function and health. These independent variables allow for an understanding of how liver health is influenced by lifestyle factors such as smoking and BMI. The y-axis, on the other hand, includes broader health outcomes like triglycerides, cholesterol, fasting blood sugar, and systolic blood pressure. These dependent variables capture the systemic impact of liver and metabolic stress, offering insights into the broader health implications.

Data Insights from the Plot

The x-axes and y-axes work together seamlessly to reveal the connections between liver function and metabolic health. For instance, higher ALT levels (x-axis) are often associated with elevated triglycerides (y-axis), particularly among individuals with higher BMI levels, indicated by darker blue dots. Similarly, GTP levels (x-axis) show a strong correlation with fasting blood sugar (y-axis), highlighting the compounded risks of liver dysfunction and metabolic stress in individuals with obesity. For example, a cluster of darker blue dots at the upper-right quadrant of the ALT and triglycerides grid illustrates individuals with both high liver enzyme levels and significant triglyceride elevations, reflecting compounding metabolic risks tied to obesity.

A notable feature of this scatter plot is the color gradient, which provides an additional layer of insight. Darker blue dots represent individuals with higher BMI levels, while lighter blue dots correspond to those with lower BMI levels. This color coding makes it easier to identify patterns, such as the clustering of darker blue dots in regions with elevated triglycerides and

cholesterol. This pattern underscores the link between obesity, liver dysfunction, and metabolic health challenges.

In summary, this scatter plot effectively integrates liver function metrics (x-axis) and metabolic health indicators (y-axis) to provide a comprehensive view of the relationships between smoking, BMI, and systemic health. The relationship between the axes, enhanced by the color gradient, makes it easier to interpret the data and draw meaningful conclusions. This visualization underscores the significant role of BMI in increasing the adverse health effects of smoking on both liver and metabolic health.

3.1.4 Cardiovascular Metrics

This section examines the relationship between BMI and cardiovascular health metrics using a detailed scatter plot. The visualization employs a color gradient, where darker blue hues represent higher BMI levels, adding depth to the data and making it easier to identify patterns and trends across key health indicators.

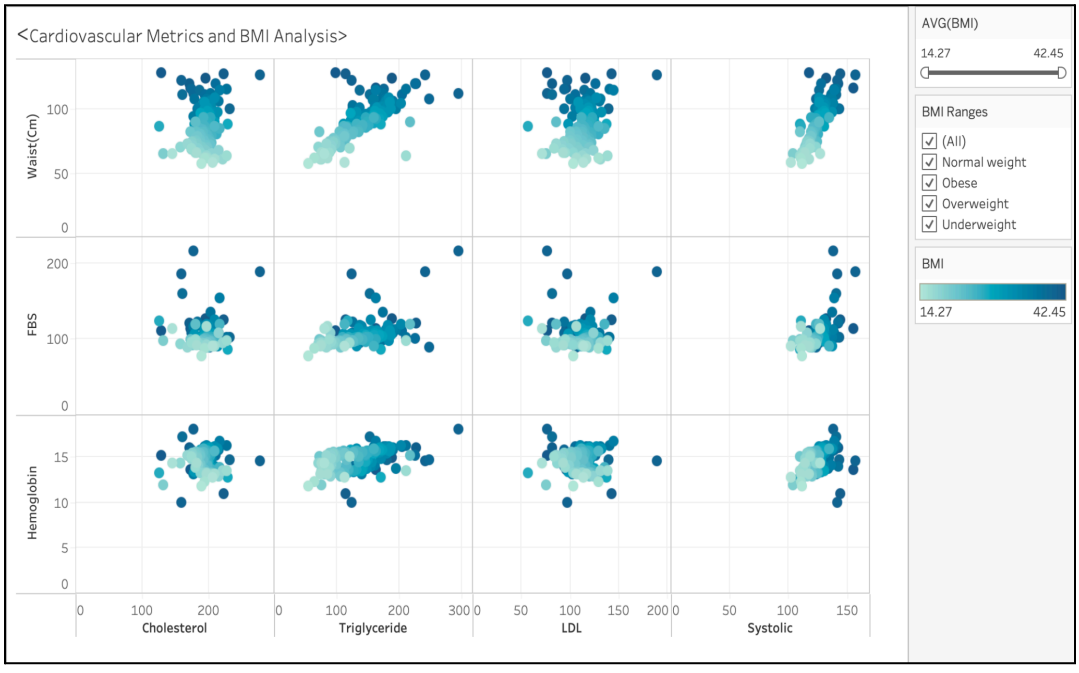


Figure 9: Scatter Plot of Cardiovascular Metrics and BMI Analysis

Understanding the Axes

The scatter plot aligns four critical cardiovascular health metrics such as cholesterol, triglycerides, LDL cholesterol, and systolic blood pressure on the x-axis, representing independent variables that influence cardiovascular and metabolic health. The y-axis focuses on outcomes such as waist circumference and fasting blood sugar, which are strongly correlated with BMI and overall metabolic conditions. This alignment creates a dynamic visualization that highlights how variations in the x-axis metrics relate to y-axis outcomes, enabling a more holistic analysis of health risks.

The combination of the x- and y-axes allows for the simultaneous examination of health stressors like triglycerides, LDL, and cholesterol alongside their associated outcomes such as fasting blood sugar and waist circumference. For example, higher triglyceride levels (x-axis) often correspond with larger waist circumferences (y-axis), particularly among individuals with higher BMI, as indicated by darker blue points. Similarly, elevated LDL cholesterol levels align with increased fasting blood sugar, illustrating how cardiovascular and metabolic health factors interact and compound the effects of obesity.

Key Observations

Cholesterol and Triglycerides: Individuals with higher BMI levels (darker blue points) tend to exhibit elevated cholesterol and triglyceride levels. These markers indicate a higher risk of cardiovascular disease, particularly in overweight and obese individuals.

LDL Cholesterol and Systolic Blood Pressure: Higher BMI levels are associated with elevated LDL cholesterol and systolic blood pressure, reinforcing the link between excess

weight, hypertension, and poor lipid profiles. These findings align with established research on the cardiovascular strain caused by obesity.

Fasting Blood Sugar and Waist Circumference: Higher BMI correlates strongly with increased fasting blood sugar, highlighting the risk of metabolic disorders like type 2 diabetes. Waist circumference, another marker of abdominal obesity, also increases significantly with BMI, further emphasizing the systemic impact of obesity on health.

3.2 - Health Metrics by BMI Categories

As highlighted in 2.2.5 - Correlation Analysis, the metrics selected for this section were chosen based on their strong correlation with BMI, making them the most relevant for analyzing the impact of smoking and BMI on health. This section organizes the analysis by dividing the health metrics into functional groups: Liver Function (AST, ALT, GTP), Cardiovascular Health (Cholesterol, LDL, Relaxation, Systolic), and Metabolic Metrics (fasting blood sugar, triglycerides). For each group, the impact of BMI is examined across two categories—(1) obese and overweight and (2) normal and underweight—while comparing the differences between smokers and non-smokers. This approach provides a structured and detailed view of how smoking and BMI interact to influence various aspects of health.

3.2.1 - Impact of BMI and Smoking on Liver Function

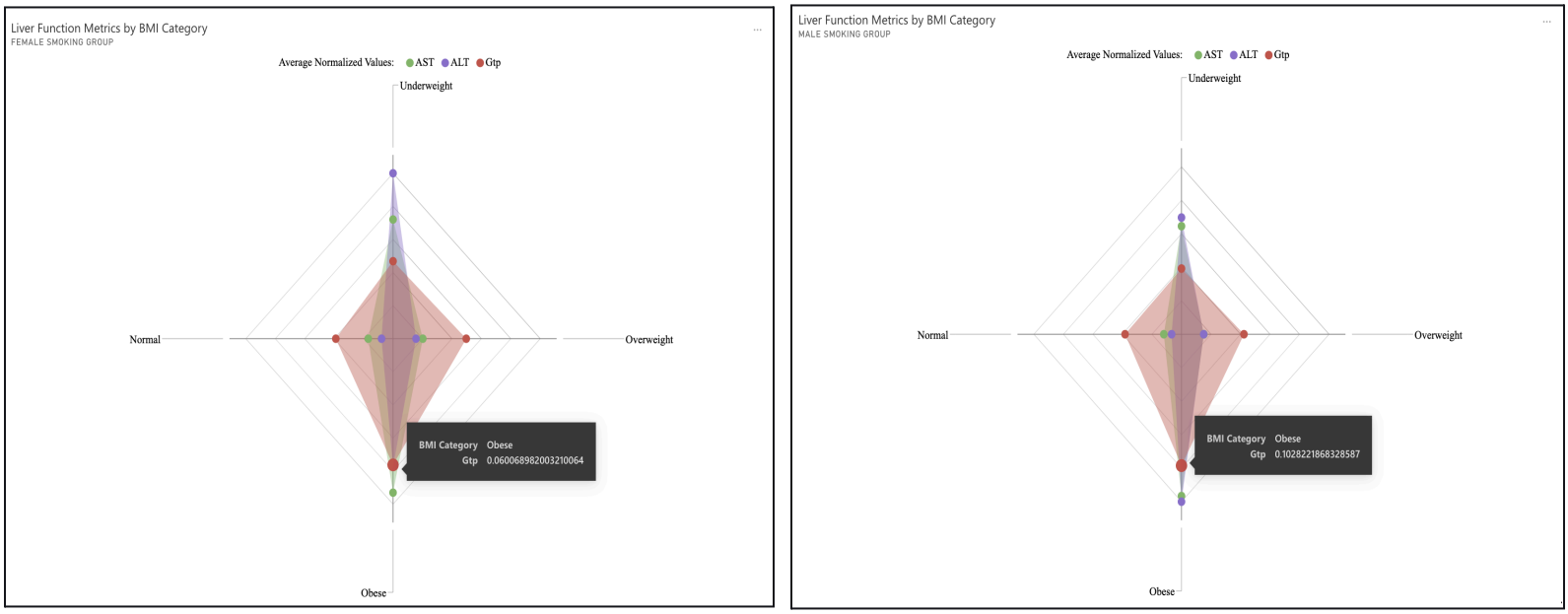


Figure 10: Radar chart comparing liver function metrics by BMI category in smoking group (female vs. male). Left: female smoking group, right: male smoking group

Radar charts are used to visualize liver function metrics (ALT, AST, and GTP) across different BMI categories—normal, underweight, overweight, and obese—while comparing smoking status and gender. Comparing the two charts, GTP levels (red) are consistently higher in men across all BMI groups, indicating that men experience greater liver stress under smoking conditions.

In the right figure 6, which represents men, the obese group shows a significant narrowing at the bottom for ALT (purple) and AST (green), reflecting a more significant reduction in these enzymes. This suggests that men's liver function may improve more in terms of ALT and AST levels, despite the higher GTP levels indicating greater liver stress.

The charts reveal that men experience higher liver stress, as shown by elevated GTP levels, but also demonstrate greater reductions in ALT and AST, suggesting better improvements

in liver enzyme metrics. These findings emphasize the importance of quitting smoking and maintaining a healthy weight to reduce liver stress and promote better liver health, particularly for men.

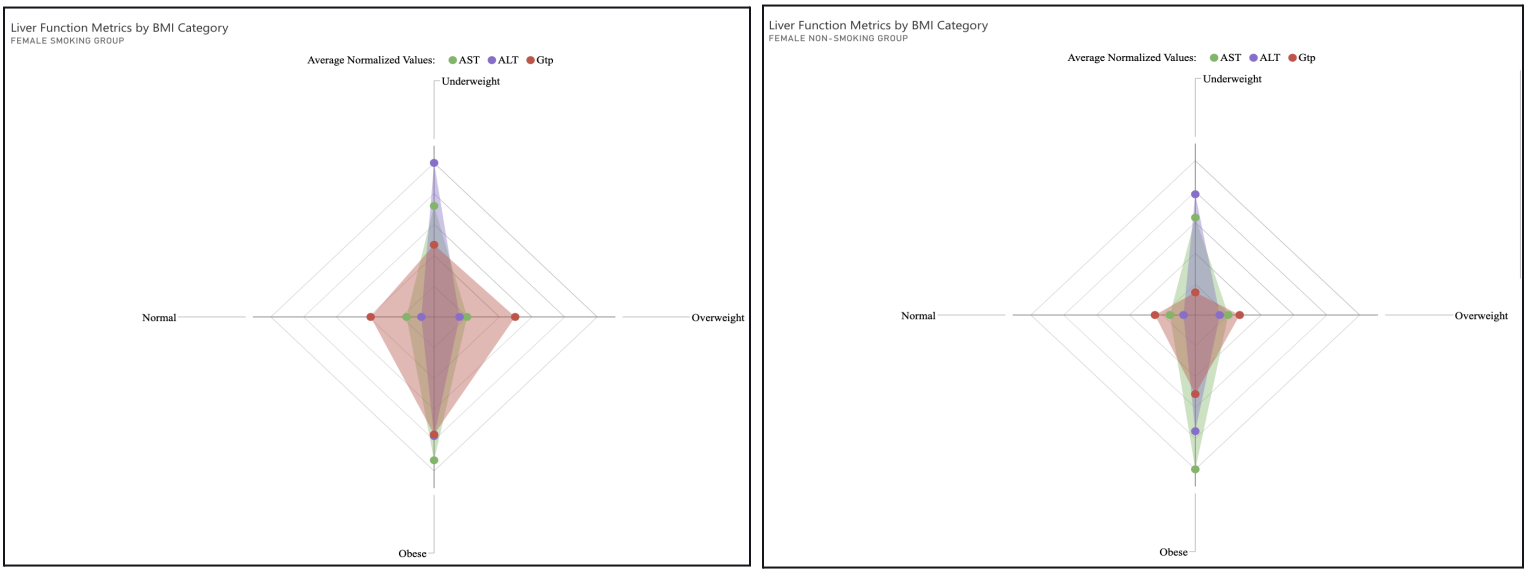


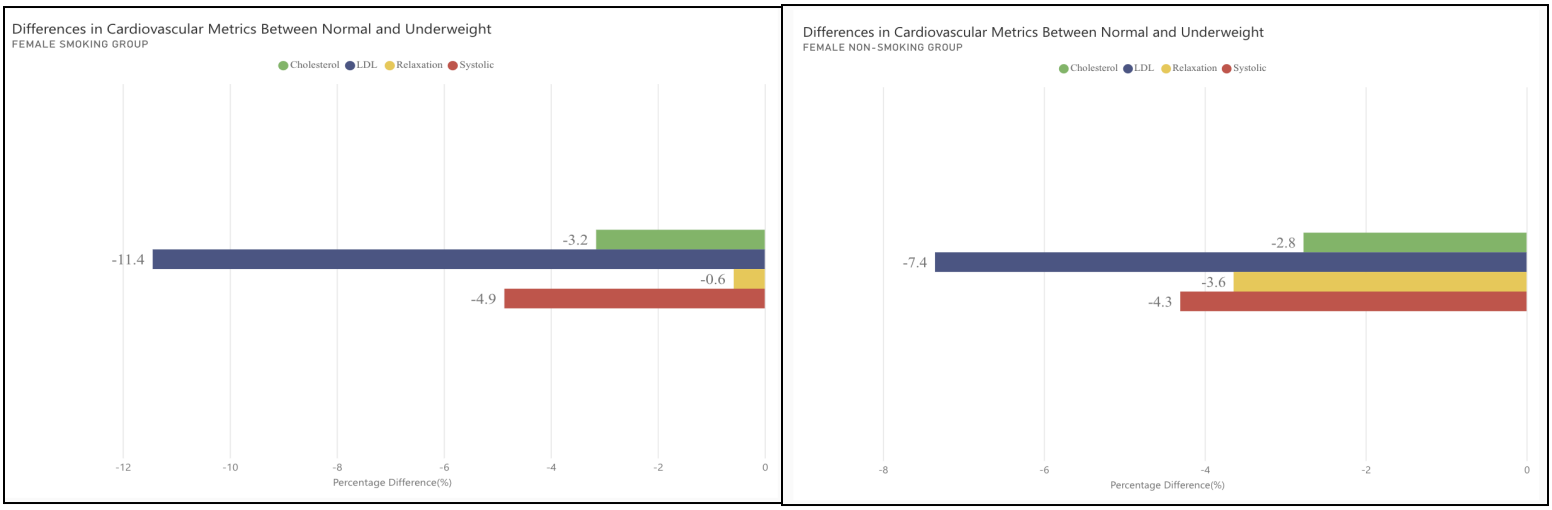
Figure 11: Radar chart comparing liver function metrics by BMI category in female group (smoking vs. non-smoking). Left: female smoking group, right: female non-smoking group.

Similarly, the radar charts compare liver function metrics between smoking females (left chart) and non-smoking females (right chart) across the same BMI categories. Obese individuals display the widest spread in both charts, signifying the most liver stress, while underweight individuals have the smallest spread, indicative of healthier liver function. Across all BMI groups, smoking females exhibit consistently higher GTP levels than their non-smoking counterparts, underscoring the harmful impact of smoking on liver health.

These findings highlight the significant influence of lifestyle factors on liver function. Smoking amplifies liver stress, particularly in individuals with higher BMI, as evidenced by elevated GTP levels. Adopting a healthier lifestyle—quitting smoking, maintaining a healthy

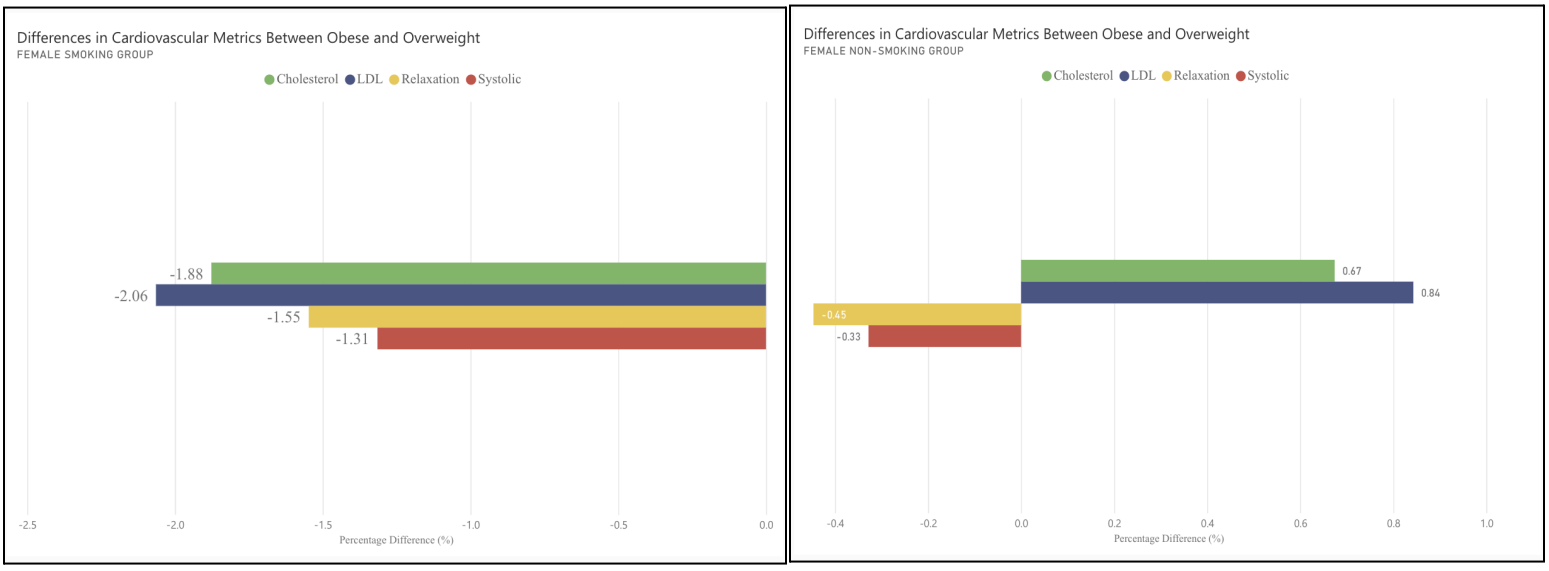
weight, limiting alcohol intake, and following a nutrient-rich diet—can significantly reduce liver stress and improve overall liver health.

3.2.2 - Impact of BMI and Smoking on Cardiovascular Function



a) Female smoking group *b) Female non-smoking group*

Figure 7: Differences in Cardiovascular metrics between normal and underweight



a) Female smoking group *b) Female non-smoking group*

Figure 12: Differences in Cardiovascular metrics between obese and overweight

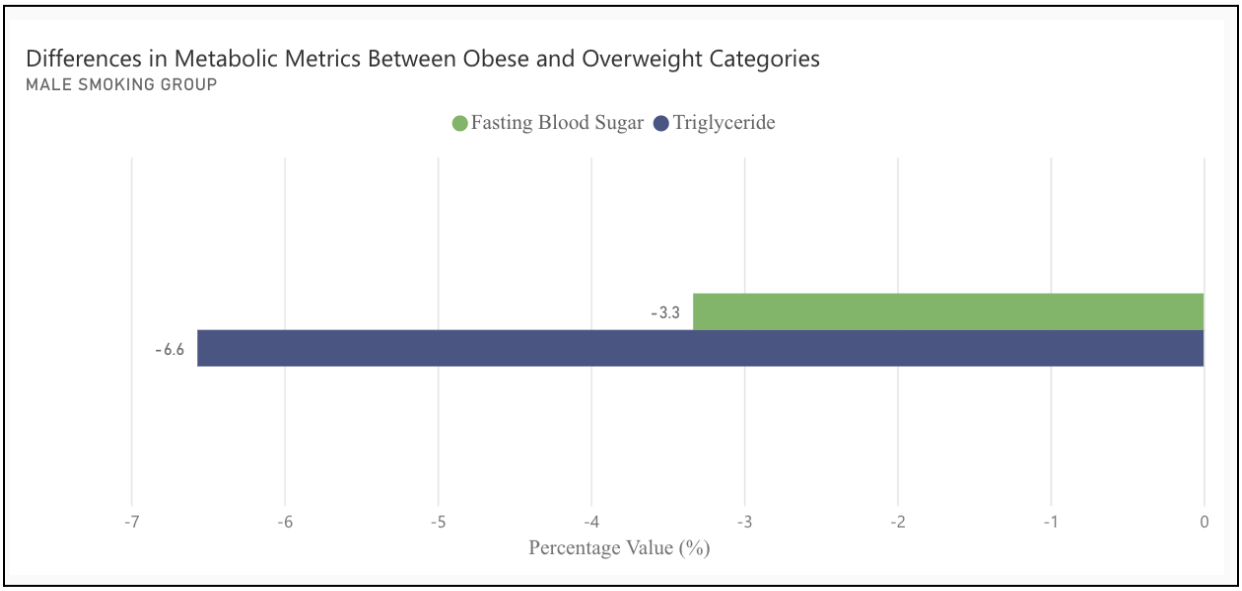
For males, the differences in cardiovascular metrics are less pronounced. Both smokers and non-smokers show similar reductions in LDL levels across BMI groups [3], indicating that smoking status has a limited impact on these metrics for men. Therefore, the focus shifts to the female group, where more notable trends emerge.

For females, the data reveals an unexpected pattern: cardiovascular metrics such as LDL, systolic blood pressure, and cholesterol decrease more in smokers than in non-smokers. For instance, in normal and underweight women, LDL levels drop by 11.4% in smokers compared to a 7.4% reduction in non-smokers. While this appears beneficial at first glance, it is important to note that these reductions may represent short-term effects and do not account for the long-term damage caused by smoking.

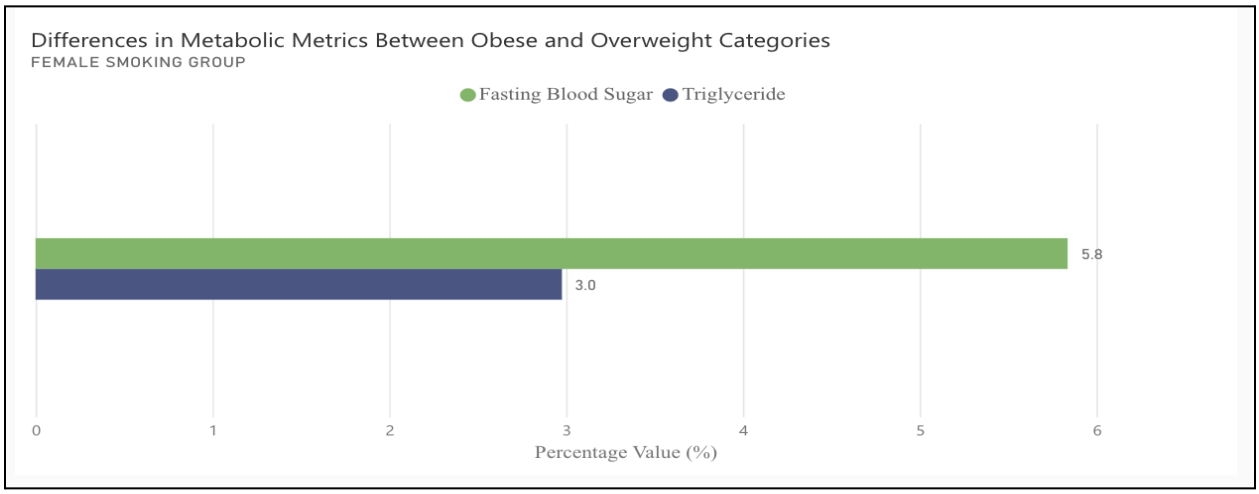
In the obese and overweight group, non-smoking women show slight increases in LDL (+0.84%) and cholesterol (+0.67%), while smoking women experience small decreases. This suggests that obesity may aggravate cardiovascular health risks, even overriding some of the benefits of a non-smoking lifestyle.

These findings highlight that while smoking may temporarily lower LDL levels, its long-term risks far outweigh any perceived benefits. To protect cardiovascular health, quitting smoking, maintaining a healthy weight, and adopting a heart-healthy lifestyle are essential, particularly for women who may face additional challenges such as obesity.

3.2.3 - Impact of BMI and Smoking on Metabolic Metrics



a) Male smoking group



b) Female smoking group

Figure 13: Differences in metabolic metrics between obese and overweight categories

Metabolic metrics, such as fasting blood sugar and triglycerides, are healthier when their levels are lower, and, as anticipated, these metrics are generally lower in non-smokers compared to smokers. However, notable gender differences emerge within the obese and overweight categories. Among smokers, fasting blood sugar shows a significant increase in females (+5.84%) but a decrease in males (-3.33%). Similarly, triglycerides increase slightly in smoking

females (+2.98%) but decrease substantially in smoking males (-6.57%). These findings highlight the complex interplay between smoking, gender, and metabolic health. Maintaining a healthy lifestyle—quitting smoking, staying active, achieving a healthy weight, and following a balanced diet—is essential, particularly for obese individuals and women, who may face greater challenges in managing these metabolic risks.

4 - Proposed Method

4.1 - Regression

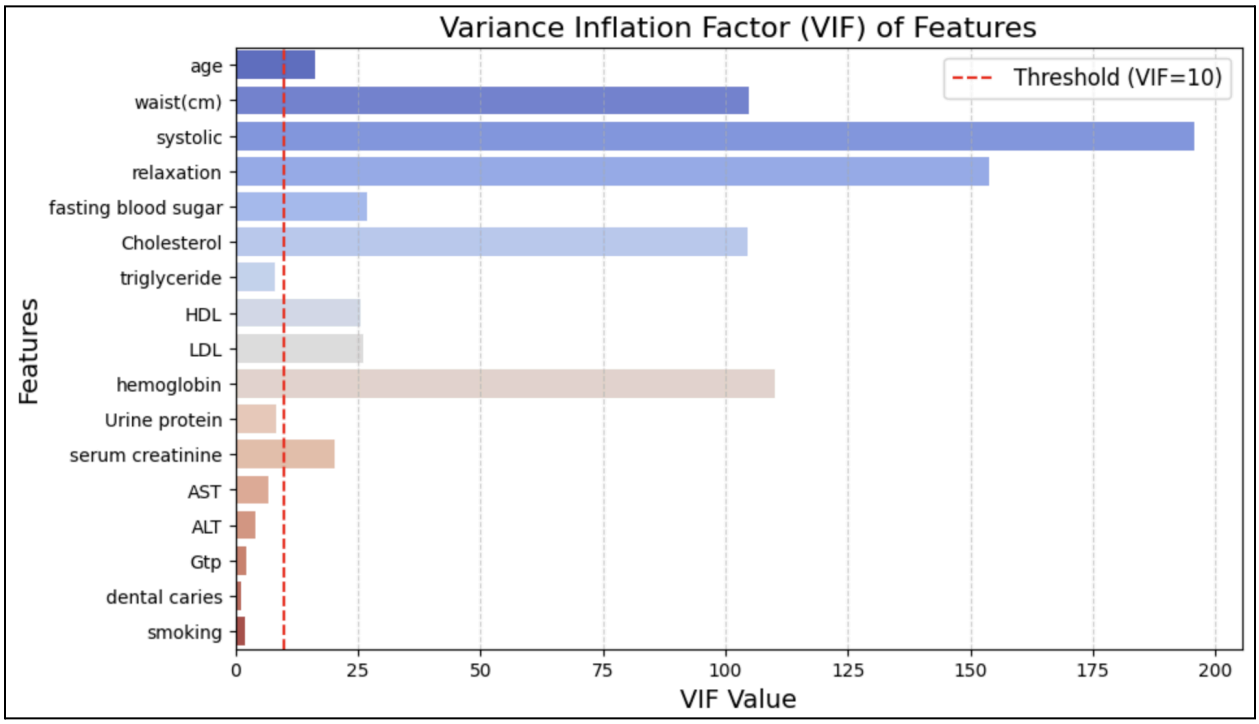
We aimed to predict BMI using smoking data. Smoking-related variables often influence BMI, and understanding these relationships can provide insights into health outcomes. But we faced some challenges such as multicollinearity and optimizations.

4.1.1 - Challenges: Multicollinearity

Multicollinearity arises when independent variables are highly correlated, which can make it challenging for a model to accurately estimate the individual effect of each variable. To detect multicollinearity, the Variance Inflation Factor (VIF) is commonly used as a diagnostic tool. A VIF value greater than 10 is typically considered a threshold indicating problematic multicollinearity, serving as a rule of thumb to identify variables that may require further examination or adjustment.

4.1.2 - Solutions and Methods

To address multicollinearity, Ridge Regression was implemented, introducing a penalty term to shrink the coefficients of correlated features, thereby reducing their impact on the model and enhancing stability. Additionally, data refinement techniques were applied to improve feature relationships and model interpretations. The dataset was grouped by gender to identify differences and provide clearer insights, while K-means clustering was used to segment the data into more homogeneous groups, improving feature alignment and strengthening the overall analysis.



4.1.3 - Model Evaluation

We compared the performance of different regression models and optimization techniques using metrics such as Mean Squared Error (MSE) and R-squared:

Model	Splitting data method	MSE	R-squared
Linear	ALL	3.8160	0.6882
Ridge	ALL	3.7127	0.6964
Ridge	By gender	3.5521	0.7107
Ridge	By using k-mean	3.6033	0.6813

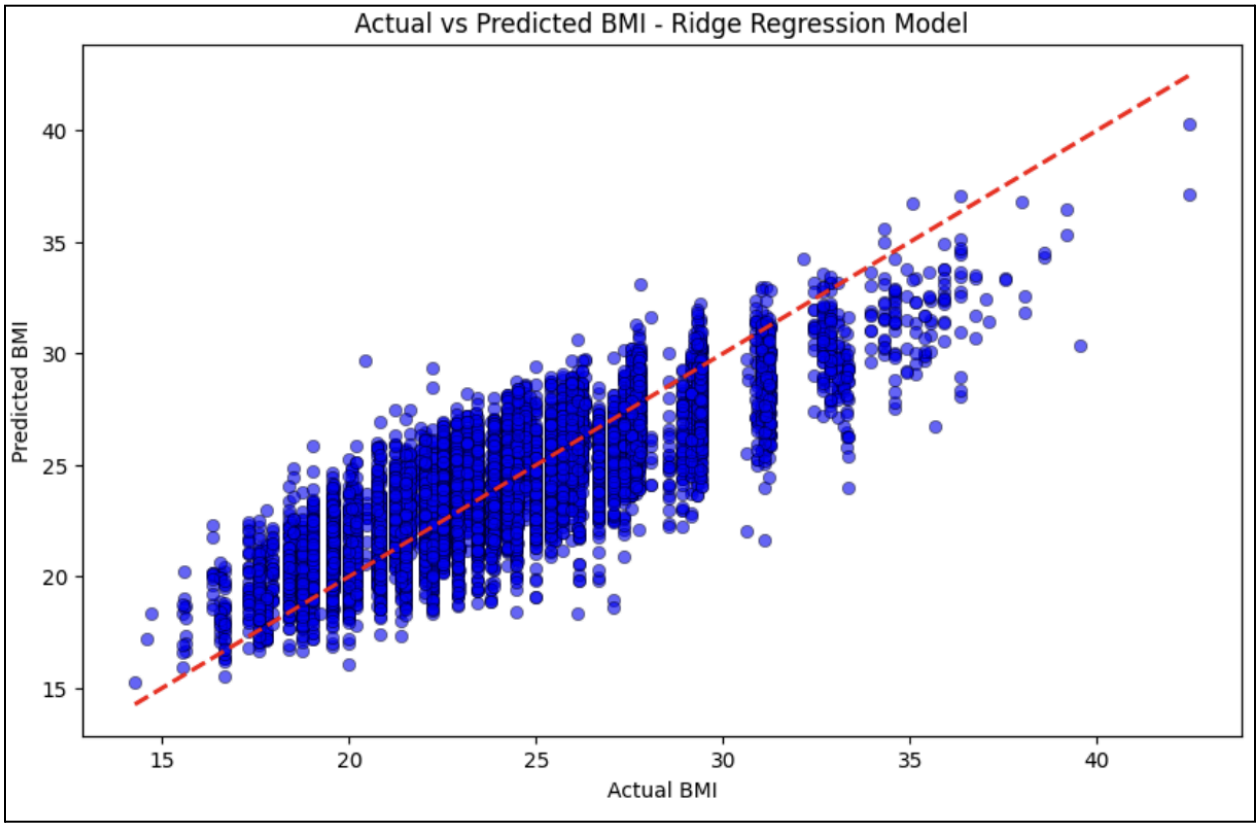


Figure 14: Comparison of Actual vs Predicted BMI Using Ridge Regression

4.1.4 - Conclusion

Ridge Regression outperformed standard linear regression by addressing multicollinearity. Grouping data by gender achieved the best performance with the lowest MSE (3.5521) and the highest R-squared (0.7107). K-means clustering also improved model performance but was slightly less effective than grouping by gender.

4.2 - Classification

4.2.1 - WHtR-Based Risk (WHtR)

WHtR-Based Risk helps assess health risks based on the Waist-to-Height Ratio (WHtR), which is a measure that estimates body fat distribution and potential health risks including cardiovascular disease, diabetes, and some cancers. WHtR equals the waist circumference divided by the height in the same unit (centimeters or inches). The following range is guidelines [4] for taking action:

$$\text{WHtR} < 0.5\text{: Low risk (healthy range)}$$

$$\text{WHtR} \geq 0.5 \text{ to } < 0.6\text{: Moderate risk}$$

$$\text{WHtR} \geq 0.6\text{: High risk}$$

4.2.2 - PCA-Based Classification:

The objective is to apply Principal Component Analysis (PCA) for dimensionality reduction while minimizing information loss and to evaluate the performance of multiple classification algorithms with reduced features. The target variable for classification was the "WHtR-Based Risk."

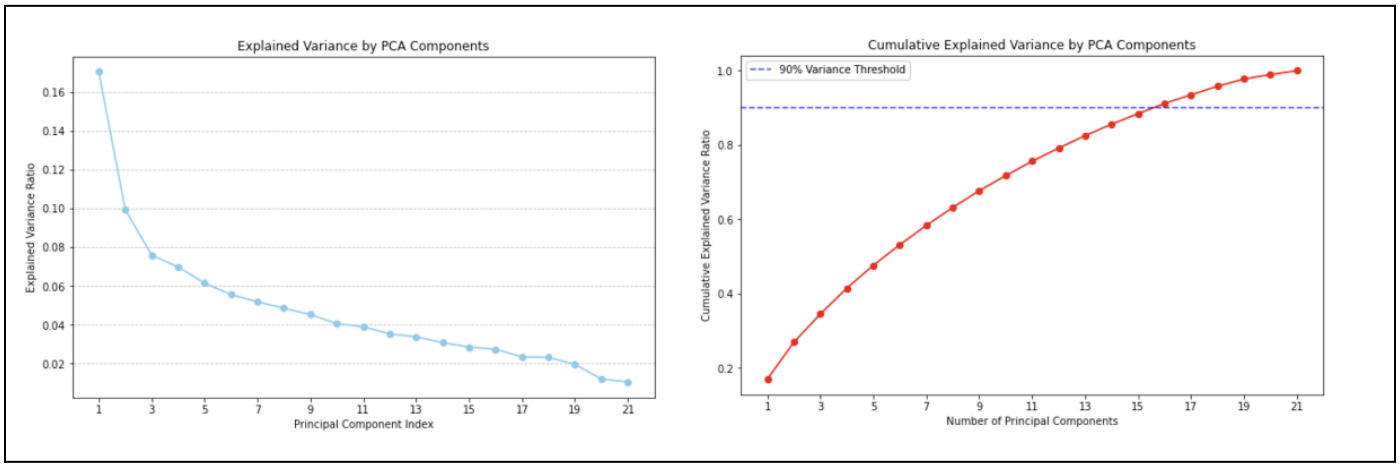


Figure 15: Explained Variance by PCA Components(left) and Cumulative Explained Variance by PCA Components(right)

According to the plots of explained and cumulative explained variance, the first few principal components captured the majority of the variance in the data. The first 16 components explained over 90% of the variance for the data information.

4.2.3 - PCA-Based Model Evaluation

The following machine learning models were evaluated using the reduced feature set (16 PCA components): Logistic Regression, Random Forest, k-Nearest Neighbors (k-NN), Decision Tree.

Model	Accuracy	Precision (1)	Precision (0)	Recall (1)	Recall (0)	F1-Score (1)	F1-Score (0)
Logistic Regression	0.82	0.82	0.82	0.79	0.85	0.81	0.83
Random Forest	0.87	0.86	0.87	0.86	0.88	0.86	0.87
k-NN	0.79	0.79	0.79	0.76	0.81	0.77	0.80
Decision Tree	0.81	0.80	0.82	0.80	0.82	0.80	0.82

Table 3: Performance Metrics for Classification Models

Random Forest has the best performance with an accuracy of 87% which may be because of its robustness with reduced feature sets. k-NN has relatively lower performance due to the fact that it has the sensitivity to high-dimensional data even after PCA.

4.2.4 - Optimizing models using feature importance for WHtR-Based Risk

The objective is to determine the most significant features influencing WHtR-based risk. A Random Forest Regressor can be used to evaluate feature importance for predicting WHtR-based risk.

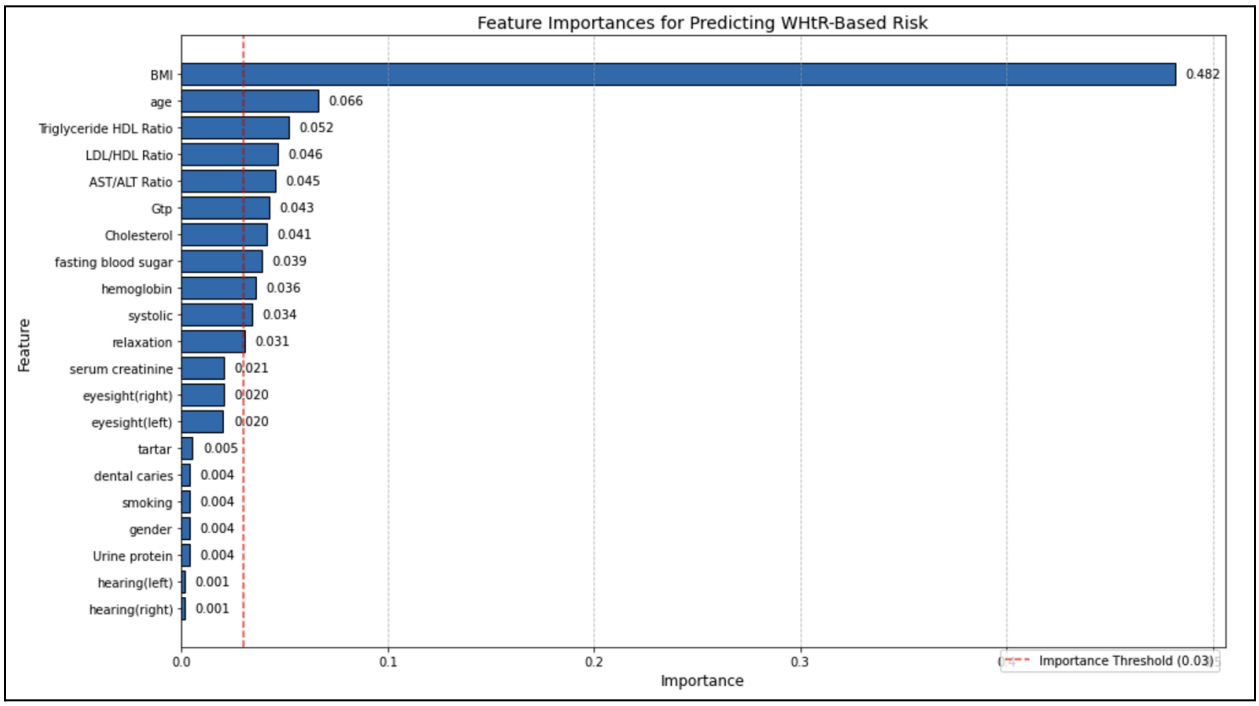


Figure 16: Feature Importances for Predicting WHtR-Based Risk.

Based on the plot of Random Forest feature importance, the following features are selected (over 0.03) for further modeling: BMI, Age, Triglyceride HDL Ratio, LDL/HDL Ratio,

AST/ALT Ratio, GTP, Cholesterol, Fasting Blood Sugar, Hemoglobin, Systolic Blood Pressure, and Relaxation.

4.2.5 - Model Evaluation

The following machine learning models were evaluated using the reduced feature set (using Random Forest feature importance): Logistic Regression, Random Forest, k-Nearest Neighbors (k-NN), Decision Tree

Model	Accuracy	Precision (1)	Precision (0)	Recall (1)	Recall (0)	F1-Score (1)	F1-Score (0)
Logistic Regression	0.83	0.83	0.83	0.81	0.85	0.82	0.84
Random Forest	0.88	0.89	0.88	0.86	0.90	0.87	0.89
k-NN	0.81	0.80	0.81	0.79	0.82	0.79	0.82
Decision Tree	0.84	0.83	0.85	0.83	0.84	0.83	0.85

Table 4: Enhanced Performance Metrics for Classification Models

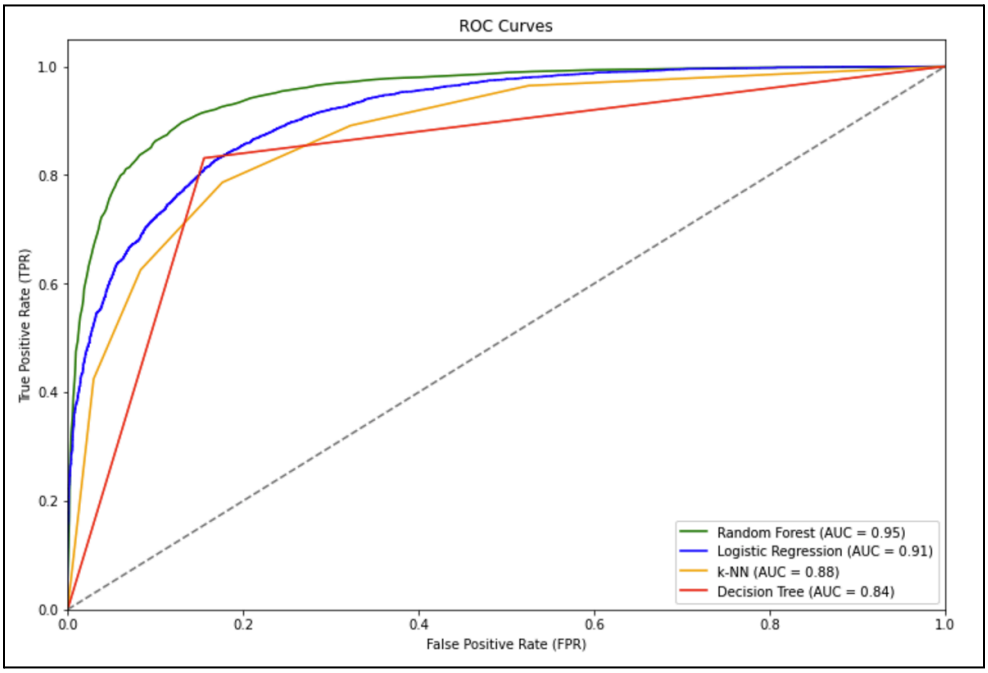


Figure 17: ROC Curves and AUC

Every model has improved after using these selected features, and the plot showcases the ROC (Receiver Operating Characteristic) for the different classification models after using the selected features with the respective AUC (Area Under the Curve).

Model	AUC	Performance
Random Forest	0.95	Best-performing model. The green curve is close to the top-left corner, indicating strong classification with low false positives and high true positives.
Logistic Regression	0.91	Second-best model. The blue curve is slightly below Random Forest but significantly better than random guessing (AUC = 0.5).
k-NN	0.88	Moderate performance. The orange curve is farther from the top-left corner compared to Random Forest and Logistic Regression.
Decision Tree	0.84	Least effective model. The red curve is closest to the diagonal, indicating weaker performance but still better than random guessing (AUC = 0.5).

The Random Forest Classifier is a reliable model for predicting WHtR-Based Risk since it has an accuracy of 88% and an AUC score of 0.95. It is suitable for identifying individuals at higher risk based on their health metrics.

5 - Discussion and Conclusion

The findings from this study reveal important findings into the relationship between smoking, BMI, and health metrics. Smokers consistently demonstrate higher BMI levels compared to non-smokers, suggesting a potential link between smoking and increased body mass. Smoking not only influences external measures like BMI but also internal markers such as liver enzymes and cardiovascular health indicators, demonstrating its extensive interconnected effects. Additionally, demographic factors, such as gender and age, significantly shape these patterns, underscoring the importance of customizing health strategies to specific population groups.

The analysis of health metrics by BMI categories highlights the combined influence of smoking and BMI on liver, cardiovascular, and metabolic health. Smoking amplifies liver stress, particularly in obese individuals, as evidenced by elevated GTP levels, while men demonstrate greater improvements in ALT and AST levels despite higher stress. In women, cardiovascular metrics revealed unexpected short-term decreases in LDL and cholesterol among smokers; however, obesity worsens risks, overriding these temporary benefits. Metabolic metrics also revealed gender differences, with smoking females experiencing increases in fasting blood sugar and triglycerides, while males exhibited decreases. These findings emphasize the need for adopting healthier lifestyles, including quitting smoking, maintaining a healthy weight, staying active, and following a balanced diet, to mitigate the risks associated with smoking and obesity.

Correlations among health indicators further support these findings. Smokers with higher BMI exhibited pronounced stress patterns in liver function, while the interaction between smoking and BMI revealed complex cardiovascular health impacts. Metabolic metrics displayed

distinct trends across population groups, highlighting the variability in health outcomes based on demographic factors. Predictive modeling, particularly using the Random Forest Classifier, achieved an 88% accuracy rate in predicting WHtR-based risks, identifying key health indicators for early detection and validating BMI and related metrics as reliable predictors of smoking-related health risks.

Together, these findings underline the urgent need for targeted health programs and personalized strategies to address the combined challenges of smoking and obesity on health. This research not only deepens our understanding of the interplay between lifestyle choices and health outcomes but also provides actionable tools, such as predictive models, to assess and prevent smoking-related health risks, supporting both healthcare professionals and individuals in improving long-term health outcomes.

References

[1] Mustanger. “Smoking Signal of Body Classification.” Kaggle, *https://www.kaggle.com/code/eisgandar/smoking-signal-of-body-classification/input*.
[2] “Body Signal of Smoking Dataset.” *e-Government Website of the Republic of Korea*. Kaggle, *https://www.data.go.kr/data/15007122/fileData.do*.
[3] “ *Impact of BMI and Smoking on Cardiovascular Function chart*”. Power Bi https://app.powerbi.com/groups/me/dashboards/4c16df0e-004b-4b28-afff-aa9dd8e55302?ctid=e85c5307-76b1-4c48-bc5d-e88373dda261&pbi_source=linkShare
[4] Marinos, Sarah. “*Why Your Waist-to-Height Ratio Is a Good Measure of Health*.” Baker Heart and Diabetes Institute, The House of Wellness, *https://www.baker.edu.au/news/in-the-media/waist-height-ratio#:~:text=According%20to%20research%2C%20a%20healthy,the%20highest%20risk%20of%20disease.3*.
[5] “*Health Metrics and Data Overview*”, Tableau https://public.tableau.com/app/profile/khac.minh.dai.vo/viz/FinalProject_17321748868400/AnthropometricMeasurements?publish=yes