Predictive Modeling Of Chronic Health Conditions Using Lifestyle And Physiological Factors

Introduction

Chronic health conditions such as diabetes, hypertension, asthma, arthritis, and obesity affect millions of people and place a major burden on healthcare systems. These conditions often develop gradually, and early signs can be difficult to detect without analyzing patterns in lifestyle and physiological factors.

With the growing availability of health data, machine learning offers an opportunity to identify these patterns more accurately and support early detection. This project applies data science techniques to predict different chronic health conditions using key behavioral and physiological indicators. The aim is to explore how well predictive models can classify individuals based on their health profiles and to identify the most important factors influencing these predictions.

This report presents the full modeling process (from data preparation to model evaluation) and provides insights on how data-driven approaches can support better decision-making in healthcare.

Problem Statement

Chronic health conditions remain a major public health concern, often progressing silently until they become difficult and expensive to manage. Traditional diagnostic methods rely heavily on clinical assessments, which may not fully capture the subtle lifestyle and physiological patterns that precede these conditions. With growing datasets containing patient information, there is a need to understand whether machine learning models can accurately predict chronic health conditions using such features. This project addresses the challenge of building reliable predictive models that can classify individuals into specific health categories based on measurable attributes.

Aim

The aim of this project is to develop and evaluate machine learning models capable of predicting chronic health conditions using lifestyle and physiological variables.

Objectives

1. To explore, clean, and structure the dataset to ensure it is suitable for machine learning analysis.
2. To perform feature engineering and selection to identify the most relevant variables influencing health outcomes.

3. To build and tune predictive models, including Decision Tree, Random Forest and XGBoost classifiers.
4. To evaluate model performance using accuracy, macro precision, macro recall, macro F1-score, and confusion matrices.
5. To interpret the results, identifying important predictors and assessing how well each model distinguishes between different chronic health conditions.
6. To provide recommendations for how predictive modeling can support early detection and healthcare planning.

Definition of Terms, Features, and Target Variable

This section explains the key features and concepts used throughout the project.

Target Variable

- **Medical Condition**
  This is the main outcome the model predicts. It is a multi-class variable with seven possible categories: Arthritis, Asthma, Cancer, Diabetes, Healthy, Hypertension, and Obesity.

Predictor Variables (Features)

These are the attributes used by the machine learning models to learn patterns that may indicate a specific health condition.

- **Age** – The individual's age in years.
- **Gender** – Biological sex (Male/Female).
- **Glucose** – Blood glucose level, an important indicator for metabolic disorders such as diabetes.
- **Blood Pressure** – Measured arterial pressure, linked to hypertension and cardiovascular risk.
- **BMI** – Body Mass Index, reflecting body weight relative to height.
- **Oxygen Saturation** – Level of oxygen in the blood, often used to assess respiratory and cardiac function.
- **Length Of Stay** – Duration of hospitalization or medical observation, where applicable.
- **Cholesterol** – Total cholesterol level, associated with heart and metabolic diseases.
- **Triglycerides** – Blood fat levels, an important risk indicator for cardiovascular health.
- **HbA1c** – A measure of average blood glucose over time; often used to diagnose and monitor diabetes.
- **Smoking** – Smoking status or frequency, an important behavioral risk factor.

- **Alcohol** – Alcohol consumption level, another lifestyle-related health indicator.
- **Physical Activity** – Frequency or intensity of exercise or movement.
- **Diet Score** – A measure of dietary quality based on nutritional patterns.
- **Family History** – Indicates whether close relatives have chronic health conditions.
- **Stress Level** – Self-reported or measured stress rating.
- **Sleep Hours** – Average number of hours of sleep per day.

Key Machine Learning Terms

- **Model:** A mathematical representation that learns patterns from data to make predictions.
- **Training Set:** Portion of the data used to teach the model.
- **Test Set:** Data used to evaluate how well the model performs on unseen cases.
- **Precision, Recall, F1-score:** Metrics used to measure how accurate and reliable the model's predictions are, especially in multi-class problems.
- **Macro Average:** An average of metrics across all classes, treating each class equally.

Data Quality Assessment

A data quality review was conducted to evaluate the integrity and readiness of the dataset prior to analysis. The dataset includes 18 variables, 17 predictors spanning demographic, lifestyle, and clinical attributes, and one target variable (Medical Condition).

Key checks performed include:

- **Missing Values:** Several features contained missing entries. These were quantified to guide the appropriate imputation strategy during preprocessing.
- **Duplicate Records:** The dataset was scanned for duplicate observations to prevent data redundancy and avoid inflating model performance. Any duplicates identified were removed.
- **Outliers:** Numerical variables such as Glucose, BMI, and Blood Pressure were assessed for extreme values that could distort trends or impair model reliability.
- **Class Imbalance:** The target variable exhibited uneven distribution across classes. This imbalance was factored into model development to ensure fair and unbiased predictions.

The assessment confirmed areas requiring corrective preprocessing and provided a clear foundation for subsequent analysis and modeling.

Data Preprocessing

The dataset was preprocessed to ensure consistency, reliability, and suitability for machine learning models. Key steps included:

1. **Handling Missing Values:**
    - Numerical features with few missing entries were imputed using the median to reduce the influence of extreme values.
    - Categorical variables were imputed with the most frequent category.
2. **Removing Duplicates:**
    - Duplicate observations were removed to ensure each patient was represented only once, preventing bias and inflated accuracy metrics.
3. **Outlier Treatment:**
    - Outliers in numerical features such as Glucose, Blood Pressure, BMI, and Triglycerides were identified using statistical thresholds (e.g., IQR) and either capped or removed to stabilize model training.

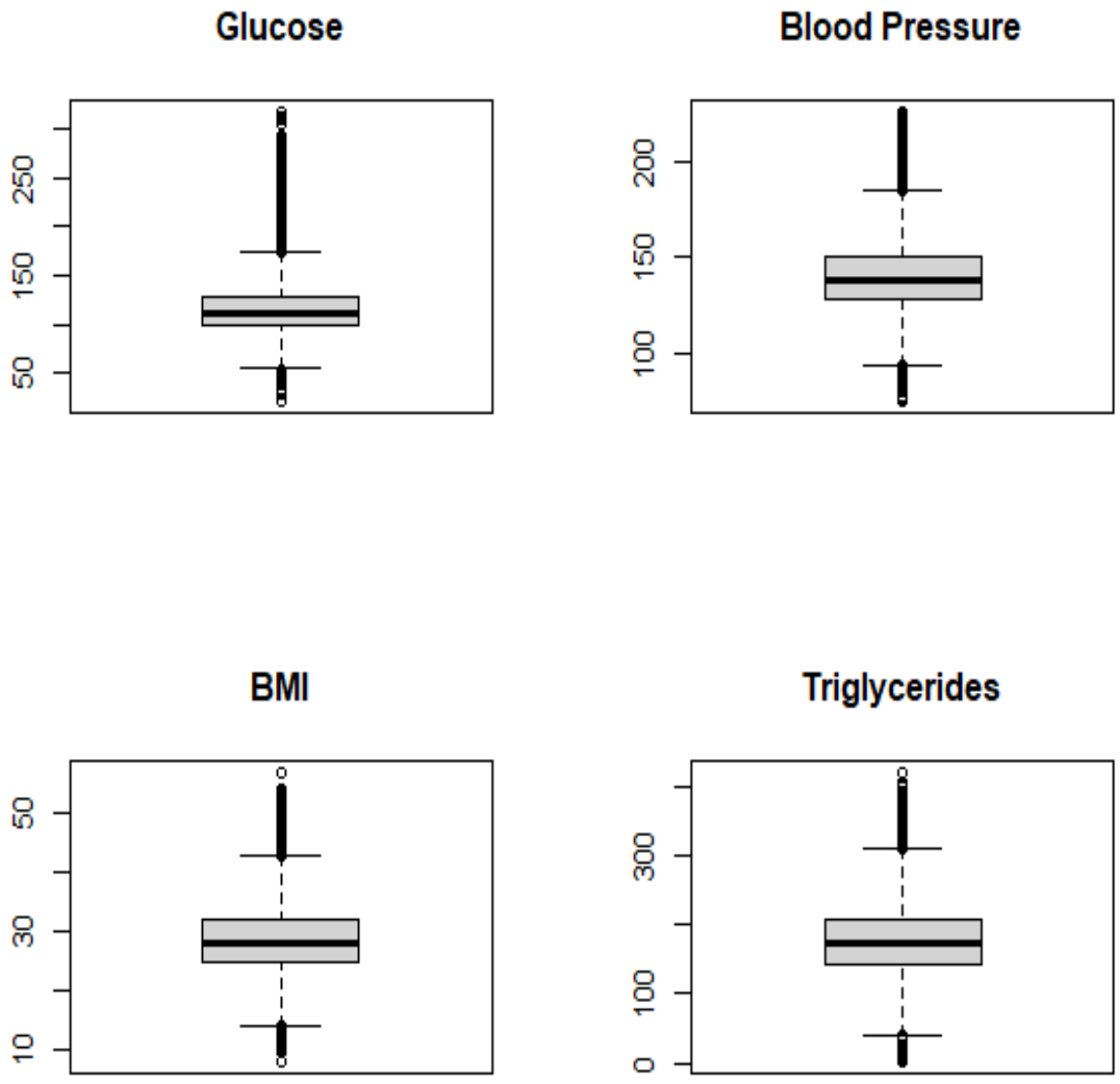

Glucose
50
150
250
Blood Pressure
100
150
200
BMI
10
30
50
Triglycerides
0
100
300

Figure 1

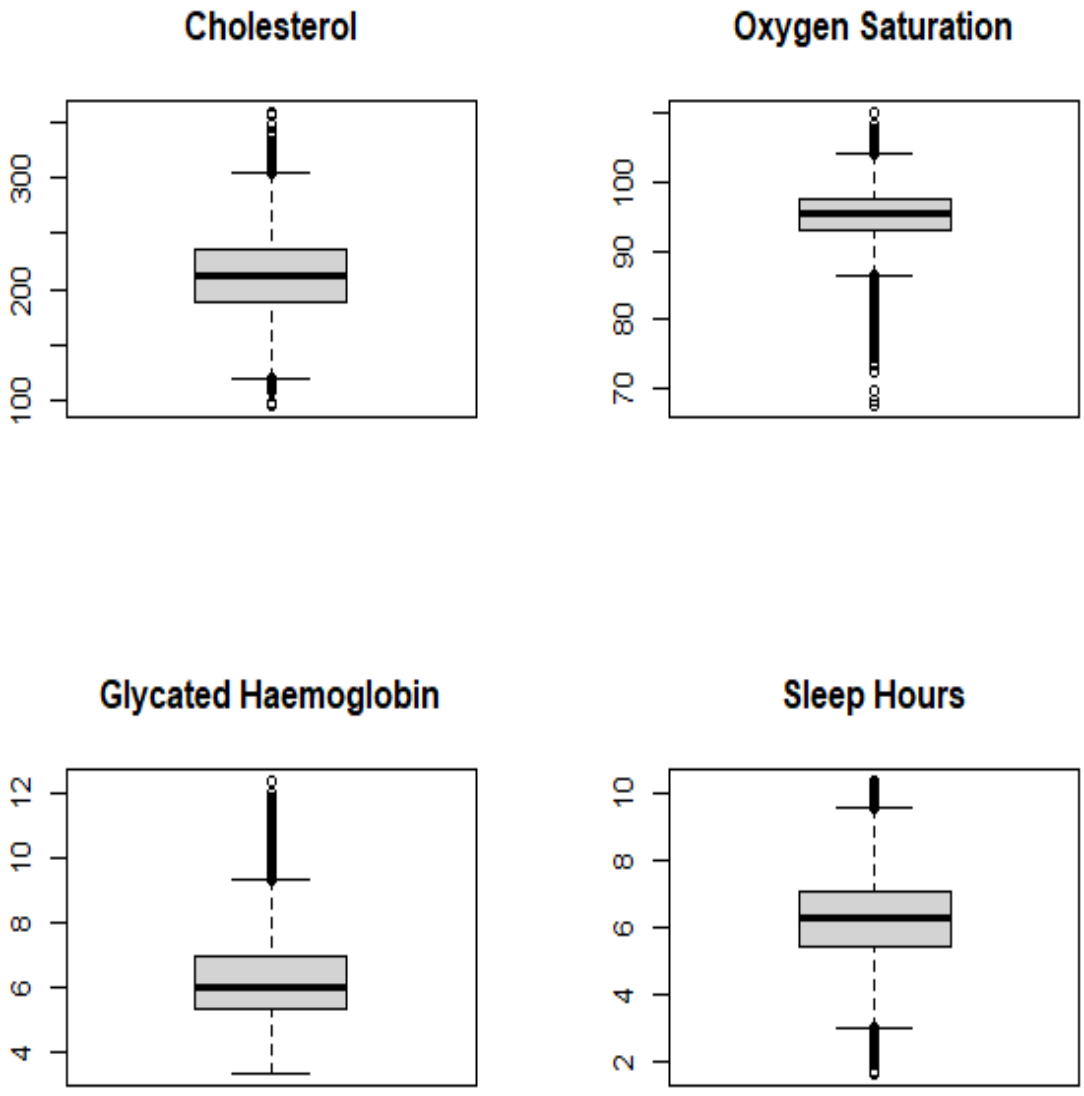

Cholesterol
100
200
300
Oxygen Saturation
70
80
90
100
Glycated Haemoglobin
4
6
8
10
12
Sleep Hours
2
4
6
8
10

Figure 2

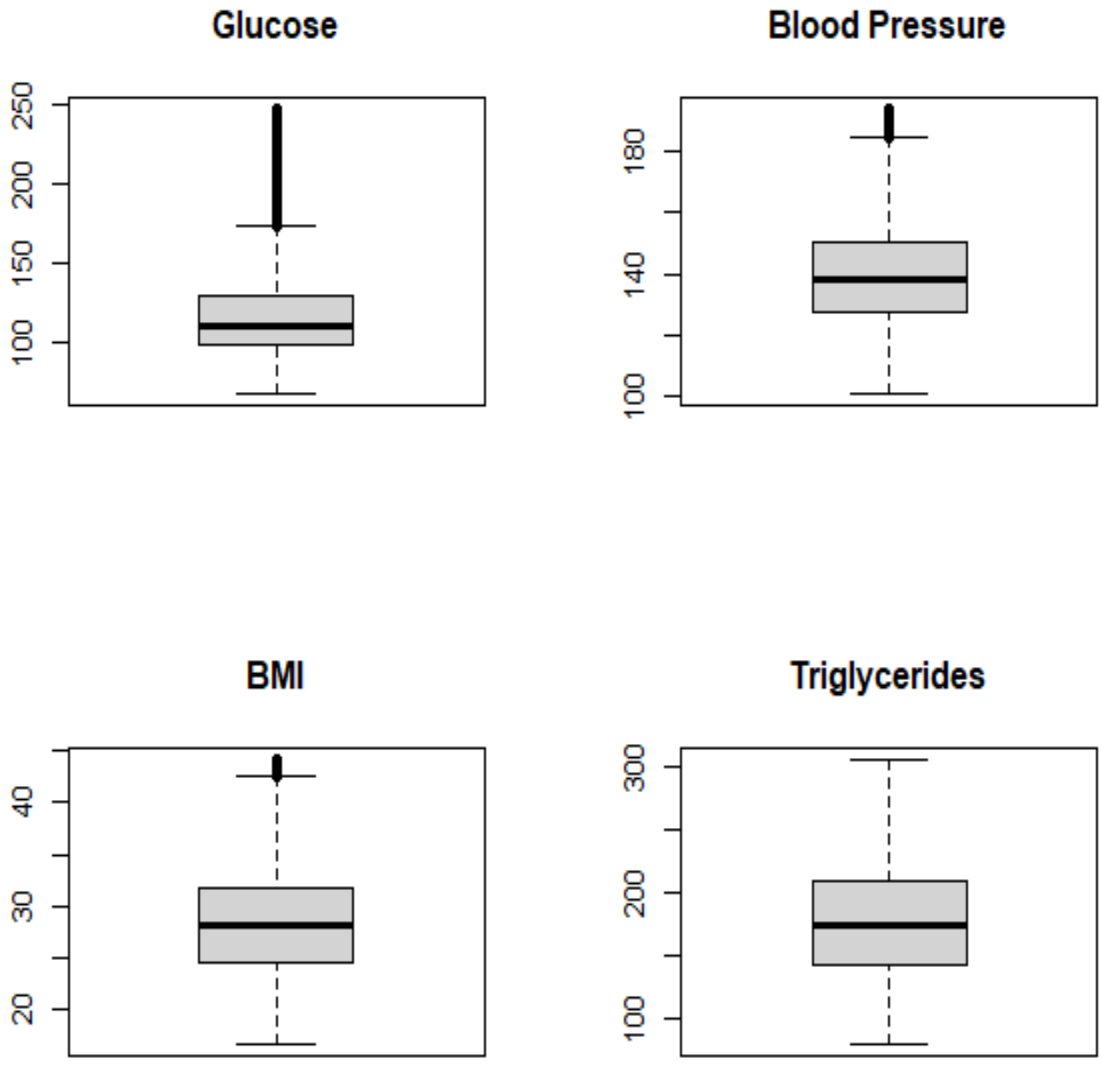

Glucose
100
150
200
250
Blood Pressure
100
140
180
BMI
20
30
40
Triglycerides
100
200
300

Figure 3

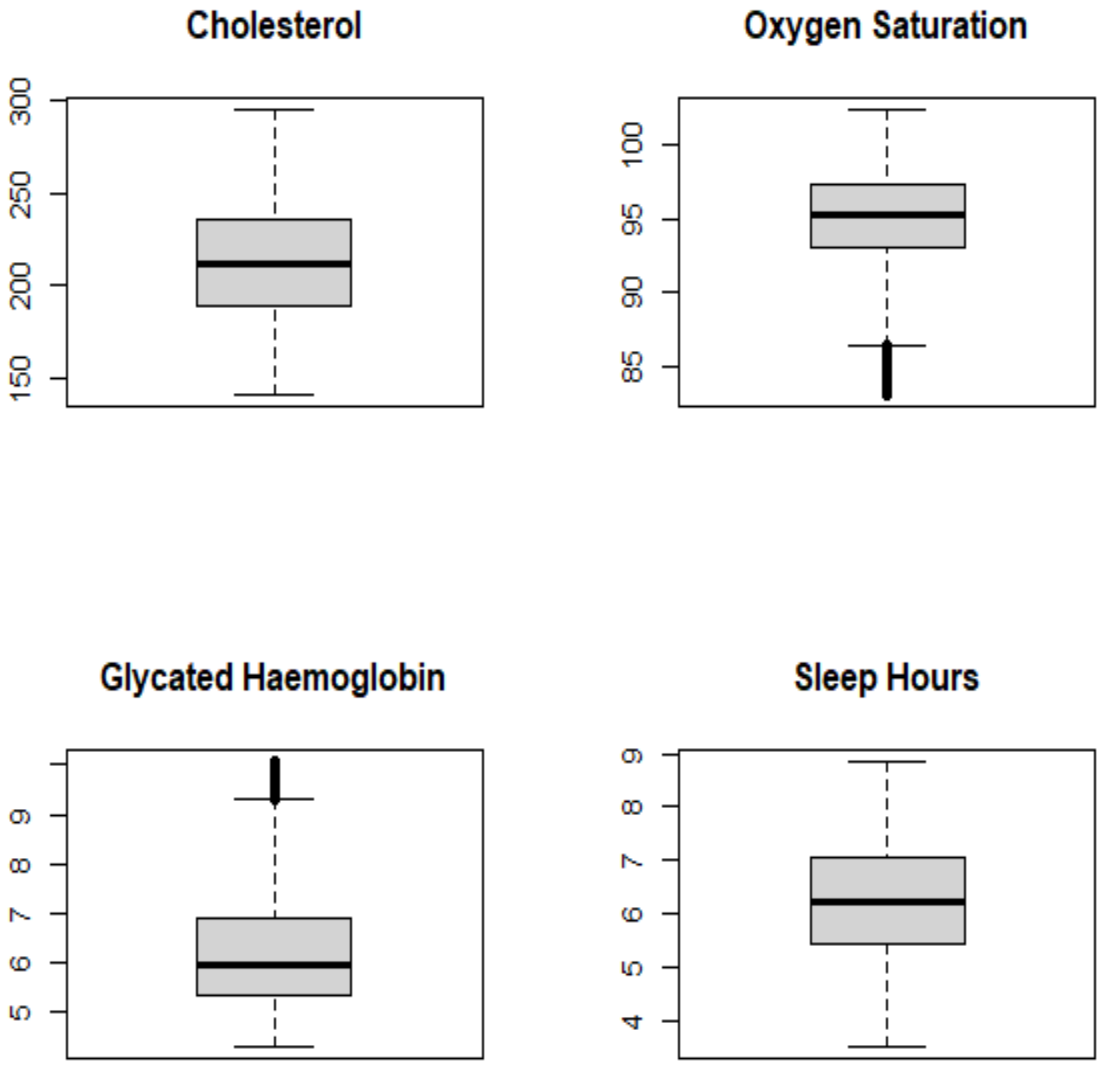


Figure 4

Figure 1& 2 shows the distribution of Glucose, Blood Pressure, BMI, Triglycerides and other numeric variables before outlier treatment. Figure 3 % 4 shows distribution after outlier treatment, highlighting how extreme values were controlled while preserving the underlying distribution. All numeric variables were assessed for outliers; the figures display representative variables for clarity.

4. **Encoding Categorical Variables:**
    - Categorical features (e.g., Gender, Smoking, Alcohol, Family History, and the target variable Medical Condition) were converted into factor formats for proper interpretation by machine learning algorithms.

Exploratory Data Analysis (EDA)

After preprocessing, a thorough exploratory data analysis was conducted to understand the underlying patterns, distributions, and relationships within the dataset. This step helped reveal insights that guided feature selection, model choice, and interpretation of results.

Univariate Analysis

Univariate analysis examined each feature independently to understand its distribution and detect potential irregularities.

1. **Numerical Features:**
   Features such as *Glucose*, *Blood Pressure*, *BMI*, *Cholesterol*, and *Triglycerides* were visualized using histograms and boxplots. These plots highlighted the central tendency, spread, and any remaining extreme values.

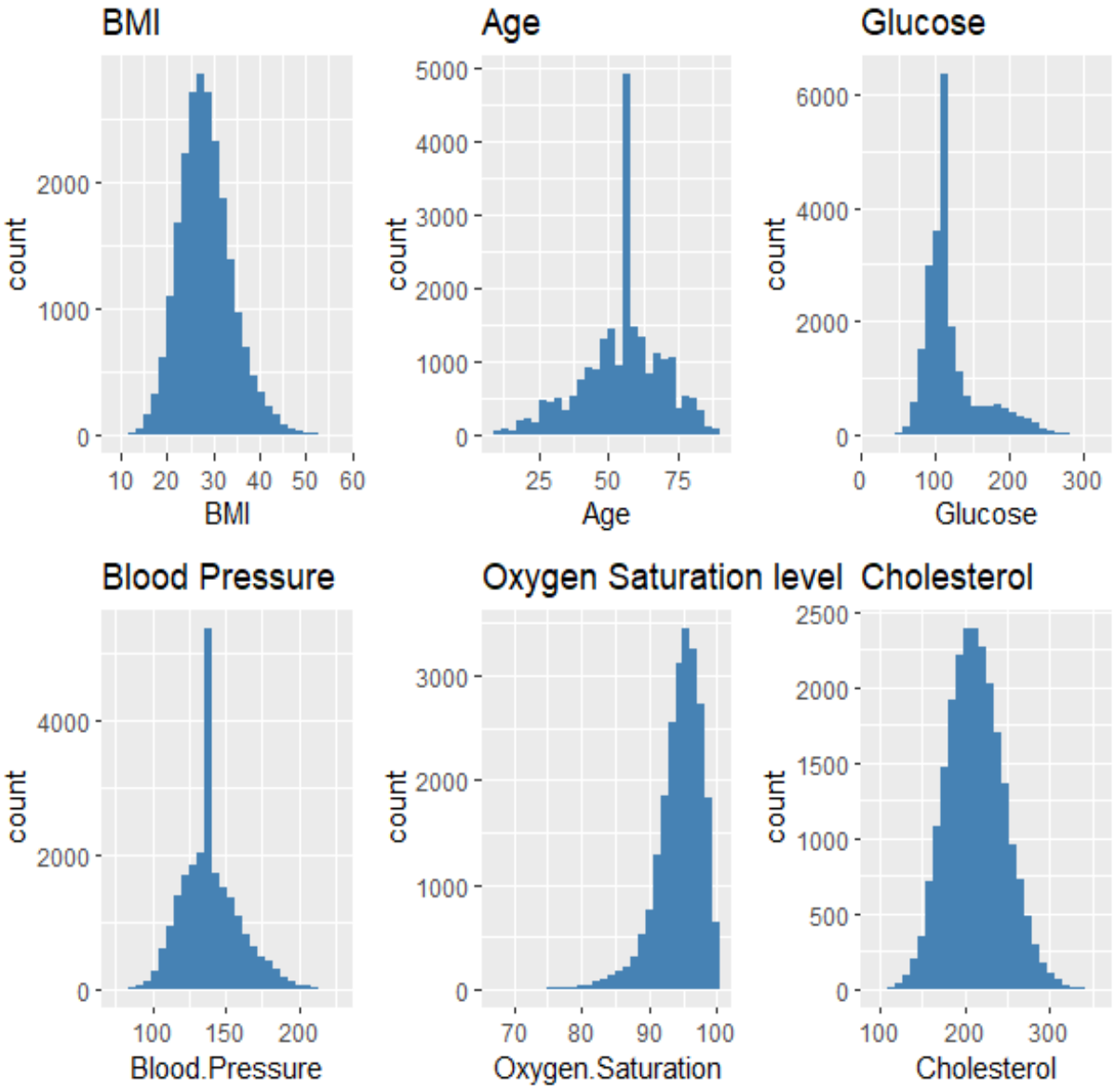

BMI
count
2000
1000
0
10 20 30 40 50 60
BMI
Age
count
5000
4000
3000
2000
1000
0
25 50 75
Age
Glucose
count
6000
4000
2000
0
0 100 200 300
Glucose
Blood Pressure
count
4000
2000
0
100 150 200
Blood.Pressure
Oxygen Saturation level
count
3000
2000
1000
0
70 80 90 100
Oxygen.Saturation
Cholesterol
count
2500
2000
1500
1000
500
0
100 200 300
Cholesterol

Figure 5

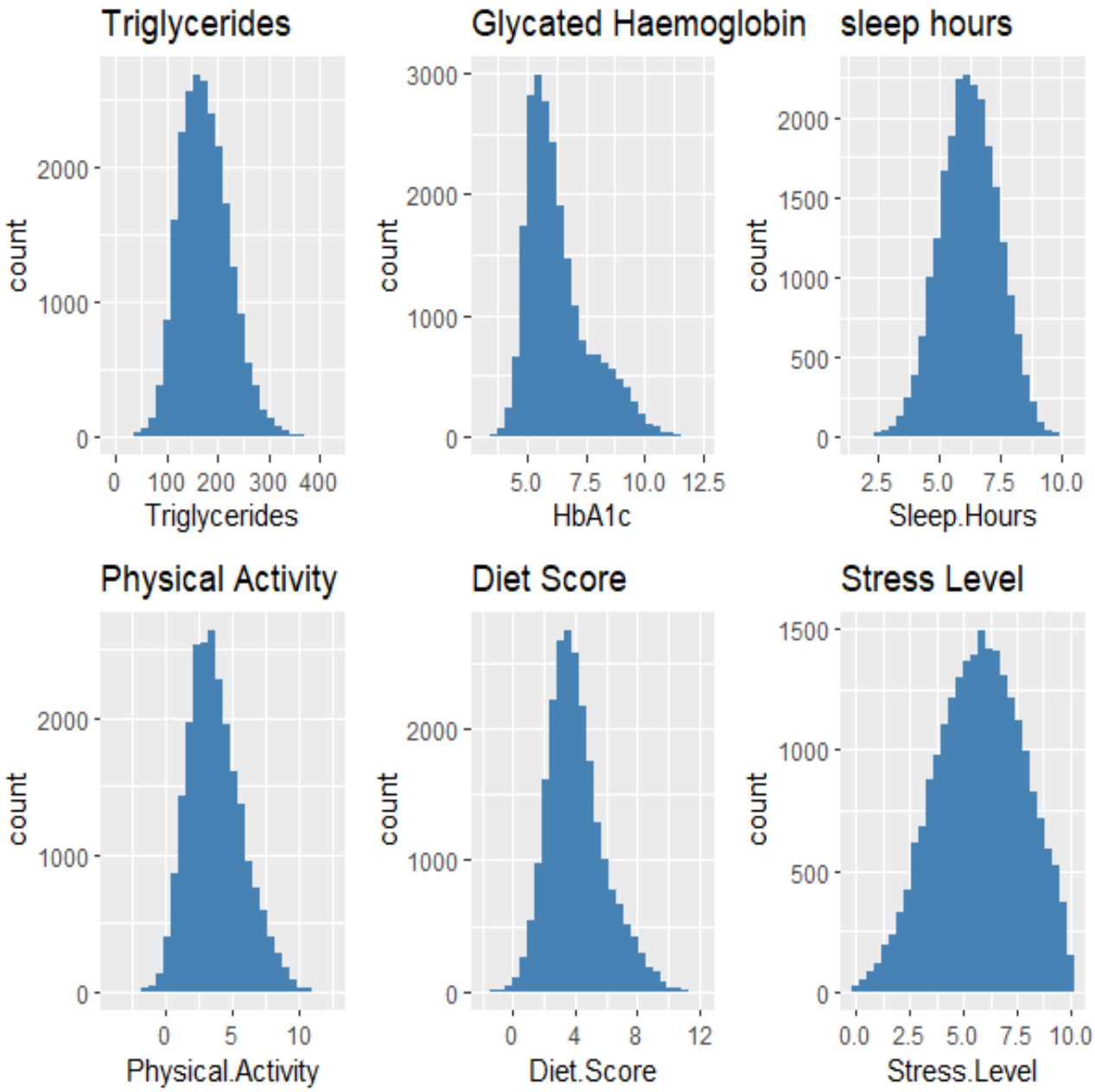


Figure 6

2. **Categorical Features**:
   Variables like *Gender*, *Smoking*, *Alcohol*, and *Family History* were visualized with bar charts. This revealed imbalances in categories. For instance, more males than females, and fewer smokers compared to non-smokers.

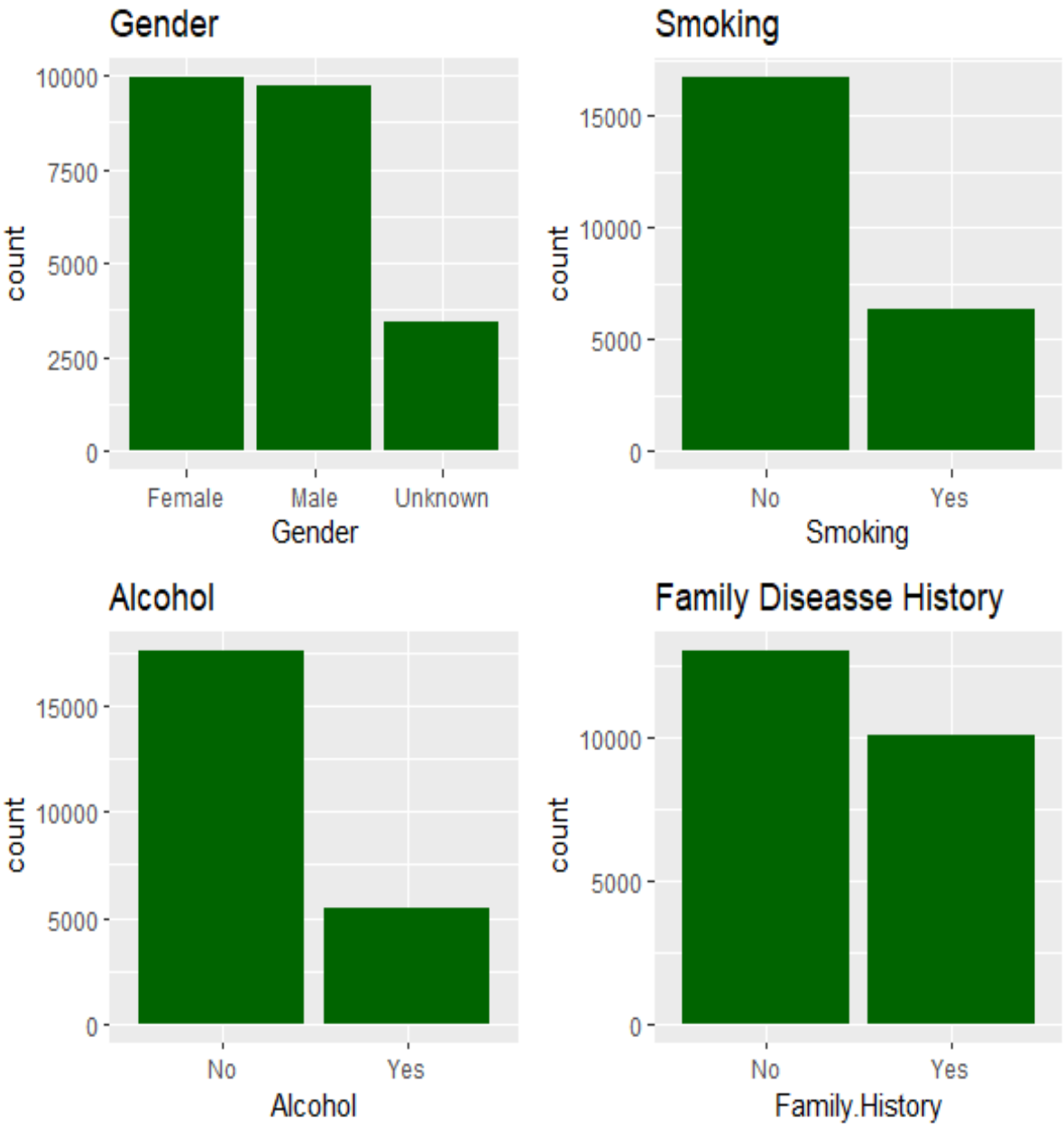


Figure 7

3. **Target Variable:**
   The distribution of *Medical Condition* was examined to confirm class imbalance, with conditions like Diabetes and Hypertension more frequent than Cancer or Asthma.

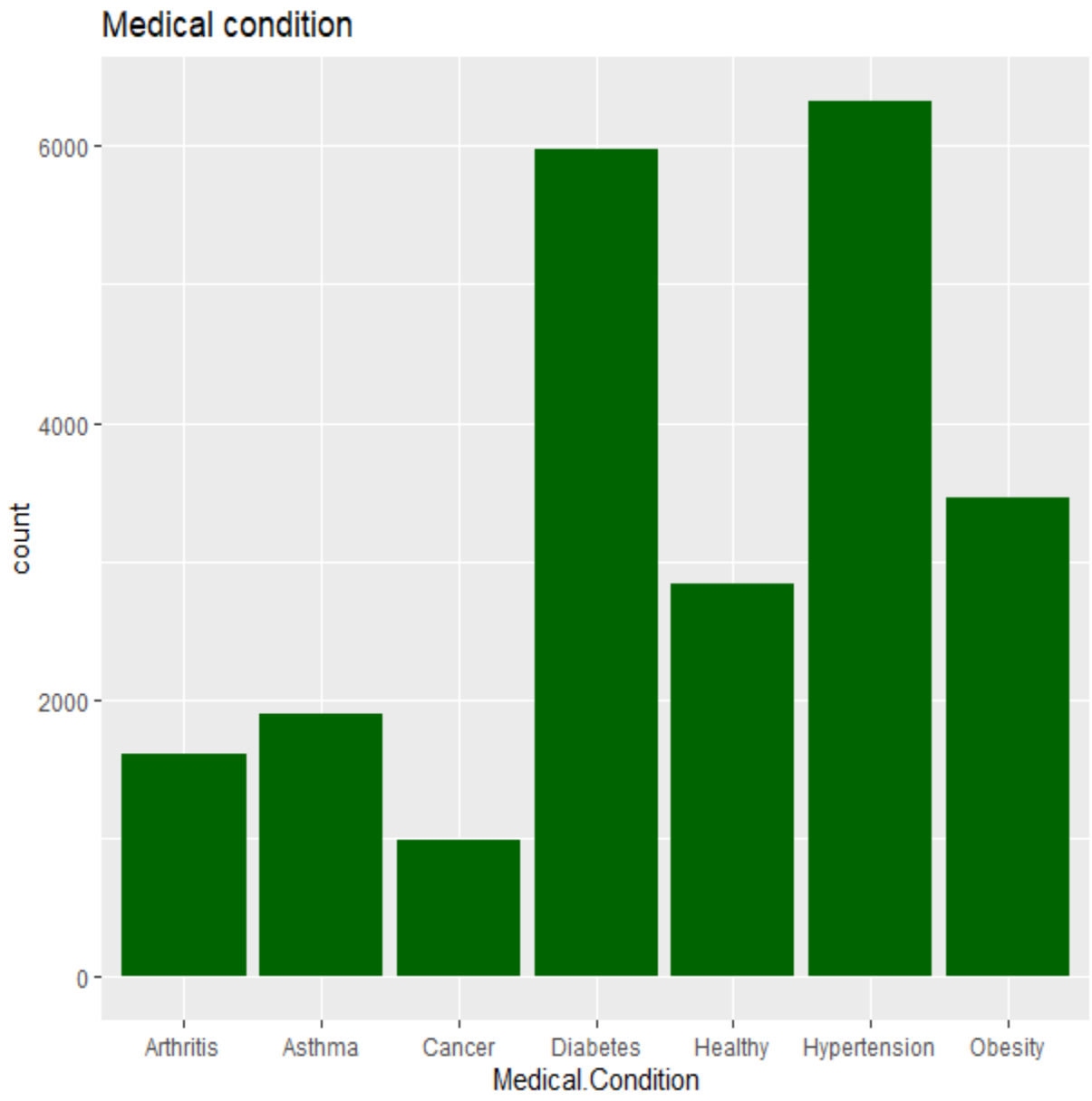


Figure 8

Bivariate Analysis

Bivariate analysis explored relationships between features and their association with the target variable.

1. **Numerical vs Target:**
   Boxplots compared distributions of *Glucose*, *BMI*, and *Blood Pressure* across different medical conditions. This highlighted meaningful differences; for example, higher average glucose levels in patients diagnosed with Diabetes.

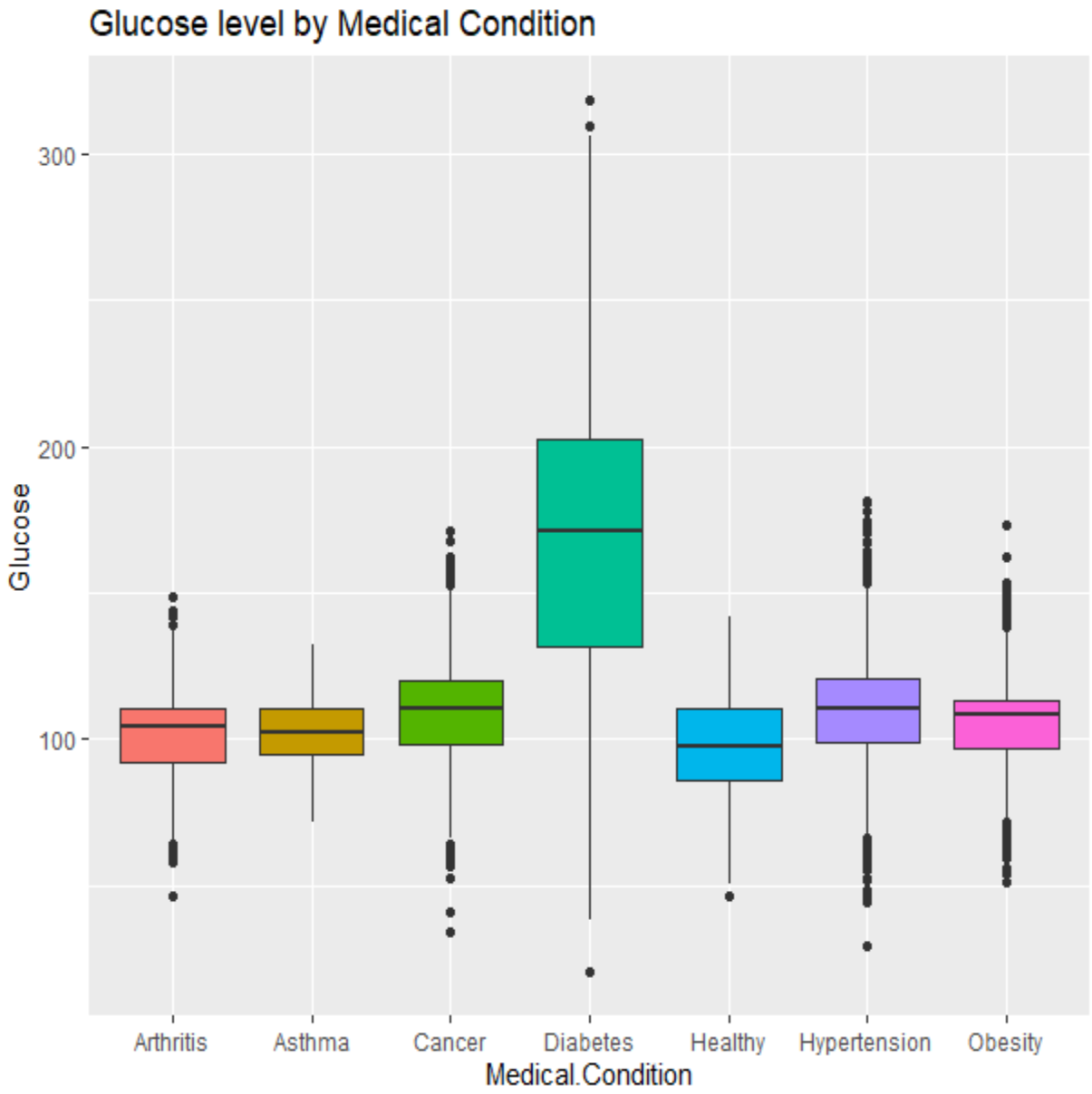

Glucose level by Medical Condition
300
200
100
Glucose
Arthritis
Asthma
Cancer
Diabetes
Healthy
Hypertension
Obesity
Medical.Condition

Figure 9

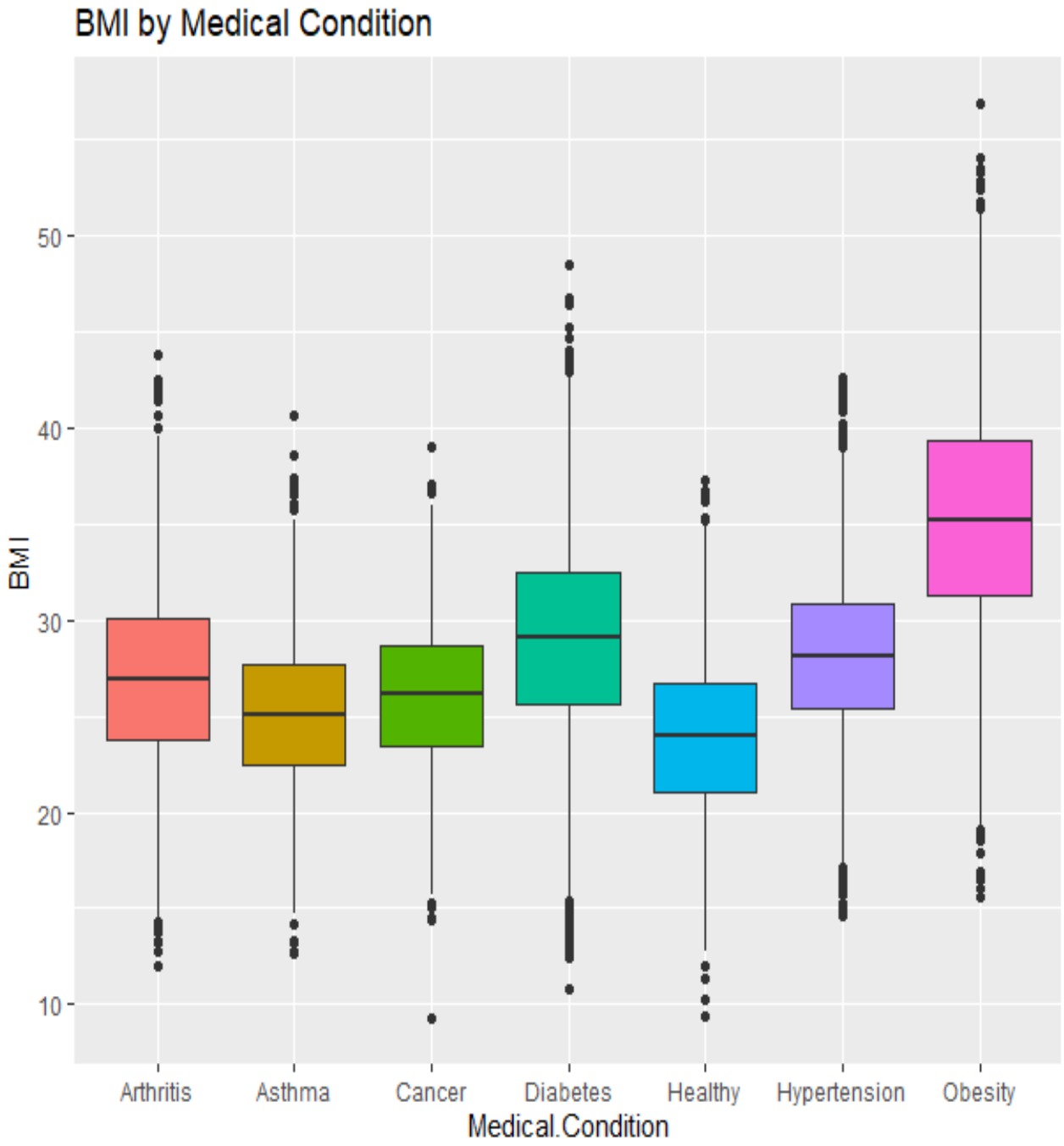

BMI by Medical Condition
BMI
50
40
30
20
10
Arthritis
Asthma
Cancer
Diabetes
Healthy
Hypertension
Obesity
Medical.Condition

Figure 10

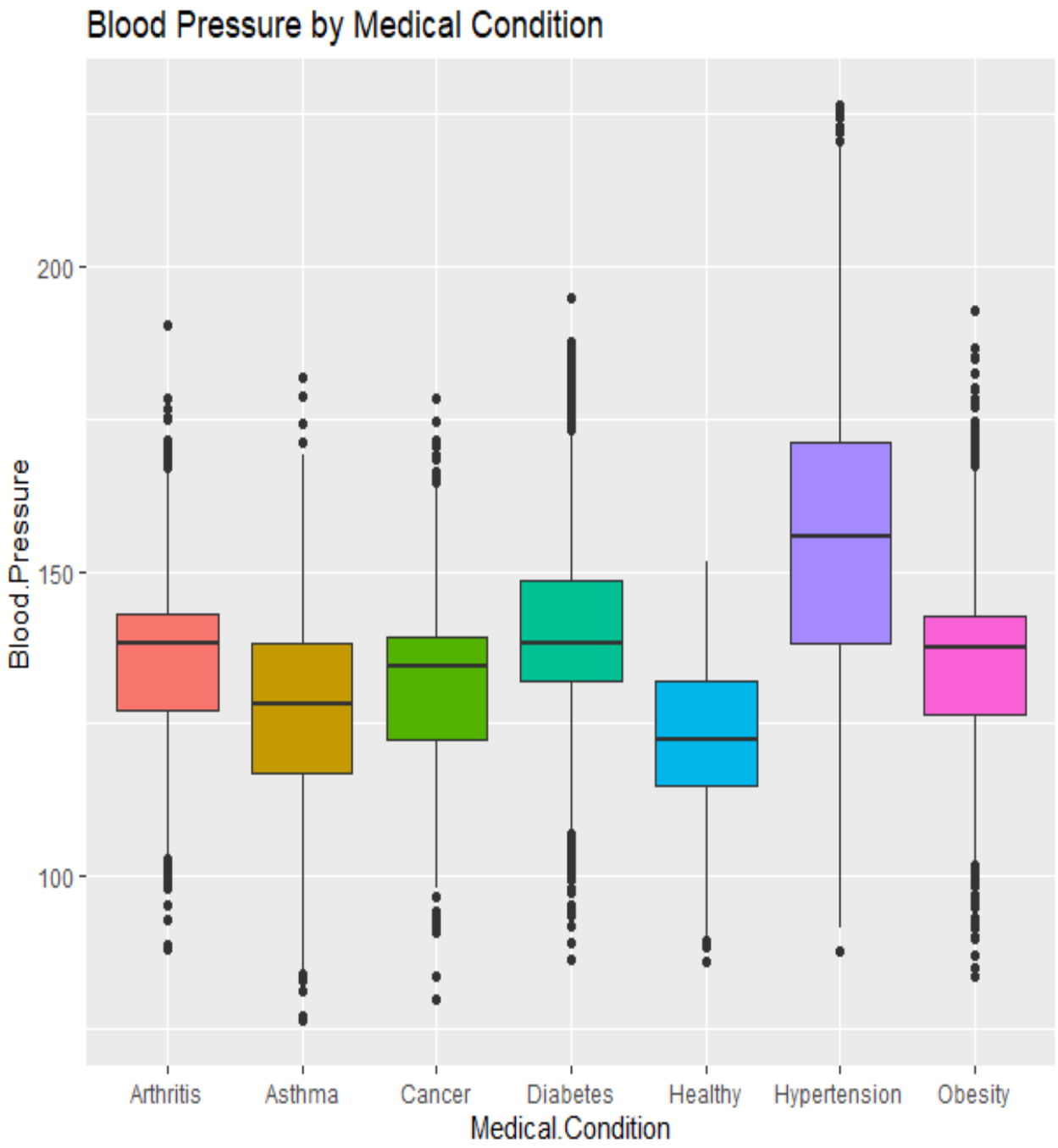


Figure 11

2. **Categorical vs Target:**
   Stacked bar charts examined how lifestyle variables such as *Smoking*, *Alcohol* and family health history related to medical conditions.

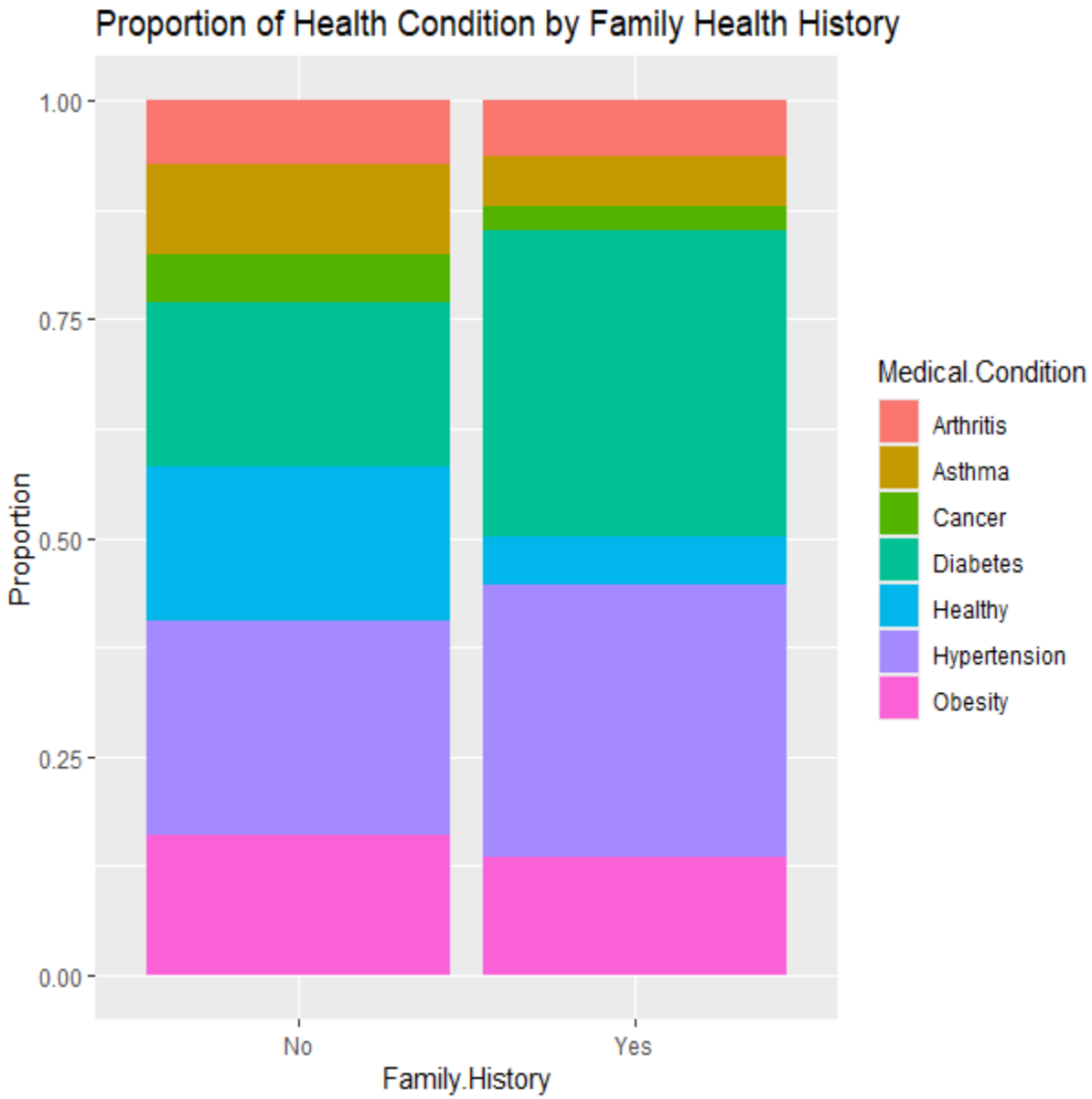

Proportion of Health Condition by Family Health History
1.00
0.75
0.50
0.25
0.00
Proportion
No
Yes
Family.History
Medical.Condition
Arthritis
Asthma
Cancer
Diabetes
Healthy
Hypertension
Obesity

Figure 12

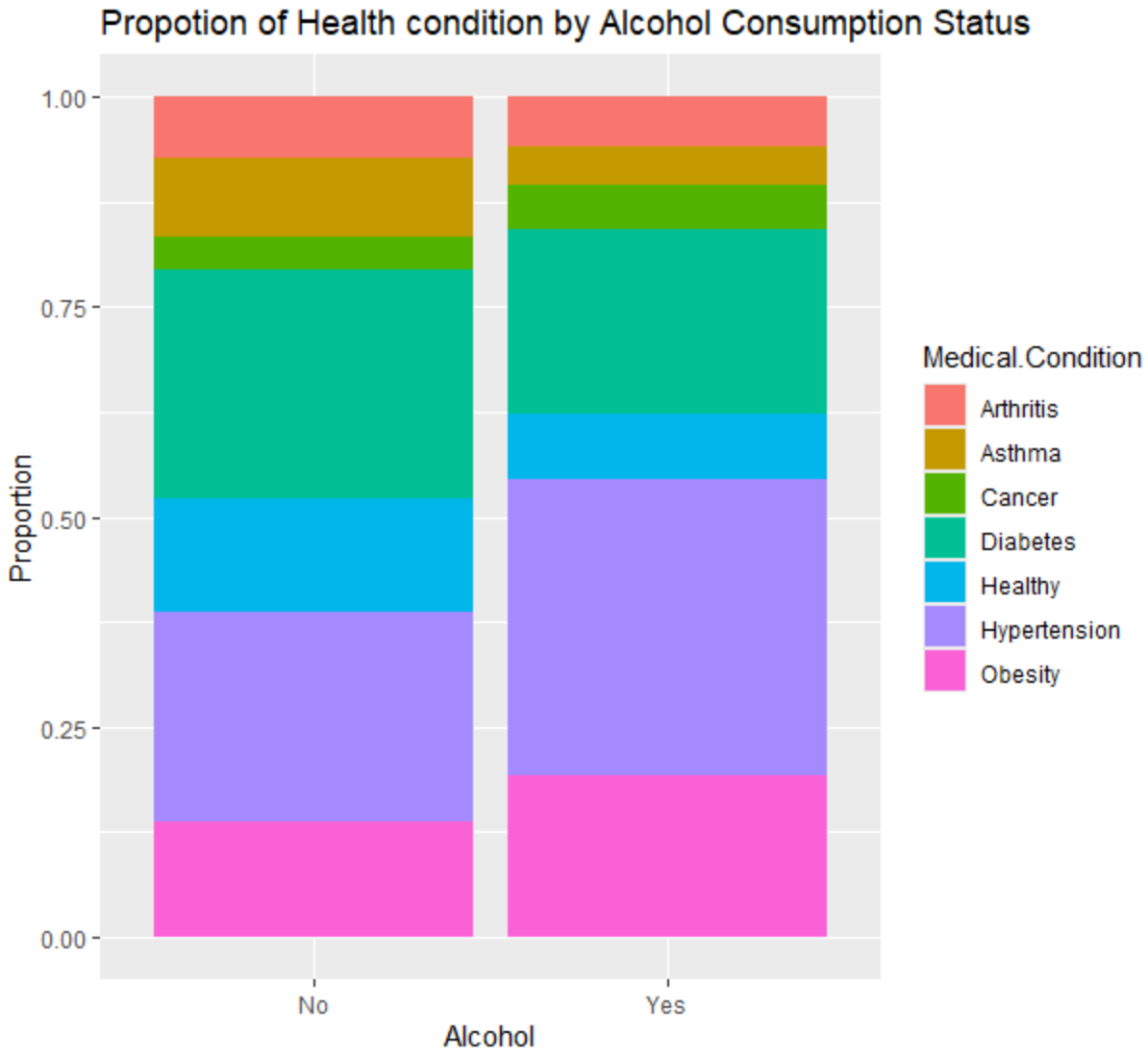

Propotion of Health condition by Alcohol Consumption Status
1.00
0.75
0.50
0.25
0.00
Proportion
No
Yes
Alcohol
Medical.Condition
Arthritis
Asthma
Cancer
Diabetes
Healthy
Hypertension
Obesity

Figure 13

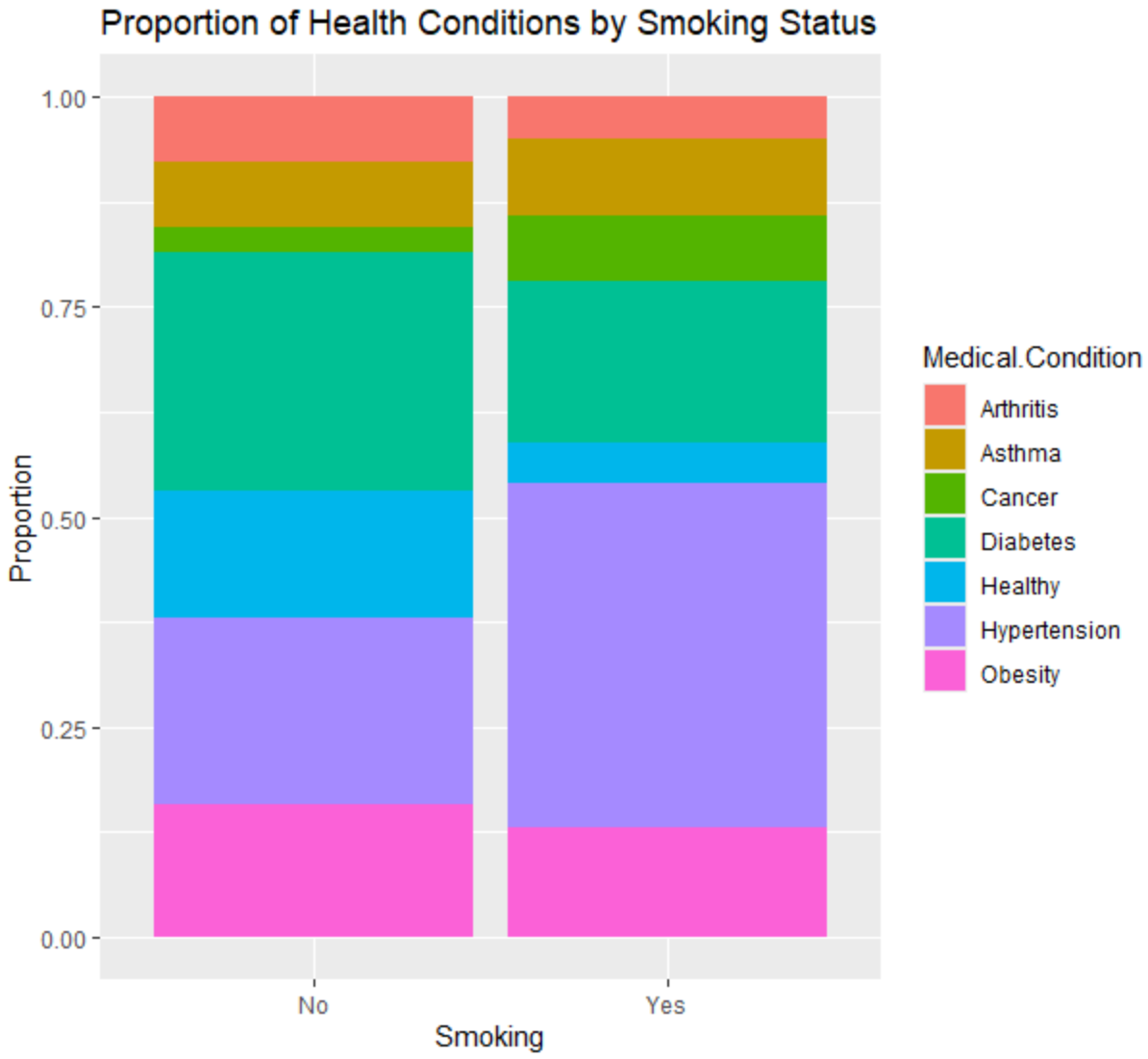


Figure 14

Correlation Analysis:
A correlation heatmap of numerical variables identified strong correlations, such as between *Cholesterol* and *Triglycerides*, and moderate correlations between *BMI* and *Blood Pressure*.

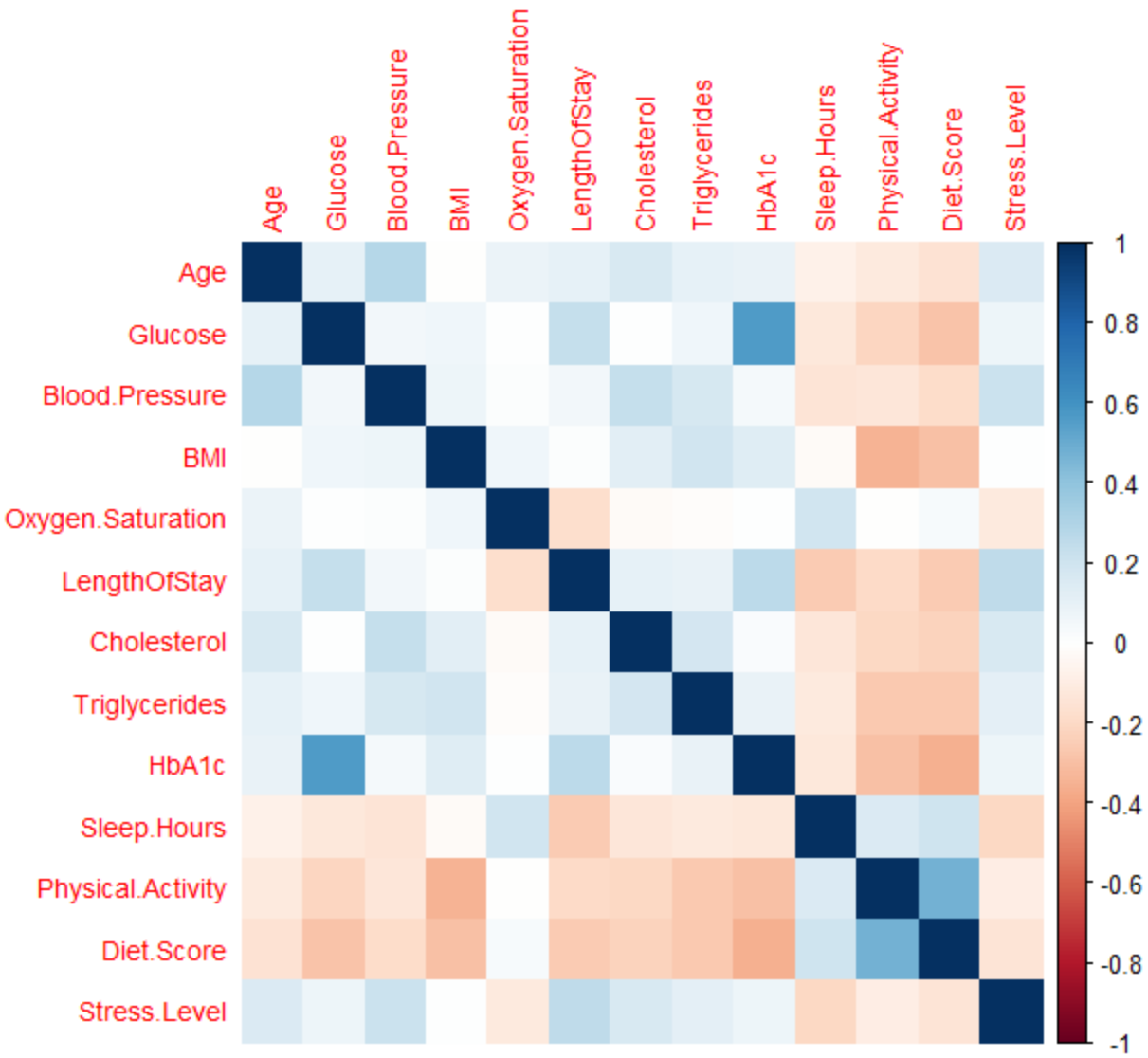


Figure 15

Key Insights from the Exploratory Data Analysis

The exploratory analysis provides several important observations that help shape the modelling strategy and highlight the behaviour of the dataset:

1. Strong Predictive Signals in Core Clinical Variables
   Variables such as Glucose, BMI, Blood Pressure, and HbA1c demonstrate clear variation across health-condition categories. Their distributions indicate that they are likely to contribute significantly to the model's predictive accuracy. These features show measurable differences between individuals with chronic conditions and those classified as healthy, making them central to subsequent modelling decisions.

2. Influence of Lifestyle-Related Risk Factors
   Lifestyle indicators (including smoking status, alcohol consumption, and physical activity

levels) add meaningful context to the dataset. Although they may not be as strongly correlated with the target variable as the clinical measurements, they provide broader behavioral insights that help the model capture non-clinical risk patterns.

3. Presence of Class Imbalance
   The distribution of the target variable reveals an imbalance across health-condition classes. This has direct implications for model training. Without proper handling, the model may become biased toward the majority class. Techniques such as resampling, class-weight adjustments, or synthetic oversampling will be required to ensure robust and fair model performance.

4. Feature Correlation and Redundancy Considerations
   Several features display moderate to high correlations with one another. This suggests the possibility of redundancy, which could introduce multicollinearity and reduce model interpretability. Approaches such as feature selection, regularization, or dimensionality-reduction methods may improve overall model stability and predictive strength.

Feature Engineering

Feature engineering was carried out to improve the predictive capacity of the dataset and ensure that the models received well-structured, informative inputs. The process covered two major areas: feature creation and feature selection.

1. Feature Creation

To enhance the representation of patients' health and lifestyle characteristics, two composite indicators were engineered. These indicators were designed to summarize multiple related variables into single, interpretable metrics.

a. Wellness Index

The *Wellness Index* was created to capture the combined metabolic and physiological risk profile of each individual. It was formed by scaling and averaging the following health-related variables:

- BMI
- Cholesterol
- Triglycerides
- Glucose
- HbA1c

This index provides a compact measure of cardiometabolic wellness and helps the model better recognize patterns associated with chronic health conditions.

b. Lifestyle Score

The *Lifestyle Score* was developed to quantify behavioral and daily habit factors known to influence long-term health outcomes. It was constructed from the scaled versions of:

- Physical activity
- Diet score
- Sleep hours
- Stress level

This new variable captures overall lifestyle quality and allows the model to evaluate how behavioral factors contribute to different medical conditions.

2. Feature Selection

Feature selection was performed to identify the variables that carry the strongest predictive signal. Two independent approaches were used to ensure robustness: the Boruta algorithm and Random Forest importance analysis.

a. Boruta Algorithm

The Boruta feature selection method (an all-relevant feature selection wrapper) was applied to confirm which variables are essential for predicting the target outcome.

- Confirmed Features: 18 attributes
- Rejected Feature: 1 attribute (Gender)
- Iterations: 11
- Runtime: ~19 minutes

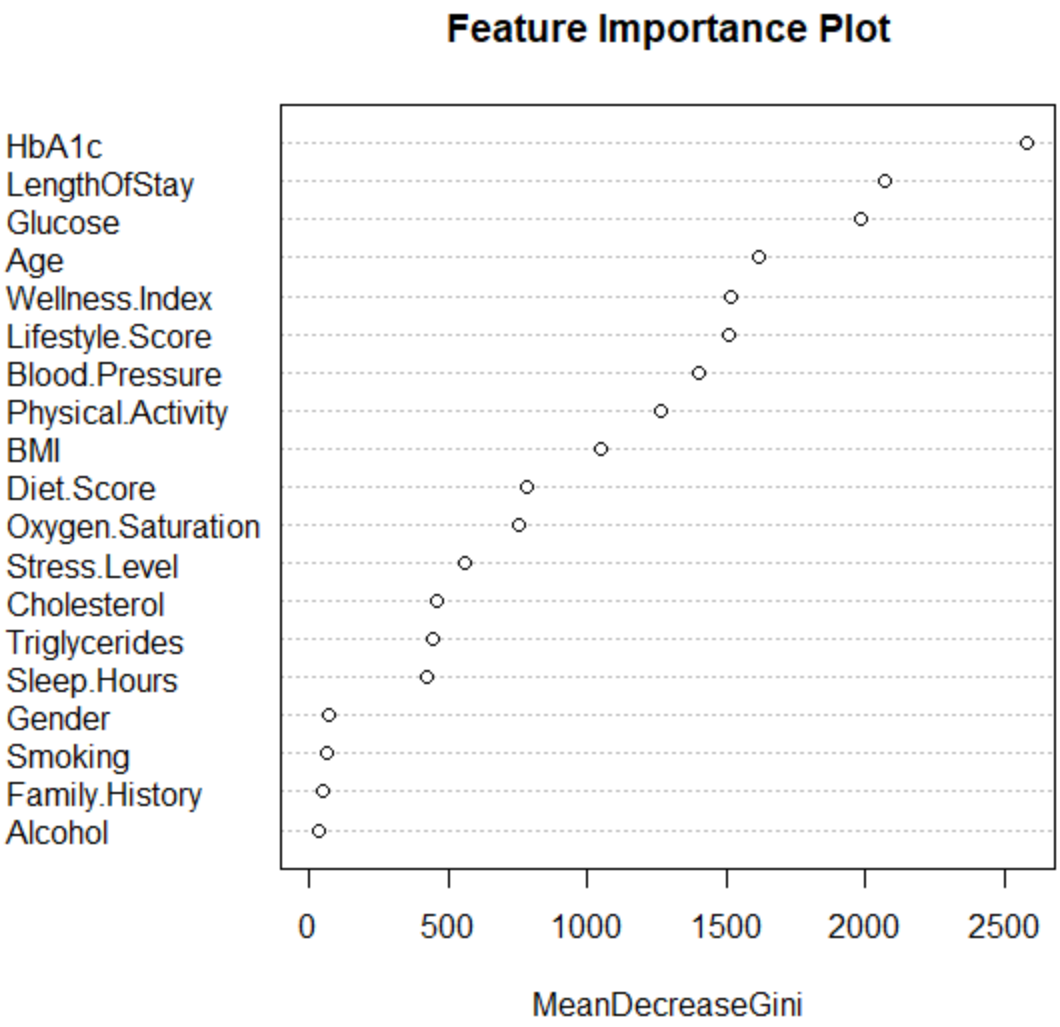
Boruta consistently identified nearly all variables as important predictors, except Gender, suggesting that gender differences did not significantly contribute to the classification of medical conditions in this dataset.

b. Random Forest Variable Importance

A Random Forest model was trained to compute the Mean Decrease in Gini Index, providing another reliable measure of feature relevance.

Key observations from the RF importance results include:

- HbA1c, Glucose, and Length of Stay were among the strongest predictors.
- Variables such as Alcohol, Smoking, and Gender showed very low importance scores.
- The engineered features (Wellness Index and Lifestyle Score) ranked highly, confirming their usefulness.



A variable importance plot (attached above) visually highlights these contributions.

3. Final Selected Features

Based on the combined evidence from Boruta and Random Forest importance, the following variables were retained for model training:

Age, Glucose, Blood Pressure, BMI, Oxygen Saturation, Cholesterol, Triglycerides, HbA1c, Physical Activity, Diet Score, Stress Level, Sleep Hours, Wellness Index, Lifestyle Score.

Modelling

1. Train–Test Split

The dataset was divided into an 80:20 train–test split to ensure that model training and evaluation were performed on independent samples.

All preprocessing and resampling strategies were applied exclusively to the training set to prevent data leakage and preserve the integrity of the evaluation process.

2. Baseline Modelling on the Imbalanced Dataset

The first stage of the modelling process involved training three baseline classifiers (Decision Tree, Random Forest and XGBoost) on the original imbalanced dataset.
This provided an initial benchmark and revealed how each model performed under the natural class distribution.

Performance on Imbalanced Data

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.7503	0.6231	0.6312	0.6247
Random Forest	0.8725	0.8337	0.7817	0.8063
XGBoost	0.8788	0.8351	0.8059	0.8202

The Decision Tree model struggled with the imbalanced data, showing lower precision and recall.
Random Forest and XGBoost delivered notably stronger results, with XGBoost achieving the most balanced performance.
The high accuracy values, however, indicated potential bias toward majority classes, reinforcing the need for resampling.

3. Modelling on the Balanced Dataset (Random Oversampling)

To address class imbalance and improve minority-class representation, random oversampling was applied to the training set.
The three models were retrained using the balanced dataset to assess how improved class distribution influenced performance.

Performance After Oversampling

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.675	0.6372	0.6662	0.6514
Random Forest	0.8708	0.8188	0.8059	0.8126
XGBoost	0.8734	0.795	0.807	0.814

As expected, oversampling reduced accuracy for the Decision Tree model due to the loss of majority-class bias.
Despite this, recall improved, indicating better recognition of minority classes.

Random Forest and XGBoost remained highly stable, with XGBoost again demonstrating the strongest balance between precision, recall and F1-Score.

4. Hyperparameter Tuning (GridSearch CV)

Given their strong baseline performance, Random Forest and XGBoost were selected for hyperparameter tuning using GridSearch with 5-fold cross-validation.
This step aimed to refine model complexity, reduce overfitting and improve the precision–recall balance.

Tuned Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Tuned Random Forest	0.8715	0.802	0.798	0.799
Tuned XGBoost	0.8741	0.785	0.806	0.806

Tuning led to modest but meaningful improvements in predictive balance.
The tuned XGBoost model preserved strong recall while maintaining competitive precision, resulting in the highest F1-Score among all tuned models.
The tuned Random Forest showed stable performance but was slightly less sensitive to minority-class patterns.

Model Selection and Discussion

XGBoost emerged as the best-performing model across all evaluation settings. It consistently delivered strong accuracy, balanced precision–recall performance, and stable results on both imbalanced and oversampled datasets. Its ability to handle complex feature interactions and maintain high recall for minority classes makes it especially suitable for medical predictions.

Random Forest also performed well but showed slightly weaker balance across classes. The Decision Tree model recorded the lowest performance due to its sensitivity to class imbalance and limited modelling capacity. Oversampling improved recall for all models, confirming its value for health-related classification tasks. Hyperparameter tuning introduced only marginal improvements, indicating that the models (particularly XGBoost) were already well aligned with the data structure.

Overall, XGBoost is selected as the final model for its strong generalization, superior class balance, and consistent performance across all conditions.

Conclusion

This project set out to build a reliable model for predicting medical conditions using a combination of lifestyle indicators, clinical measurements, and patient history. After evaluating several algorithms under different sampling conditions, XGBoost emerged as the most dependable model. It delivered the best balance of accuracy and sensitivity while remaining

consistent across imbalanced and balanced datasets. Overall, the results show that thoughtfully engineered features such as the Wellness Index and Lifestyle Score, combined with advanced machine learning methods, can meaningfully support early detection efforts in healthcare.

Recommendations

1. Retrain the model periodically with new patient records to maintain accuracy and capture evolving health trends.
2. Expand the dataset to include additional clinical variables that may improve predictive strength.
3. Test the model in real clinical environments to assess how well it performs on diverse populations.
4. Develop a monitoring system that tracks the model's performance over time and highlights any decline in accuracy.