A human body with a heart

Description automatically generated

**Comparative Evaluation of Machine Learning Algorithm for Heart Attack Risk Prediction Using R Programming System**

# **ABSTRACT**

Cardiovascular diseases (CVD) pose a global health concern that requires advanced tools for precise diagnosis and prediction. Diagnosing and predicting cardiovascular disease are vital medical duties that are necessary to guarantee accurate classification and allow cardiologists to administer appropriate care. Because machine learning algorithms can identify patterns in data, they are becoming increasingly popular in the medical industry. This opens new possibilities for enhancing the classification of cardiovascular diseases and decreasing the number of incorrect diagnoses.

The objective of this research is to develop a forecasting model that precisely determines the likelihood of a heart attack occurring. In line with the general healthcare trend, the study incorporates lifestyle and health factors into the predictive model, where sophisticated analytics and machine learning play a critical role in enhancing predictive skills. This integration aims to offer insightful information into personalized risk assessments to promote early detection and preventive measures for individuals with a higher likelihood of heart attack.

The study primarily aims to assess the efficacy of the K-Nearest Neighbours (KNN) and Support Vector Machine (SVM) algorithms. Utilising a diverse Kaggle dataset comprising observations from multiple continents with a particular focus on Europe, the research endeavours to offer region-specific perspectives on cardiovascular well-being. Precise categorization is crucial when it comes to creating efficient treatment and preventive plans for cardiovascular disease.

The research provides intriguing findings. Notably, the KNN model outperforms the SVM model with an accuracy of 89.08%, which is significantly higher than the latter's accuracy of 60%. A significant Kappa value of 77.7% for KNN, which shows its efficacy in identifying patterns beyond chance, further supports this robust performance. These results highlight the KNN model and machine learning's promise as a useful tool for precise risk assessment and prediction in cardiovascular illnesses.

# **OUTLINE SPECIFICATION OF THE PROBLEM**

Cardiovascular disease (CVDs) is the major cause of death and disability globally, leading to over 70% of all fatalities. The development of machine learning algorithms that can accurately predict an individual's risk of suffering a heart attack is the aim of this research. Although heart attack problems are common worldwide, the European population is the focus of this study.

High-income countries like the countries in Europe have high rates of heart disease risk factors, such as smoking, eating a poor diet, high sugar consumption, and being overweight or obese. The main issue to be addressed in this research is the creation of a precise machine-learning model for the prediction of an individual’s risk of having a heart attack based on their health and lifestyle characteristics that are common in European nations.

Heart attack risk prediction in the European population is important because of this continent's distinct socioeconomic, cultural, and lifestyle characteristics. The research helps to enhance tailored healthcare methods by attending to the unique demands of this population. This allows for timely interventions and lowers the burden of heart-related ailments in Europe.

# **TABLE OF CONTENT**

**Contents**

[**ABSTRACT** 1](#_Toc155255612)

[**OUTLINE SPECIFICATION OF THE PROBLEM** 2](#_Toc155255613)

[**TABLE OF CONTENT** 3](#_Toc155255614)

[**INTRODUCTION** 4](#_Toc155255615)

[**Research Objective** 5](#_Toc155255616)

[**Research Questions** 6](#_Toc155255617)

[**LITERATURE REVIEW** 6](#_Toc155255618)

[**METHODOLOGY** 8](#_Toc155255619)

[**Description of the data set** 8](#_Toc155255620)

[**Data Pre-processing** 9](#_Toc155255621)

[Plotting to check for missing values. 9](#_Toc155255622)

[Continent Selection 9](#_Toc155255623)

[Data Visualisation 10](#_Toc155255624)

[Data Cleaning 14](#_Toc155255625)

[Data Distribution 15](#_Toc155255626)

[Feature Selection 16](#_Toc155255627)

[Feature Scaling/ Normalisation 17](#_Toc155255628)

[Smote Balancing 17](#_Toc155255629)

[Data Preparation 18](#_Toc155255630)

[Machine Learning Classifications 18](#_Toc155255631)

[**Machine Learning Application** 19](#_Toc155255632)

[**Performance Evaluation** 20](#_Toc155255633)

[**RESULT AND CONCLUSIONS** 24](#_Toc155255634)

[**Accuracy** 24](#_Toc155255635)

[**ROC-AUC Curve** 25](#_Toc155255636)

[**RECOMMENDATION** 27](#_Toc155255637)

[**REFERENCES** 28](#_Toc155255638)

# **INTRODUCTION**

Cardiovascular diseases (CVDs) are a broad term encompassing disorders affecting the heart and blood vessels. It covers everything from hereditary or inherited disorders to illnesses that occur later in life, like vascular dementia, heart failure, stroke, and coronary heart disease.

The World Health Organisation (WHO) published a report on cardiovascular diseases (CVDs) in 2023. With 17.9 million deaths annually, CVDs are the leading cause of mortality worldwide. Heart and blood vessel disorders commonly referred to as cardiovascular diseases, or CVDs, include rheumatic heart disease, coronary heart disease, and cerebrovascular illness. Heart attacks and strokes account for more than four out of every five fatalities from CVD, and one-third of these events happen too soon among those under the age of seventy.

The main lifestyle risk factors for heart disease and stroke include excessive alcohol consumption, poor eating habits, smoking, and inactivity. Individuals with this lifestyle risk factors may experience elevated blood pressure, blood sugar, blood cholesterol, and obesity. In primary care settings, these "intermediate risk factors" can be assessed to detect an elevated risk of heart attack, stroke, heart failure, and other consequences.

Cardiovascular diseases continue to be the primary cause of death in the European Union and the world's top cause of death despite significant advancements in medicine over the past few decades. In the EU, cardiovascular disease results in about 6 million new cases and the death of about 1.8 million individuals annually. According to estimates from the European Heart Network, the annual cost of all of this to the EU economy is more than €210 billion.

One of the deadliest cardiovascular diseases (CVDs) is acute myocardial infarction (AMI), sometimes referred to as a "heart attack." It happens when the blood supply to the heart muscle is cut off, either permanently damaging or destroying the heart muscle. It results from inadequate blood supply to the cardiac muscle. Although there are numerous other possible causes for the reduced blood flow, the most common one is usually an obstruction in one or more of the heart's arteries. The damaged heart muscle will start to perish without blood supply. A heart attack may be fatal or cause irreversible cardiac damage if blood flow is not quickly restored.

To replicate human learning, machine learning (ML), a subfield of artificial intelligence (AI) and computer science, uses data and algorithms to gradually increase model accuracy. It is an essential part of the quickly evolving data science profession. Algorithms in data mining projects are trained using statistical techniques to find relevant information and create predictions or classifications. Subsequently, these insights inform business and application decisions that ideally affect key performance measures.

The application of machine learning (ML) in the healthcare industry is dependent on the gathering of relevant patient data. By identifying patterns in datasets, machine learning algorithms can assist medical personnel in diagnosing new illnesses and predicting treatment outcomes. Data is filtered and categorized by these systems and technologies.

Even within a state or nation, the amount of data gathered from patients in medical facilities is enormous. Ensuring that all medical records and gadgets are part of a central network that enables data scientists to identify patterns and trends is the only way to bring everything into sync.

Algorithms are used in machine learning for heart attack prediction to examine a variety of data sources, such as genetics, lifestyle, and medical history. Based on unique risk profiles, these models allow for risk categorization, early identification, and tailored interventions. ML can continually monitor real-time patient data for dynamic risk assessment and can identify pertinent variables influencing the likelihood of heart attacks. The precision and generalizability of the model are guaranteed by the incorporation of historical datasets and validation procedures. The goal of ongoing research is to increase cardiovascular risk management through the ethical application and effectiveness of machine learning models in healthcare environments.

## **Research Objective**

The goal of this research is to use machine learning techniques to create a forecasting model that considers lifestyle and health indicators to figure out a person's risk of having a heart attack.

## **Research Questions**

The questions that will be focused on in this research are:

1. How do health and lifestyle factors affect a person's likelihood of having a heart attack?
2. What machine learning methods can be used to create a forecasting model that accurately assesses the risk of a heart attack?

# **LITERATURE REVIEW**

The fields of data mining and machine learning have made noteworthy progress in the healthcare sector in recent years. Numerous articles have been written about the application of data mining methods to survival research. These methods have been extensively used and shown in several healthcare settings, most notably in the speciality of medical cardiology. Researchers in this sector now have an unparalleled chance to create and evaluate novel algorithms due to the quick collection of medical data. This literature review covers related works on machine learning-based heart attack prediction, offering an in-depth overview of the main ideas, approaches, and conclusions in the area. Some of the related works are explained below:

In a study conducted by Bhatt et al (2023), on Effective Heart Disease Prediction Using Machine Learning Techniques. This study created a model that accurately predicts cardiovascular illnesses to lower the death rate associated with these conditions. To increase classification accuracy, this research suggested a k-mode clustering technique with Huang beginning. It made use of models like XGBoost (XGB), random forest (RF), decision tree classifier (DT), and multilayer perceptron (MP). The applied model's parameters were hyper-tuned using Grid Search CV to maximize the outcome. A real-world dataset of 70,000 cases from Kaggle is used to test the suggested model. The research led to the conclusion that the multilayer perceptron with cross-validation has achieved the highest accuracy compared to all other algorithms. With an accuracy of 87.28%, it was the most accurate.

In a study conducted by Riyaz et al (2022) on heart disease prediction using machine learning techniques: a quantitative review. To effectively anticipate, diagnose, and treat a variety of heart ailments, this study provided a comprehensive overview of several machine learning algorithms and evaluated their efficacy. Support vector machines (SVM), decision trees (DT), Naïve Bayes (NB), K-nearest neighbour (KNN), artificial neural networks (ANN), etc. are a few machine learning techniques that are used to forecast the development of cardiac illnesses and are reviewed in the suggested study. Next, to determine which technique performed the best overall, the average forecast accuracy for each technique was calculated. The results showed that the ANN technique had the highest average prediction accuracy (86.91%), while the C4.5 decision tree technique had the lowest average prediction accuracy (74.0%).

In a study published by Shah et al (2020), the author aimed to investigate creating a machine learning-based model for cardiovascular disease prediction. The study took advantage of the pre-existing Cleveland database of the UCI heart disease patient repository. The study report lists several heart disease-related characteristics and a model based on supervised learning techniques, including algorithms like random forest, decision trees, K-nearest neighbour, and Naïve Bayes. According to the results, the K-nearest neighbour yields the highest accuracy score.

A study was conducted towards comparing and using machine learning techniques for detecting and predicting Heart Attacks and Diseases by Obasi et al (2019). The suggested solution was based on currently used methods such as logistic regression and random forest Bayesian classification, which gave medical practitioners a decision support system to identify and forecast heart attacks and heart diseases in people or in people who used heart disease risk factors. The comparison findings demonstrated that the accuracy and performance of the system are satisfactory, with 92.44% accuracy for Random Forest, 61.96% accuracy for Naïve Bayes Classifier, and 59.7% accuracy for Logistic Regression in the classification of heart disease.

In the research conducted by S Manikandan (2017), on Heart Attack Prediction System. With the aid of a binary classifier, the research study automates risk prediction in an effort to save the doctor's time and effort. This paper presented a prototype implementation of such a system with a user-friendly interface. The classifier was constructed using the Naïve Bayes technique, and the graphical user interface is online based. The accuracy score provided by the final system was 81.25%.

Two categorization algorithms will be used in this study to determine an individual's risk of having a heart attack based on lifestyle and health factors.

# 

# **METHODOLOGY**

## **Description of the data set**

This dataset offers a wide range of elements that are pertinent to heart health and lifestyle decisions. It includes lifestyle factors, medical aspects, socioeconomic aspects, and patient-specific information. The collection comprises 26 columns of different features and 8763 records from patients worldwide all kept in CSV format. This dataset consists of records of patients from various continents of the world, but this analysis would only focus on records within Europe.

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Features | Data Type | Description |
| 1 | Patient ID | Categorical | Unique ID for each patient |
| 2 | Age | Numerical (Continuous) | Age of Patient |
| 3 | Sex | Categorical (Nominal) | Patient’s Sex [Male or Female] |
| 4 | Stress Level | Numerical (Continuous) | Patient’s Stress Level [1/10] |
| 5 | BMI | Numerical (Continuous) | Body Mass Index (BMI) of the patient (derived from patient’s weight and height) |
| 6 | Family History of Heart Attack | Categorical (Binary) | Patient’s family history of heart attack [Yes;1 or no; 0] |
| 7 | Obesity | Categorical (Binary) | Patient’s Obesity Status [Obese;1 or Not Obese;0] |
| 8 | Country | Categorical (Nominal) | The Country the patient lives in |
| 9 | Continent | Categorical (Nominal) | The Continent the patient lives in |
| 10 | Hemisphere | Categorical (Nominal) | The Hemisphere the patient lives in |
| 11 | Income | Numerical (Continuous) | Patients Annual Income |
| 12 | Smoking | Categorical (Binary) | Patient’s smoking status [Yes;1 or no;0] |
| 13 | Exercise Hours | Numerical (Continuous) | Patient’s exercise hours per week |
| 14 | Diet | Categorical (Ordinal) | Patient’s diet (healthy, unhealthy, average) |
| 15 | Sleep Hours | Numerical (Continuous) | Patient’s Sleep hours per day |
| 16 | Physical Activity | Numerical (Discrete) | Patient’s physically active days Per Week |
| 17 | Sedentary Hours | Numerical (Continuous) | Patient’s hours of inactivity per day |
| 18 | Alcohol Consumption | Categorical (Binary) | Patient’s Level of Alcohol Consumption [High;1 or Low;0] |
| 19 | Cholesterol | Numerical (Continuous) | Patient’s Cholesterol Level |
| 20 | Triglyceride | Numerical (Continuous) | Level of lipid fat in the body |
| 21 | Blood Pressure | Numerical (Continuous) | Patient’s blood pressure (systolic & diastolic) |
| 22 | Heart Rate | Numerical (Continuous) | Patient’s heart rate |
| 23 | Diabetes | Categorical (Binary) | Patient’s diabetes status [Yes;1 or no;0] |
| 24 | Previous Heart Problems | Categorical (Binary) | Previous Heart Problems [Yes;1 or no; 0] |
| 25 | Medication Use | Categorical (Binary) | Medication use by patients [Yes;1 or no; 0 |
| 26 | Heart Attack Risk | Categorical (Binary) | Heart Attack Risk in Patients |

Table 1: Data Features Description

## **Data Pre-processing**

R programming and its libraries were used to analyse and display the dataset to look for relationships between the different attributes. To do this, the data must first be cleaned, which entails looking for duplicates, missing values, and rows with "Null" values.

### Plotting to check for missing values.

While plotting the features to show if there are missing rows in the dataset, it was discovered that the dataset had no missing values as shown in Fig 1 below.

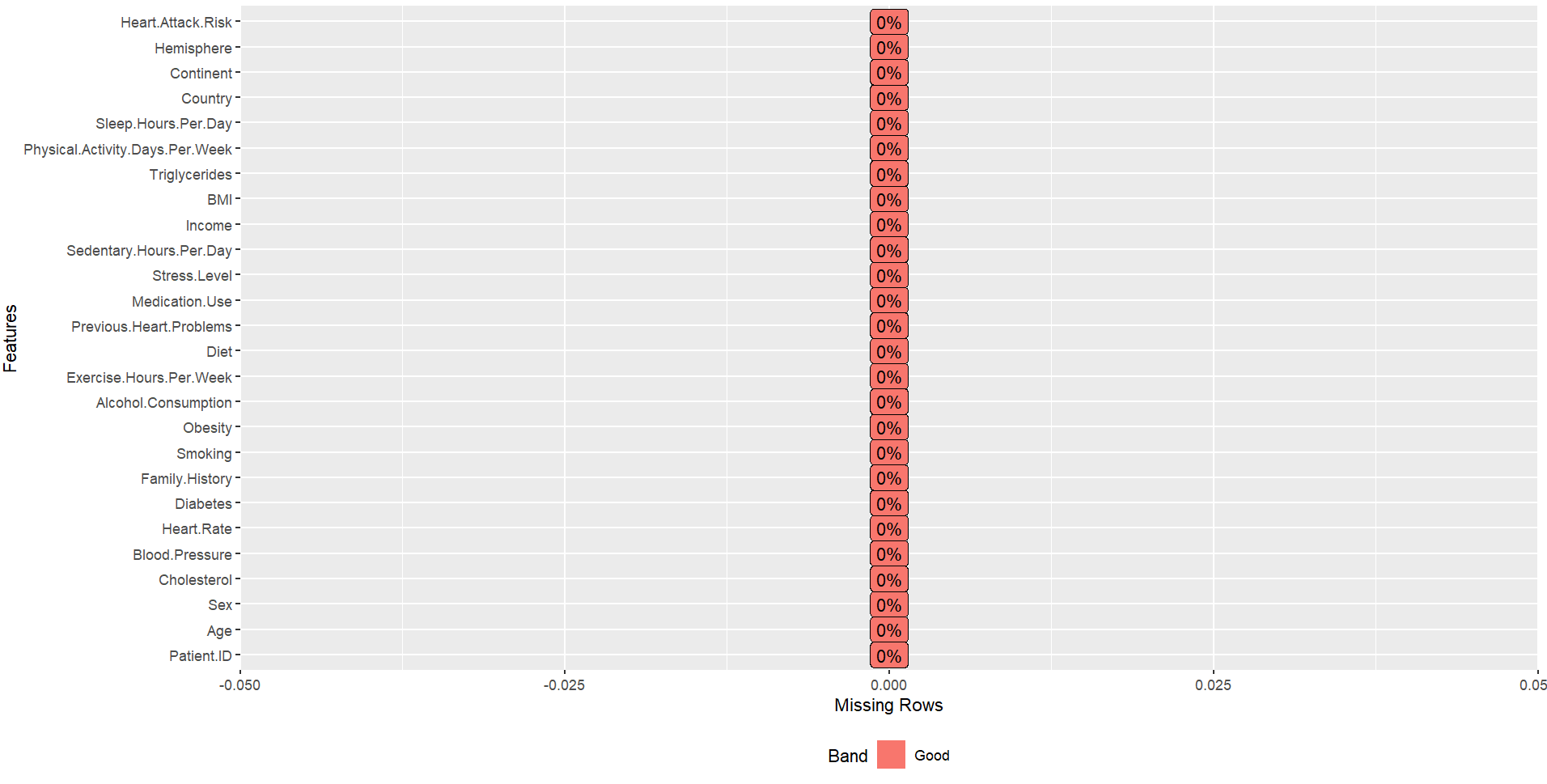


Figure 1: Plotting Missing Values

### Continent Selection

This dataset consists of records from 6 Continents which are: Asia, Africa, Australia, Europe, North America & South America). This analysis would focus on just one continent from the dataset which is “Europe.” Europe had 2241 observations and 26 observations.

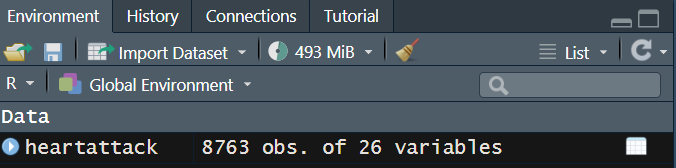


Figure 2: Observations before the selection of Europe

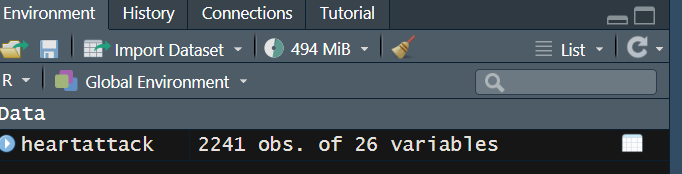


Figure 3: Observation after the selection of Europe

### Data Visualisation

To ensure proper understanding of the dataset various visualisations were carried out. A bar chart was plotted to show the distribution of each categorical features in the dataset, while a box plot was plotted to also show the distribution of other numerical features in the dataset.

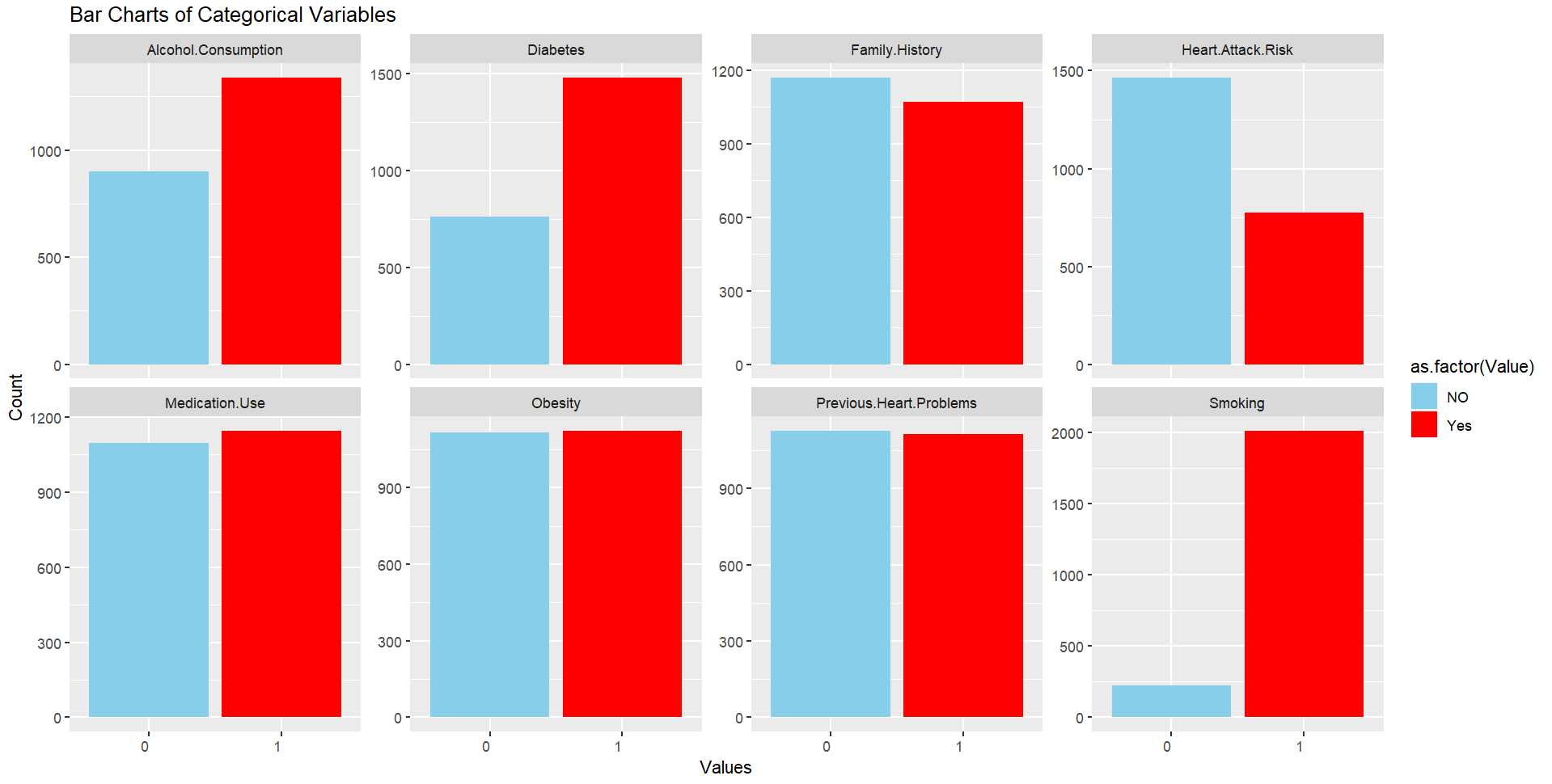


Figure 4: Bar Chart of Categorical Features

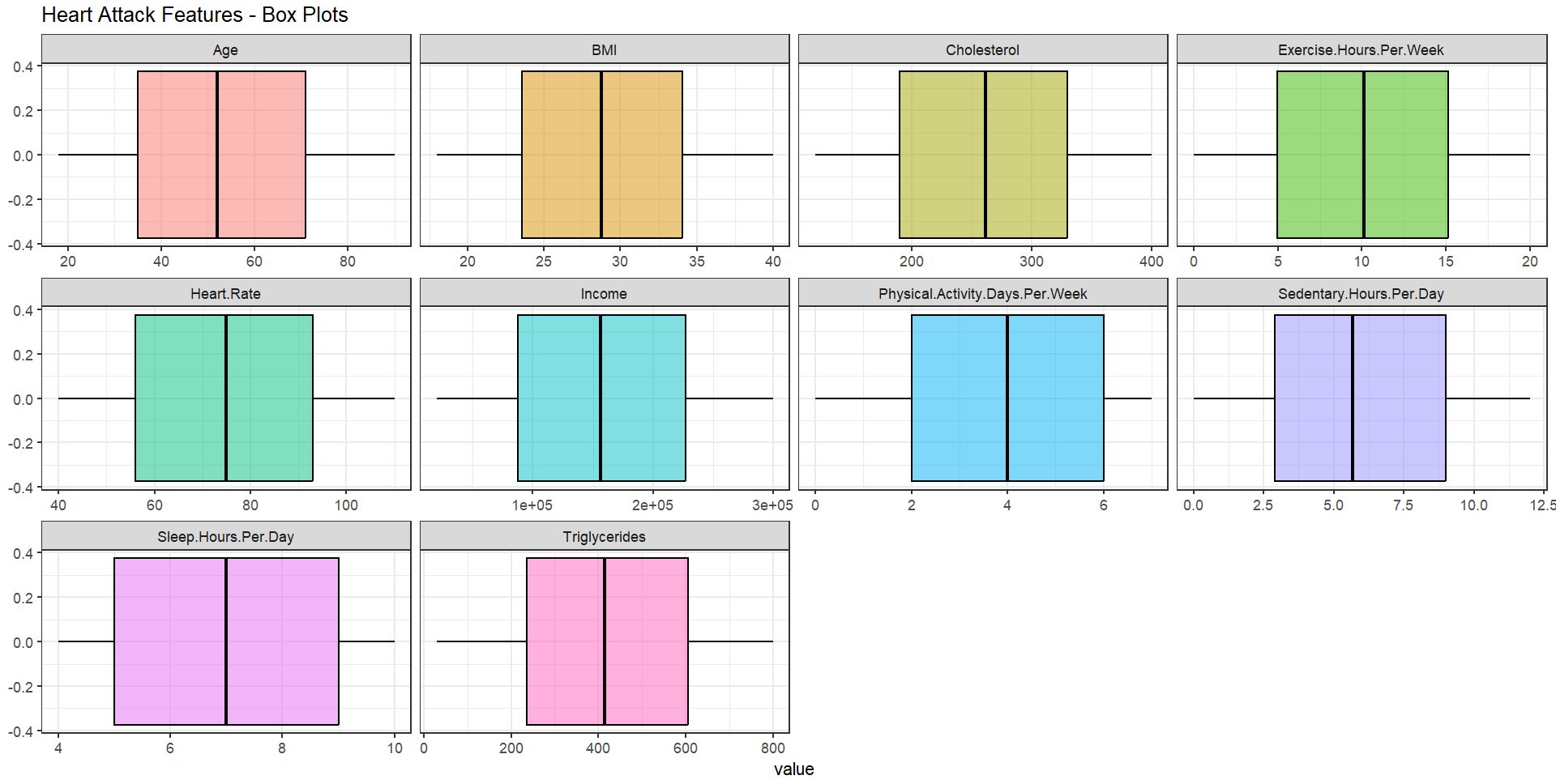


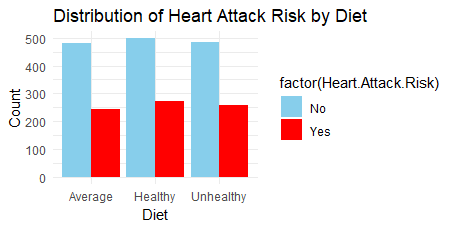
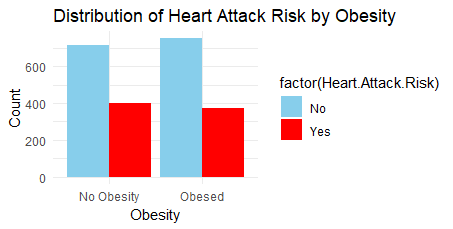
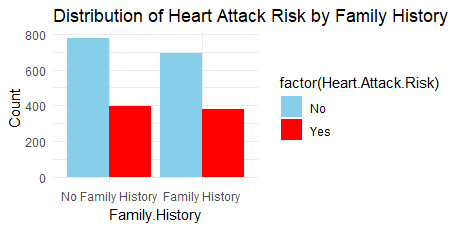
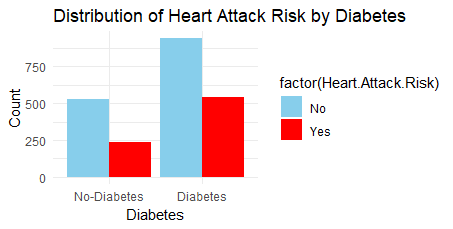
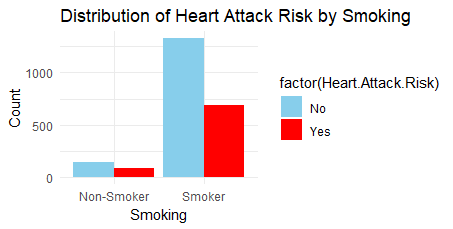
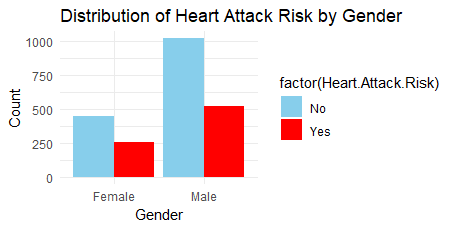
Figure 5: Box Plot showing the distribution of other numerical features in the dataset.

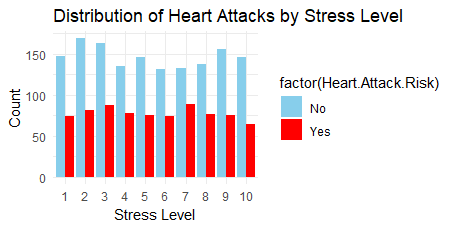
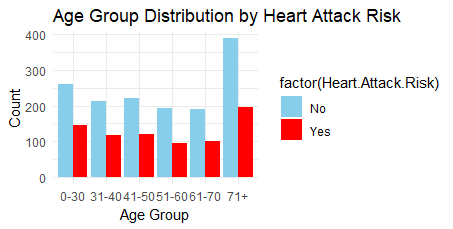
To further understand the relationships between various features in the dataset and the target variable which is the “Heart Attack Risk”, further visualisations were performed. The visualisations shown below show that the report had a record of more males than females, but males had more risk of a heart attack than females. The record also included more smokers than non-smokers, while also showing that smokers had more risk of a heart attack than smokers. It was also shown that people with diabetes had a higher risk of a heart attack than people without diabetes. The visualisation also shows that people with high alcohol consumption have a higher risk of a heart attack than people with low alcohol consumption.

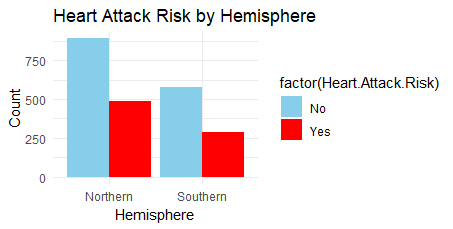
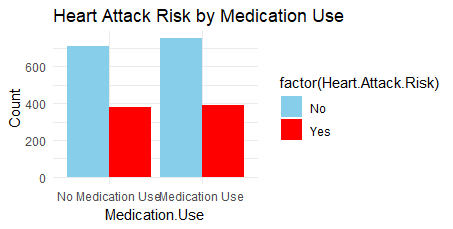
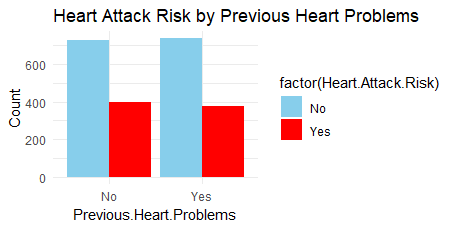
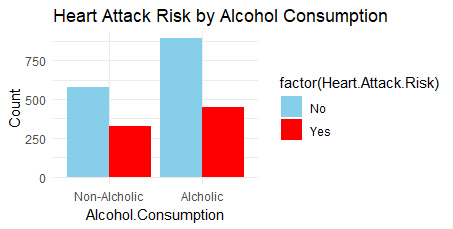
The visualisations didn’t give much insight as to whether a family history of heart attack put an individual at the risk of a heart attack as the values were the same for both, the same also goes for whether the patient is obese.

The Age and Income columns were grouped to ensure proper visualisation. It was shown that people within the age range of 71+ had a higher risk of a heart attack than people within the other age groups, while people with a higher income group had a higher risk of a heart attack.

The charts also didn’t give much insight as to whether an individual having a previous heart problem or being on medication use affected their risk of having a heart attack as the outcomes showed a balanced outcome. This shouldn’t be true as it is expected that an individual with a previous heart problem has a higher risk of a heart attack as well as an individual on medication use.







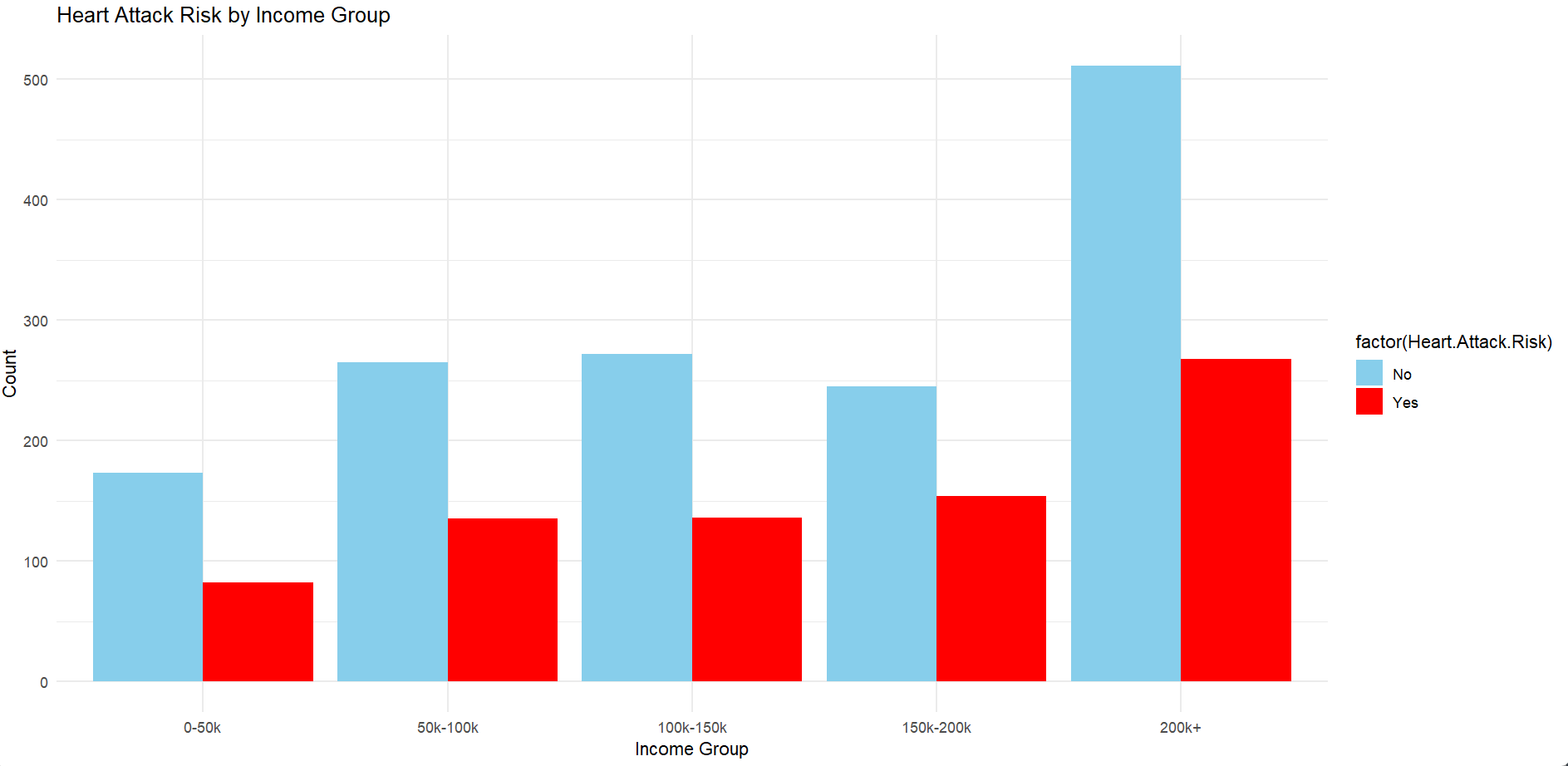


Figure 6: Figures showing the distribution of heart attack risk among various features.

### Data Cleaning

The process of fixing or removing erroneous, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset is called data cleaning. It increases overall productivity and gives you the greatest information to use when making decisions.

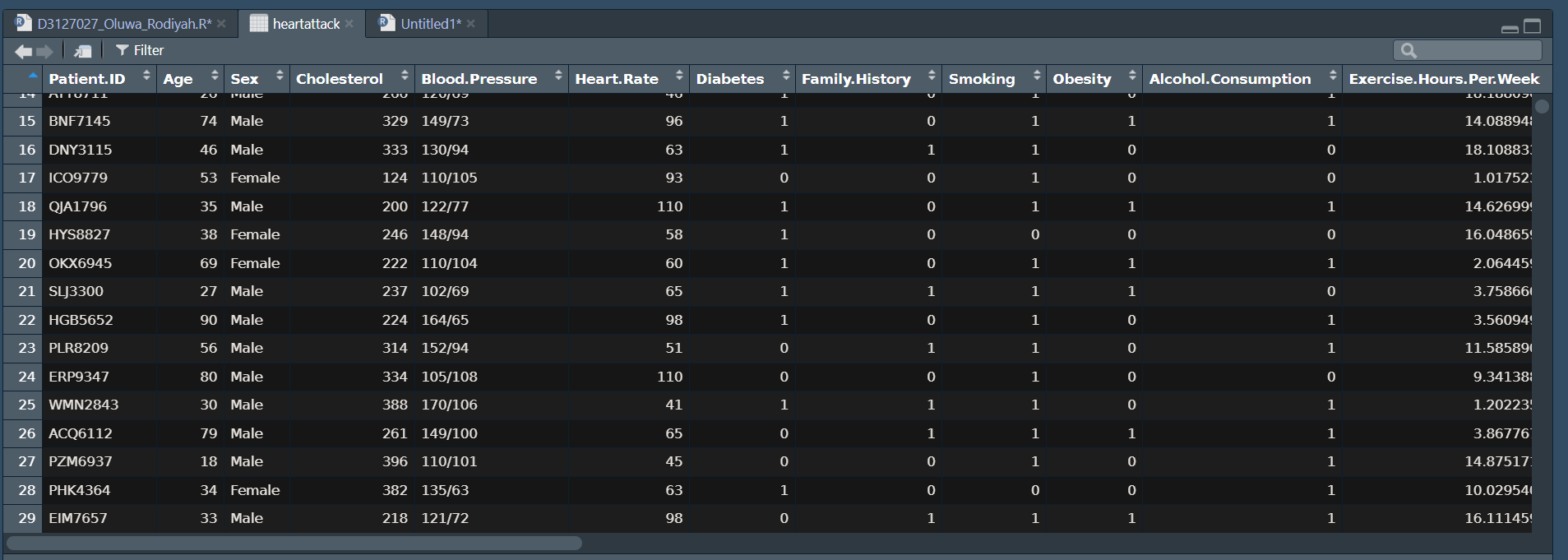


Figure 7: Dataset before cleaning

In this process, columns like Continent would be removed because the dataset only includes reports around Europe. The patient ID column was also removed because it wasn’t of any importance in this analysis, Country as well as Hemisphere were also removed from the dataset.

The blood pressure column was divided into Systolic & Diastolic, after dividing into columns, the features were in the character string format, so they were converted into numerical format.

Gender & Diet Columns were converted from character strings to numerical, while the Age Group and Income Group Columns created for visualisation were also removed.

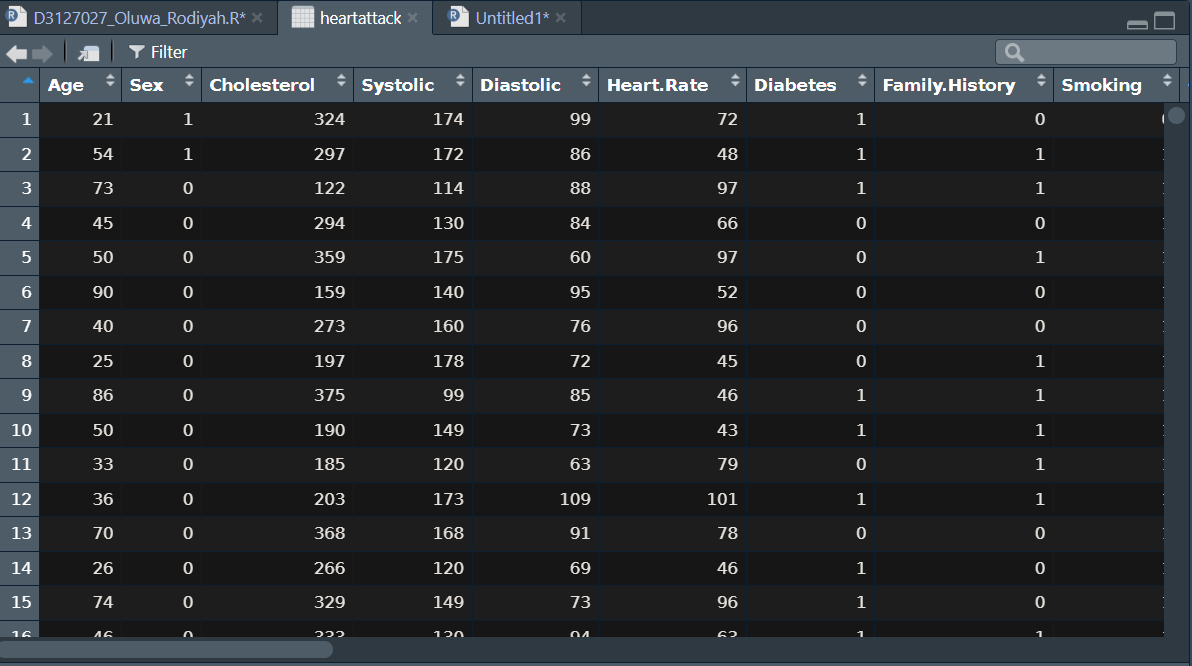


Figure 8: Dataset after Cleaning

### Data Distribution

After Data Cleaning, a distribution of the dataset was shown using a box plot.

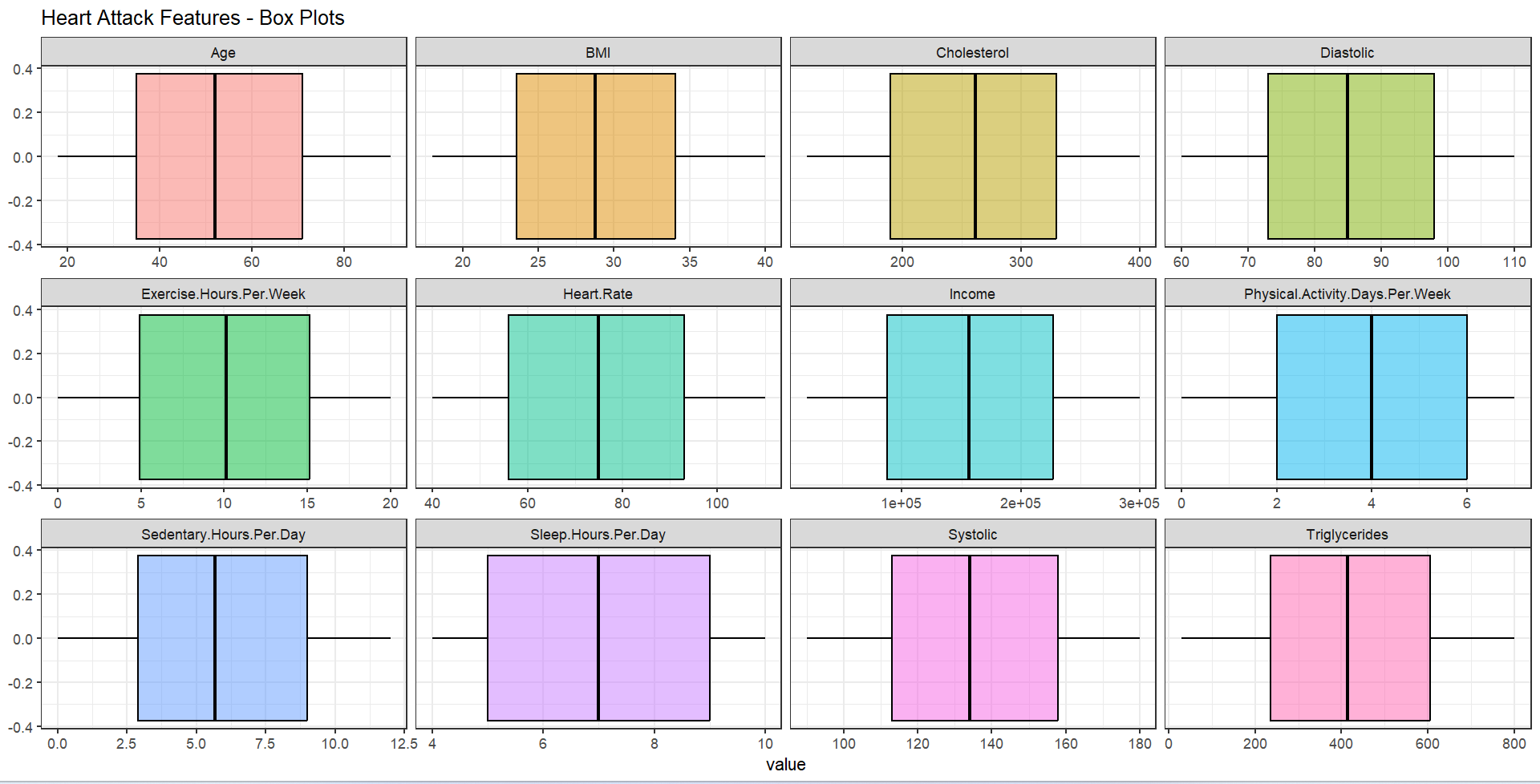


Figure 9: Box Plot showing the distribution of numerical features after data cleaning. The average age of the patients in this analysis is around 50. The average BMI of the patients is 28. The average cholesterol level of the patients is 270. The average blood pressure of the patients in systolic and diastolic are 130 and 85 respectively.

### Feature Selection

Feature selection was carried out using a correlation plot. To avoid multicollinearity, which can disrupt a model's coefficients and hinder data interpretation, highly correlated variables are removed. Additionally, this procedure lowers the chance of overfitting and improves computing efficiency.

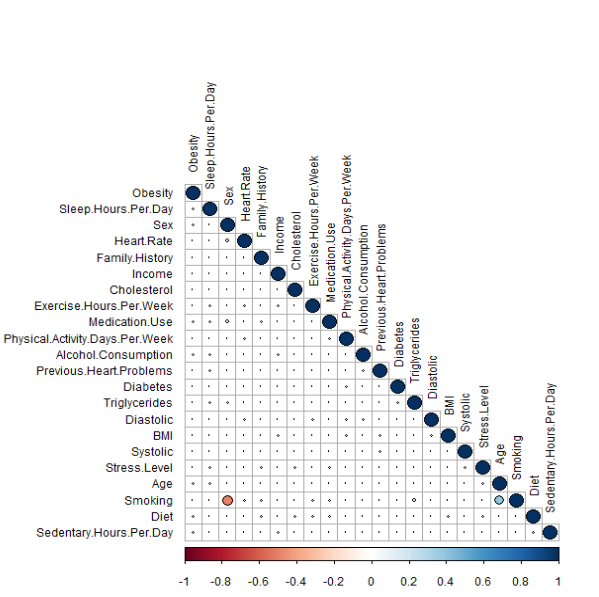


Figure 10: Correlation Plot

The correlation plot is interpreted using the colours in the plots while using the scale in the plot. The scale shows that from -0.6 to 0.6, the colours are light, and deeper colours start to show between -0.6 and -1 & between 0.6 and 1. The deeper colours represent highly positive and negative correlations between the features. The plot shows that there’s a notable negative correlation between smoking and sex, while there’s a moderate positive correlation between smoking and age. Using 0.6 as the cut-off, the correlation plot showed that no feature in the dataset was highly correlated.

### Feature Scaling/ Normalisation

Feature scaling, sometimes referred to as normalisation, is a data pre-processing technique whereby numerical characteristics are normalised to a standard range. This ensures a fair and uniform influence on machine learning models, particularly in cases when features have disparate scales or units. It helps enhance the performance and convergence of the model as well as the efficient application of methods that are sensitive to different input magnitudes.

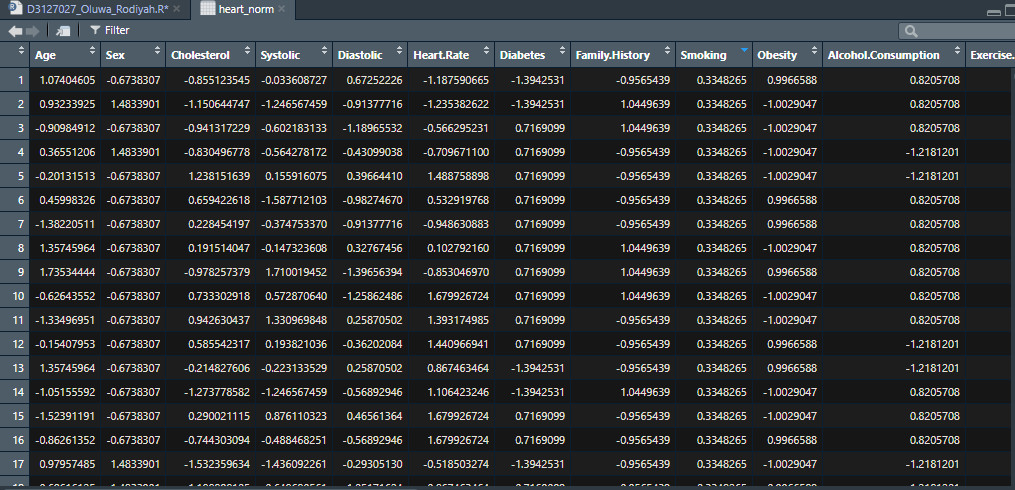


Figure 11: Scaling Features to ensure uniformity in the dataset.

### Smote Balancing

The SMOTE (Synthetic Minority Over-sampling Technique) algorithm is an effective method for handling the issue of unbalanced datasets in machine learning. SMOTE improves generalization and prediction performance by assisting classifiers in capturing patterns in the minority class more effectively. This is accomplished by smoothing the class distribution. This method is especially useful in situations where unbalanced datasets can impair the efficacy of predictive models, such as fraud detection, medical diagnosis, and other applications. The SMOTE balancing carried out in this research was scaled on the target variable on a 60/40 balancing.

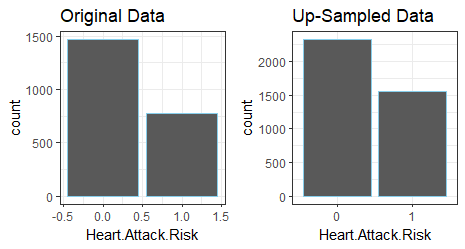


Figure 12: SMOTE Algorithm.

### Data Preparation

#### **Data Splitting**

The dataset underwent the appropriate preparations in the first phase of the study to make the modelling tasks that followed easier. The target variable, which represented the risk of an individual having a heart attack, was converted into a factor, as is frequently done in classification problems. The dataset was then split into training and testing sets, with 30% of the data set aside for testing and 70% of the data set for training. The goal of this division was to guarantee an objective evaluation of the model's performance on unobserved data. A separate data frame was created for the target variable for both the testing and training dataset. The data were also separated using the x and y axis, where the dependent variable (target class) was represented by the y-axis and the remaining attributes as independent variables.

### Machine Learning Classifications

In this study, an individual's risk of a heart attack based on lifestyle and health factors will be predicted using two machine learning (ML) algorithms. K-nearest neighbour (KNN) and Support Vector Machines (SVM) are the two ML algorithms that would be carried out.

When it comes to supervised machine learning, Support Vector Machines (SVMs) are a potent paradigm that can be applied to both regression and classification tasks. Notably, SVMs have clear advantages over modern algorithms such as neural networks, especially when faced with small datasets. They are faster and more efficient, especially when dealing with a few thousand samples. This effectiveness is especially noticeable in text classification tasks, where SVMs perform well even with a small collection of examples that have been tagged. One of SVM's inherent advantages is its unique training strategy, which reduces the impact of vector dimensionality on the model's performance when compared to other classifiers. This allows SVMs to handle extremely high-dimensional environments with ease. As a result, SVMs have shown to be extremely useful in large-scale, categorized settings, which has led to their widespread use in various applications.

Another effective supervised machine learning approach for both regression and classification applications are the K-Nearest Neighbour (KNN). In contrast to Support Vector Machines (SVM), KNN learns by memorization of the whole training dataset, rather than by intentionally creating a model during training. Using the average value or majority class of their k-nearest neighbours in the feature space, this straightforward method predicts or classes new occurrences. The critical value 'k' controls how many neighbours are considered, which affects how sensitive the algorithm is to noise. Though conceptually straightforward and easy to comprehend, KNN can have significant processing requirements, particularly when dealing with big datasets, because it must calculate the distances between each new instance and all the training instances.

## **Machine Learning Application**

The architectural structure of a machine learning model can be established using a variety of design options. It is necessary to try out a variety of options because the best model architecture for a certain model isn't always known. To do this, the computer is given complete control over the exploration and model architecture selection processes. Hyperparameters are those that, in the context of models, affect the model design. Determining the optimal model architecture is therefore a process called "hyperparameter tuning." Cross-validation, on the other hand, is a sampling technique that is used to analyse machine learning algorithms and forecast their potential performance on alternative test datasets.

Support Vector Machine (SVM) modelling was used in the study to forecast the risk of heart attacks. To maximize its performance, the SVM model's hyperparameters were set up differently. To robustly evaluate the model across various subsets of the training data, frequent cross-validation was used during the tuning phase. The training dataset was used to train the SVM model, and the confusion matrix was computed to evaluate the model's performance on the test set. The matrix shed light on how well the model predicted the likelihood of heart attacks in certain situations.

The usefulness of K-Nearest Neighbours (KNN) as a substitute classification strategy was also investigated. Several values of 'k', which denotes the number of nearest neighbours considered during classification, were used to develop the first KNN models. The precision of every model was assessed, and a hyperparameter tuning procedure was started to determine the ideal 'k' value. The accuracy of each iteration of the systematic iteration across potential "k" values was noted. To assist in determining the best configuration, a plot was created to show the relationship between accuracy and 'k' values.

## **Performance Evaluation**

In this analysis, a confusion matrix will be used to evaluate the performance of each ML Algorithm carried out. A Confusion Matrix is a tabular representation used in machine learning to evaluate a classification model's performance. Another term for it is an error matrix. The rows of the matrix show instances from real classes, and the columns show instances from predicted classes or the other way around. It does this by comparing the model's predictions to the actual results found in a dataset. There are several parts to the matrix but just a few metrics from them would be used to evaluate this model. They are Accuracy, Kappa, Sensitivity, Specificity, Positive predictive value & a Negative predictive value. ROC-AUC Curve would also be plotted for each model for clarification of which model outperforms chance. The performance metrics selected for this analysis are explained further. The Accuracy shows to what extent is the model accurate. Cohen’s Kappa Shows to what extent the model outperforms Chance. True Positive Rate or Sensitivity shows what the model's detection rate is for positive cases. True Negative Rate or Specificity shows what the model's detection rate is for negative cases. The model's Positive Predictive Value/Precision indicates the frequency of true predictions made by the model. How frequently the model is accurate when it makes a negative prediction is indicated by the negative predictive value. The trade-off between the true positive rate (sensitivity) and the false positive rate (specificity), which is visually represented by the ROC curve, is determined by the threshold settings of a binary classification model. An illustration of a model's performance over a range of classification thresholds is given by the ROC curve. An indicator of better performance is a curve that approaches the upper left corner. AUC stands for area under the curve, and it represents by the line beneath the ROC curve. It gives an overview of a binary classification model's performance over all feasible thresholds in a single scalar value. A model with a discrimination (or better separation of positive and negative cases) AUC closer to 1 is better.

The interpretation for the AUC is that when:

AUC = 1: Perfect classifier.

AUC > 0.5: Better than random guessing.

AUC = 0.5: Model performs no better than random chance.

AUC < 0.5: Model performs worse than random chance.

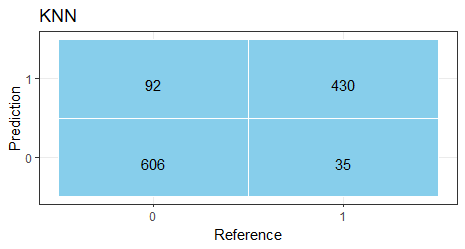
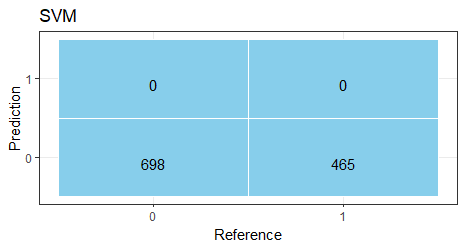


Figure 13: Confusion Matrix Plot for both models

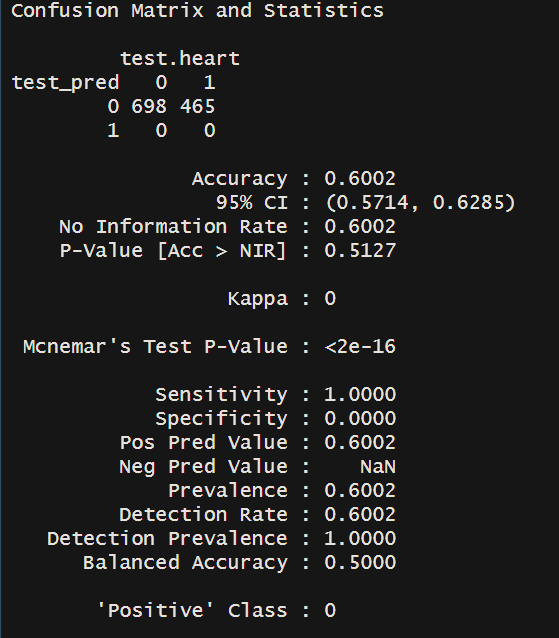


Figure 14: Confusion Matrix Table for SVM

The confusion matrix table for the SVM shows that the accuracy of the model is 60%, It also shows that Kappa is 0 indicating that the model doesn’t outperform chance. When the Kappa value of a model is zero, it means that its performance is no better than chance.

It shows that the True Positive Rate / Sensitivity is 100%, indicating that the model only detects positive values. True Negative Rate / Specificity is 0%, indicating that the model didn’t detect any negative value correctly, this might point to a problem in determining who is not at risk of having a heart attack.

Positive Predictive Value/ Precision is 60%, indicating that the model was 60% correct when it detected a positive value. The negative Predictive Value in the model was “NaN”, this shows that the model didn’t predict any negative value.

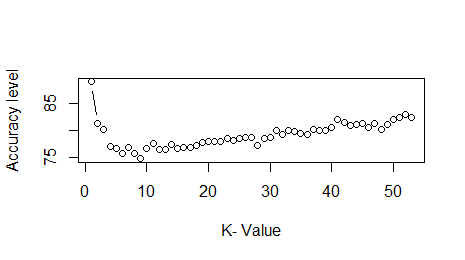


Figure 14: Plot showing accuracy for different values of “K”.

The plot shows the accuracy for different values of K after hyper-parameter tuning was carried out. It shows that accuracy was at its highest when K=1, so a confusion matrix table would be printed to show the other metrics for when k=1.

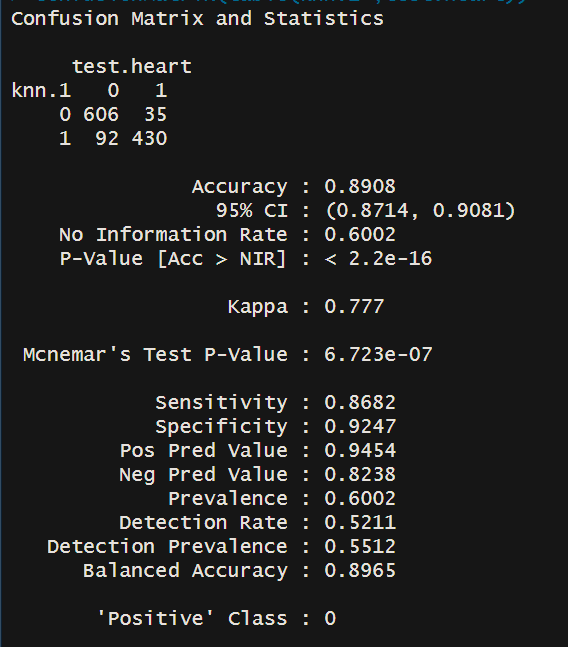


Figure 14: Confusion Matrix Table for KNN

The confusion matrix table for the KNN model shows that the accuracy of the model is 89.08%, It also shows that Kappa is 77.7% indicating that the model outperforms chance by 77%. A high Kappa value indicates that the actual results and the model's predictions accord significantly.

True Positive Rate / Sensitivity is 86.82%, indicating that the model detection rate for positive values is high. This implies that the model is effective in accurately identifying people who are at risk of having a heart attack.

True Negative Rate / Specificity is 92.47%, indicating that the model’s detection rate for negative value is 92.47. This implies that the model is effective in accurately identifying people who are not at risk of having a heart attack.

Positive Predictive Value/ Precision is 94.54%, indicating that the model was 94.54% correct when it predicted a positive value, while the negative Predictive Value of the model is 82.38%, this shows that the model was 82.38% correct when it predicted a negative value.

# **RESULT AND CONCLUSIONS**

## **Accuracy**

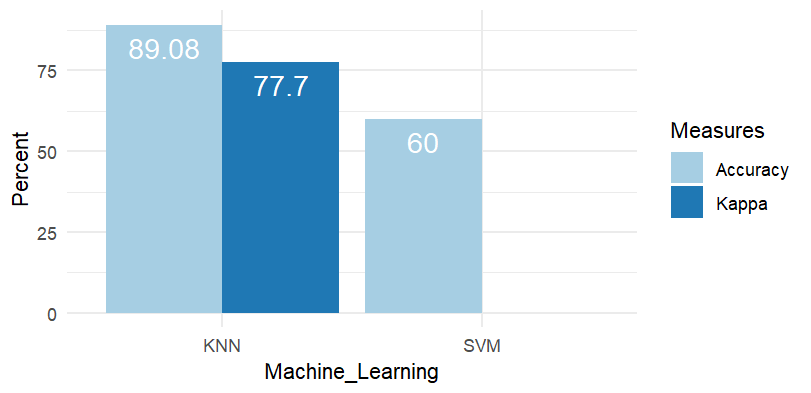


Figure 16: Accuracy and Kappa for SVM AND KNN

The accuracy level of the two algorithms was compared on the given dataset after tuning the parameters. K-Nearest Neighbour at 89.08% is the most accurate.

The SVM model shows great sensitivity in identifying affirmative cases, but its total accuracy is only 60%, and the Kappa value indicates that it does not surpass chance. The model's shortcomings in terms of negative predictive value and specificity should be considered. Additional research may be required to improve the model's performance, particularly in terms of identifying those who are not in danger of having a heart attack.

The KNN model on the other hand performs exceptionally well in terms of accuracy and Kappa values, demonstrating its strong predictive ability. The model demonstrates competency in recognizing both positive and negative cases by skilfully balancing sensitivity and specificity. These results imply that the KNN model has the potential as a trustworthy instrument for estimating the risk of a heart attack, with a noteworthy capacity to produce relevant and accurate predictions across a range of outcomes.

## **ROC-AUC Curve**

The performance of the two algorithms was also compared on the given dataset using the ROC-AUC Curve.

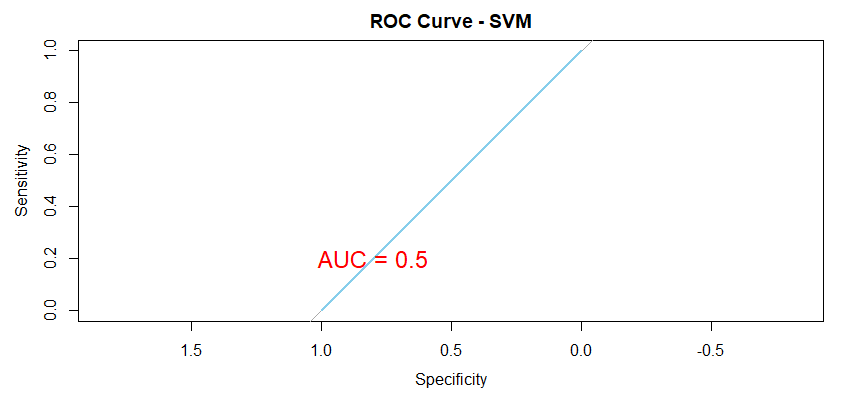


Figure 17: ROC-AUC for SVM

The SVM model's ROC curve revealed a flat line that was precisely on the diagonal (AUC = 0.5), indicating that the model has no capacity for discrimination and performs no better than chance. The diagonal line, which shows the trajectory of a random classifier, indicates that, given the available features, the model is unable to discriminate between positive and negative classes. An AUC of 0.5 is a crucial cut-off point for evaluating models since it shows that the predicted performance is meaningless and contributes nothing more significant than guesswork.

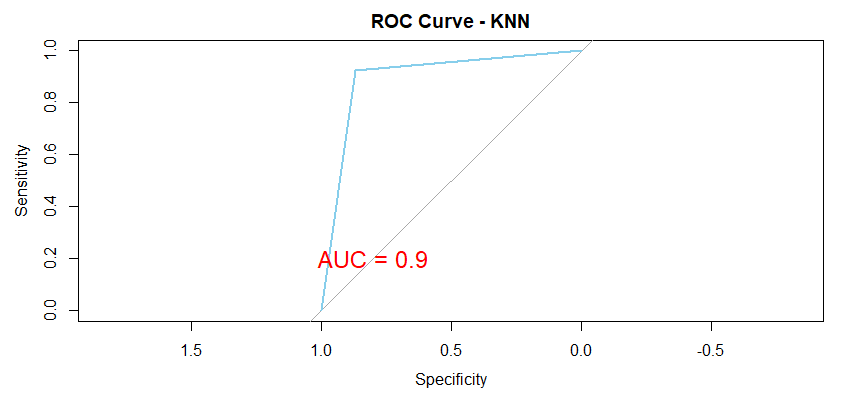


Figure 18: ROC-AUC for KNN

With an Area Under the Curve (AUC) of 0.9, the k-Nearest Neighbours (KNN) model demonstrated a highly favourable Receiver Operating Characteristic (ROC) curve with a noticeable upward sweep to the left. This performance indicates a high degree of discriminatory power, demonstrating how well the model can distinguish between positive and negative instances at different threshold values. The model's accuracy and dependability are supported by the increased AUC of 0.9, which shows a high true positive rate and low false positives. Practically speaking, the KNN model has strong prediction skills, which makes it an effective tool for precise categorization jobs.

In conclusion, the KNN model outperforms the SVM model in terms of discriminatory performance, showing great promise for real-world application in classification problems.

# **RECOMMENDATION**

This study offers opportunities for additional research since its goal was to create a forecasting model for estimating the risk of heart attacks based on lifestyle and health variables. In-depth testing, concentrating on optimizing hyperparameters and utilizing group techniques like stacking and bagging, should be part of subsequent research to improve model performance. Gaining understanding of the significance of each characteristic would improve the model's interpretability by revealing important details about the major factors that influence the risk of heart attacks. To improve the predictive models' resilience and relevance, it is advised to supplement the analysis by adding genetic features to the dataset. Given the importance of genetic traits for cardiovascular health, using them in risk assessment could provide insightful information. It is also recommended to expand the research to include datasets from different continents to take regional differences in cardiovascular health into consideration. Moreover, finding the best model configurations requires optimising hyperparameters using more sophisticated methods or thorough searches.  Investigating model ensembling techniques like stacking or bagging may improve prediction robustness and accuracy. These suggestions open the door to a more thorough comprehension of heart attack risk factors and the creation of predictive models that are applicable everywhere. To make sure the model is dependable across various populations, it is advised to use distinct datasets for external validation. An evolving understanding of the risk factors for heart attacks over time would result from expanding the study to incorporate longitudinal analysis. The creation of a customized risk assessment tool that provides individualized health advice might also be the focus of the work. Incorporating the model into clinical settings would be made easier by working together with healthcare professionals, guaranteeing its usefulness and efficacy in actual situations.

# **REFERENCES**

Dataset used for this research - <https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset>

Bhatt, C.M., Patel, P., Ghetia, T. and Mazzeo, P.L., 2023. Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), p.88.

Manikandan, S., 2017, August. Heart attack prediction system. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 817-820). IEEE.

Obasi, T. and Shafiq, M.O., 2019, December. Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attacks and Diseases. In *2019 IEEE international conference on big data (big data)* (pp. 2393-2402). IEEE.

Riyaz, L., Butt, M.A., Zaman, M. and Ayob, O., 2022. Heart disease prediction using machine learning techniques: a quantitative review. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 3* (pp. 81-94). Springer Singapore.

Shah, D., Patel, S., and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, pp.1-6.

World Health Organisation., 2023. Cardiovascular Diseases.

European Heart Network.,2023. Cardiovascular Diseases