Predicting Cardiovascular Disease Risk with Machine Learning

By Lara Oriol Cabrera

INDEX

This is my second case study for my data portfolio, where I am focusing on predicting heart disease risk using two machine learning methods. After my university experience in Business Administration and Management, and my first case study as part of the Google Data Analytics Professional Certificate, I have been able to learn new techniques and improve my approach to data analysis and model implementation. In this project, my goal is to apply the skills I've learned to develop effective models that predict the risk of heart disease using a clinical dataset. Throughout this work, I will put my knowledge of statistical analysis, and machine learning into practice, aiming to provide a precise and valuable solution.

MOTIVATION

Cardiovascular disease remains one of the most prevalent and life-threatening health conditions globally, continuing to be a leading cause of death across various populations. In the World Health Organization European Region, Cardiovascular Diseases are responsible for over 42.5% of all deaths annually, equating to approximately 10,000 lives lost each day. A significant portion of these fatalities is considered preventable, emphasizing the crucial role of early detection, precise risk assessment, and timely intervention in reducing mortality rates. Despite advancements in medical research, traditional risk prediction models often struggle to effectively identify individuals who would benefit the most from preventive treatments, while at the same time, they may lead to unnecessary interventions for others who may not be at risk.

Machine learning presents an opportunity to improve the accuracy of cardiovascular risk prediction by analyzing the non-linear relationships between multiple risk factors. Unlike traditional statistical models, machine learning algorithms can detect complex patterns within datasets, enabling more refined and individualized risk assessments. Recent studies have demonstrated the superiority of machine learning approaches in cardiovascular risk prediction, with one particular study analyzing 378,256 patients from UK general family practices. The findings revealed that machine learning models outperformed conventional risk prediction algorithms, enhancing predictive accuracy by an estimated 1.7% to 3.6%.

By developing a robust machine learning-based model for heart disease risk prediction, we aim to contribute to a more precise and personalized approach to cardiovascular health. Enhanced predictive capabilities could facilitate better-targeted preventive strategies, improve patient outcomes, and ultimately help reduce the healthcare burden associated with cardiovascular diseases. The integration of machine learning into risk assessment frameworks holds significant potential to reshape preventive cardiology, paving the way for more efficient and data-driven medical decision-making.

HYPOTHESIS

The primary objective of this study is to evaluate the effectiveness of machine learning techniques, specifically logistic regression and decision tree models, by predicting the likelihood of an individual developing heart disease based on a range of established risk factors. By systematically analyzing and comparing the predictive performance of these two models, this study aims to assess their accuracy, reliability, and overall suitability for cardiovascular risk assessment. The central hypothesis is that machine learning algorithms can enhance the precision of heart disease prediction, thereby providing valuable insights that can facilitate early detection and the implementation of targeted preventive measures. Furthermore, it is anticipated that one of these models will demonstrate superior predictive capabilities compared to the other, ultimately surpassing the accuracy of conventional risk assessment methods.

BENEFITS OF THE STUDY

Machine learning algorithms have demonstrated a substantial advancement in predictive accuracy compared to traditional risk prediction methods, making them a powerful tool in cardiovascular disease prevention and management. These advanced models have the ability to generate more reliable and precise predictions by effectively distinguishing individuals at high risk for cardiovascular disease from those at low risk. This enhancement in both sensitivity and specificity is crucial for the early identification of at-risk individuals, ultimately enabling healthcare providers to implement more targeted and effective preventive interventions.

A key advantage of machine learning techniques lies in their capacity to integrate and analyze a broader range of risk factors, including demographic information, lifestyle behaviors, genetic predispositions, laboratory test results, and pre-existing medical conditions. Unlike conventional models, which may be limited in their ability to account for complex, non-linear interactions between variables, machine learning algorithms can process large datasets and uncover intricate patterns that might otherwise go undetected. This results in a more comprehensive and nuanced risk assessment, allowing for highly personalized predictions tailored to an individual's unique health profile.

Moreover, the automation of data processing and analysis through machine learning significantly reduces the burden on healthcare professionals, streamlining the risk assessment process and enabling faster, more efficient clinical decision-making. By leveraging these capabilities, machine learning has the potential to enhance patient outcomes, optimize resource allocation in healthcare settings, and contribute to the broader goal of reducing the global burden of cardiovascular disease. As these technologies continue to evolve, their integration into clinical practice could lead to more proactive and data-driven approaches to cardiovascular disease prevention and treatment, ultimately improving public health outcomes on a larger scale.

DATABASE PRESENTATION

The “Heart Disease Risk Prediction Dataset” is a synthetic dataset specifically designed to assist in predicting the likelihood of individuals developing heart disease. This dataset includes a range of symptoms, lifestyle factors, and medical history indicators, which collectively contribute to the prediction of heart disease risk. Each entry in the dataset represents a patient and includes binary indicators (Yes/No) for various symptoms and risk factors. Additionally, each entry is labeled with a computed risk label that categorizes the patient as either being at high or low risk of developing heart disease.

With a total of 65.535 observations, this dataset is well-suited for training machine learning models aimed at classification tasks, making it an ideal resource for researchers, data scientists, and healthcare professionals looking to explore predictive modeling in the context of cardiovascular health.

The dataset was developed by students at the Vellore Institute of Technology (VIT-AP) as part of the EarlyMed initiative. EarlyMed seeks to leverage data science and machine learning techniques to promote the early detection and prevention of chronic diseases, particularly heart disease. By providing a clean and structured dataset, the project aims to empower individuals and organizations to improve healthcare outcomes through predictive analytics.

The input features in this dataset include both symptoms and risk factors. The symptoms, which are binary in nature (Yes/No), include indicators such as chest pain, shortness of breath, unexplained fatigue, palpitations, dizziness or fainting, swelling in legs or ankles, radiating pain in the arm/jaw/neck/back, and cold sweats and nausea. These symptoms are commonly associated with heart disease and play a significant role in the early detection of cardiovascular conditions.

For the risk factors, both binary (Yes/No) and continuous variables are present, such as age, hypertension, high cholesterol, diabetes, smoking history, obesity, and a family history of heart disease. These factors have been well-documented as contributing to an individual's risk of developing cardiovascular diseases. The output label, denoted as the heart disease risk label, is a binary classification where a value of "No" indicates low risk and a value of "Yes" signifies high risk.

Variable	Description	Data Type
Chest_Pain	Presence of chest pain, a common symptom of heart disease.	Binary (Yes/No)
Shortness_of_Breath	Difficulty of breathing, often associated with heart conditions	Binary (Yes/No)
Fatigue	Persistent tiredness	Binary (Yes/No)
Palpitations	Irregular or rapid heartbeat	Binary (Yes/No)
Dizziness	Episodes of lightheadedness	Binary (Yes/No)
Swelling	Swelling in legs/ankles due to fluid retention.	Binary (Yes/No)
Pain_Arms_Jaw_Back	Radiating pain in arms, back and jaw.	Binary (Yes/No)
Cold_Sweats_Nausea	Symptoms commonly associated with acute cardiac events.	Binary (Yes/No)
High_BP	History of high blood pressure	Binary (Yes/No)
High_Colesterol	Elevated cholesterol levels	Binary (Yes/No)
Diabetes	Diagnosis of diabetes	Binary (Yes/No)
Smoking	Whether the patient is a smoker	Binary (Yes/No)
Obesity	Obesity status	Binary (Yes/No)
Sedentary_Lifestyle	Whether the patient has a sedentary lifestyle (lack of physical activity)	Binary (Yes/No)
Family_History	Family history of cardiovascular conditions	Binary (Yes/No)
Chronic_Stress	Whether the patient suffers from chronic stress	Binary (Yes/No)
Gender	Patient’s gender	Binary (Yes/No)
Age	Patient’s age	Continuous
Heart_Risk	Heart disease risk	Binary (Yes/No)

The data generation process for this dataset was carried out using popular Python libraries such as numpy and pandas. The generation process ensured that the dataset contained a balanced distribution of both high-risk and low-risk cases, while maintaining realistic correlations between the various features. For example, individuals who exhibited multiple risk factors, such as smoking, hypertension, and diabetes, were more likely to be classified as high-risk. Symptom patterns in the dataset were modeled after clinical guidelines and well-established research studies on heart disease.

The design of this dataset was influenced by several authoritative sources, including textbooks such as Harrison's Principles of Internal Medicine and Mayo Clinic Cardiology, which provide valuable insights into the symptoms and risk factors of heart disease. Research studies like the Framingham Heart Study and the guidelines from the American Heart Association (AHA) have also informed the creation of this dataset. Furthermore, the dataset was inspired by existing publicly available resources, such as the UCI Heart Disease dataset and Kaggle's various heart disease datasets.

This dataset is part of a larger student-driven project, EarlyMed, which is a collaborative initiative to explore the role of data science in the early detection and prevention of chronic diseases. The dataset was created by Mahatir Ahmed Tusher, Saket Choudary Kongara, and Vangapalli Sivamani, who are students from VIT, and aims to support the growing field of data-driven healthcare innovation.

Data Limitations

While the dataset serves as a valuable resource for predictive modeling, it has several limitations that should be considered when using it for analysis or machine learning tasks.

Firstly, the dataset is synthetic, meaning it was artificially generated rather than derived from real-world clinical data. Although it is designed to simulate real-world patterns, it may not fully capture the complexity of actual patient data. The correlations between features in the dataset are approximations and might not reflect the true relationships observed in clinical practice. As such, while it offers a structured and balanced dataset for modeling, it may not account for all the nuances and variability found in real healthcare data.

In addition, the output label in this dataset is binary, classifying patients as either yes or no risk for heart disease. In reality, risk stratification is often more nuanced, involving multiple levels of risk (e.g., low, moderate, high). This simplification of the risk label might limit the ability to develop models that handle more detailed and layered risk assessment that healthcare professionals would need in real-world settings.

Finally, the dataset does not include missing values, which are a common occurrence in real-world healthcare data. Missing data can arise due to various reasons, such as incomplete patient records, data entry errors, or patients withholding certain information. The absence of missing values in this dataset means that it may not reflect the challenges often faced when working with real-world healthcare data, such as the need for data imputation or handling incomplete information.

Data ROCCC approach

1. Reliable:
The reliability of the dataset is a bit mixed. The dataset is synthetic, which means it does not come from real-world clinical data but is modeled to mimic the patterns of heart disease risk factors and symptoms. While the dataset uses known clinical guidelines and sources, the accuracy and consistency of the data are approximations rather than reflections of real patient data. In machine learning tasks, synthetic datasets like this can still provide useful insights, but they may not always capture all the subtleties and complexities of real-world data.

2. Original:
The dataset is original in the sense that it was created specifically for this project by students at Vellore Institute of Technology (VIT-AP) as part of the EarlyMed initiative. It is not a direct copy or derivative of other datasets but was designed based on real-world clinical guidelines and studies. However, while it is original in its construction, it is still synthetic and does not come from direct clinical practice or real patient records.

3. Comprehensive:
The dataset is comprehensive within the scope of heart disease risk factors. It includes a wide range of symptoms (e.g., chest pain, shortness of breath, fatigue) and known cardiovascular risk factors (e.g., age, high blood pressure, diabetes, smoking). This makes it suitable for the task of predicting heart disease risk. However, it simplifies the output to a binary "low" or "high" risk label, which may not fully capture the complexity of real-world cardiovascular risk stratification, where multiple levels of risk (low, moderate, high) would be more typical.

4. Current:
The dataset was updated a month ago. However, since it is synthetic, it might not reflect the most up-to-date medical research or the latest clinical guidelines. The data generation process mimics real-world relationships but may not include the latest advancements in medical understanding or incorporate newer research findings. In that sense, the dataset can be considered current within its scope, but it does not have the same level of freshness as datasets derived directly from recent clinical data.

5. Cited:
The dataset is properly acknowledged in terms of its origin, as it is part of the EarlyMed initiative developed by students at VIT-AP. Additionally, it draws inspiration from established medical resources like the Framingham Heart Study, Mayo Clinic Cardiology, and the American Heart Association (AHA) guidelines. However, because it is a synthetic dataset, it is important to note that the specific details of the data generation process and the sources used to generate the data are not fully cited in a formal academic manner. While the project credits its academic contributors and inspiration sources, there may not be a formal citation in the dataset itself.

Data Cleaning

Since the dataset is synthetic, it has already been cleaned, eliminating common issues such as missing values and inconsistencies. This pre-processing ensures that the data is ready for immediate use in machine learning tasks without the need for additional cleaning or handling of incomplete entries.

DESCRIPTIVE ANALYSIS

Overall inspection of the dataset

In this section, we begin by loading the dataset into RStudio, which will serve as the primary software for the entire analysis. The dataset was uploaded into RStudio using the readxl library to import the data from an Excel file, and I will use several packages such as dplyr, ggplot2, and summarytools for data manipulation, visualization, and generating descriptive statistics. RStudio is a powerful tool that allows for flexible data analysis and visualization, and it is ideal for the objectives of this project.

Following the loading of the dataset, I performed the necessary steps to prepare it for analysis. This included converting selected columns into factors, specifically binary factors (No, Yes), which are relevant for the analysis of categorical variables. Below is the code used for this step:

After that, I examined the output using the str() function, which reveals the structure of the dataset, as shown below:

```
> str(dataset)
tibble [65,535 × 19] (S3: tbl_df/tbl/data.frame)
 $ Chest_Pain         : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 2 2 2 1 ...
 $ Shortness_of_Breath: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 2 2 1 ...
 $ Fatigue            : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
 $ Palpitations       : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 2 2 2 1 ...
 $ Dizziness          : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 1 2 1 ...
 $ Swelling           : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 2 2 1 ...
 $ Pain_Arms_Jaw_Back : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 2 ...
 $ Cold_Sweats_Nausea : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 2 2 1 ...
 $ High_BP            : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 2 1 2 2 2 ...
 $ High_Cholesterol   : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 1 2 2 1 ...
 $ Diabetes           : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 2 2 2 1 1 ...
 $ Smoking            : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 2 2 2 1 ...
 $ Obesity            : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 2 2 1 2 ...
 $ Sedentary_Lifestyle: Factor w/ 2 levels "No","Yes": 2 1 2 2 1 2 2 2 1 1 ...
 $ Family_History     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 1 ...
 $ Chronic_Stress     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 2 1 1 ...
 $ Gender             : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 2 1 2 2 ...
 $ Age                : num [1:65535] 48 46 66 60 69 55 51 67 71 65 ...
 $ Heart_Risk         : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 2 2 1 ...
```

The output indicates that the dataset consists of 65,535 rows and 19 columns, as indicated previously. The columns include a combination of binary factors, such as "Chest Pain", "Shortness of Breath" or "Fatigue", and continuous variables, being “Age”. The factor columns are appropriately labeled with two levels: "No" and "Yes". This structure suggests that the dataset is largely focused on binary outcomes, which will be useful for analyzing health-related symptoms and conditions.

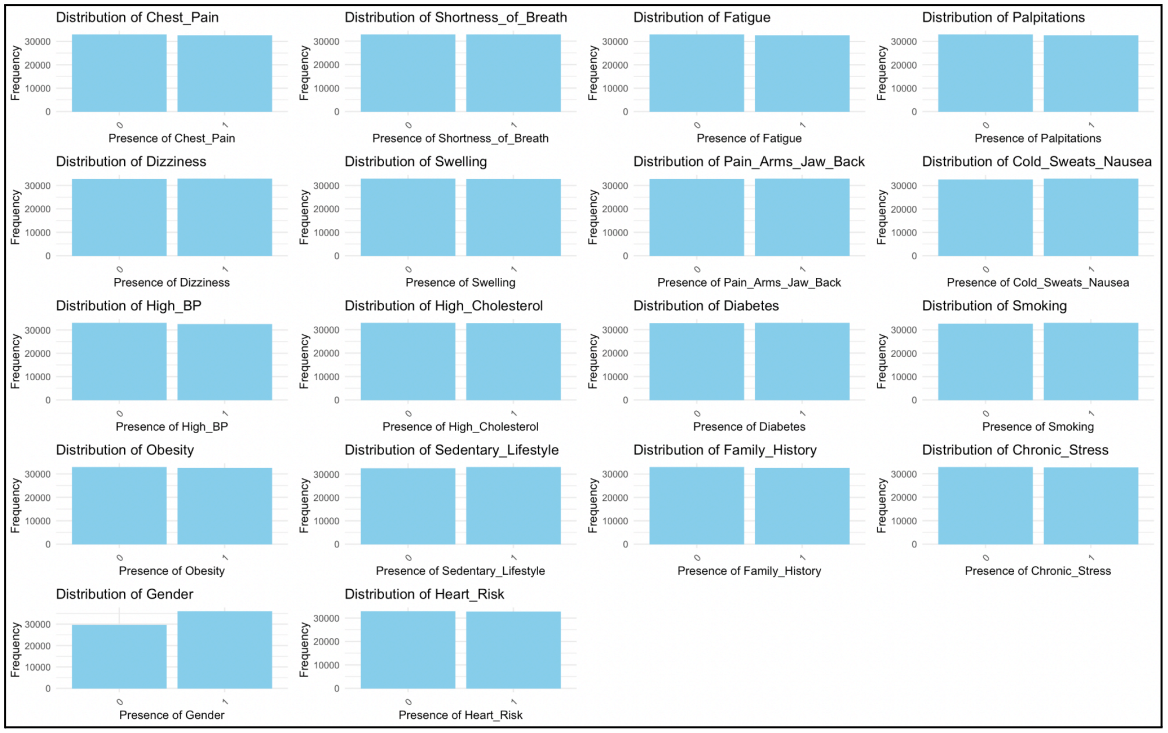
Additionally, the summary() function provides a detailed overview of the distribution of values across these variables:

```
> summary(dataset)
 Chest_Pain  Shortness_of_Breath Fatigue     Palpitations Dizziness
 No :32893   No :32773           No :32936   No :32878    No :32714
 Yes:32642   Yes:32762           Yes:32599   Yes:32657    Yes:32821



 Swelling    Pain_Arms_Jaw_Back Cold_Sweats_Nausea High_BP
 No :32847   No :32701          No :32630          No :32954
 Yes:32688   Yes:32834          Yes:32905          Yes:32581



 High_Cholesterol Diabetes    Smoking     Obesity     Sedentary_Lifestyle
 No :32845        No :32695   No :32610   No :32865   No :32569
 Yes:32690        Yes:32840   Yes:32925   Yes:32670   Yes:32966



 Family_History Chronic_Stress Gender           Age        Heart_Risk
 No :32926      No :32784      No :29595   Min.   :20.00   No :32834
 Yes:32609      Yes:32751      Yes:35940   1st Qu.:45.00   Yes:32701
                                           Median :56.00
                                           Mean   :54.43
                                           3rd Qu.:67.00
                                           Max.   :84.00
```

From the summary, we can observe that the majority of the variables have a roughly equal distribution between "No" and "Yes" responses, though some variables, like "Chest Pain" and "Shortness of Breath", show a slightly higher count for "No". For continuous variables like "Age", the summary provides insights into the range and central tendency (mean, median, quartiles), which helps us understand the distribution of age within the dataset.

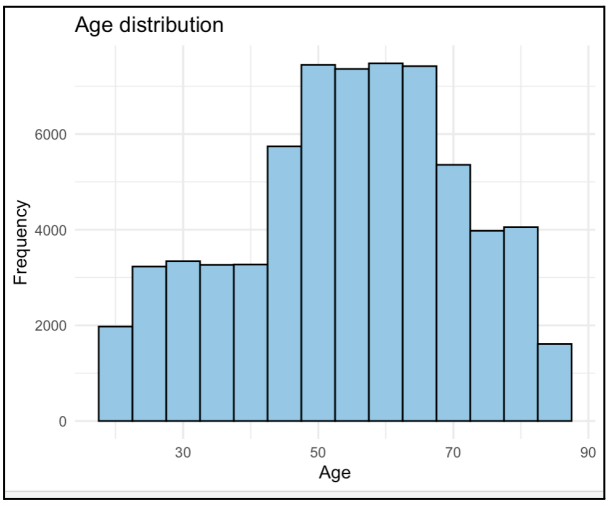
Univariate descriptive analysis

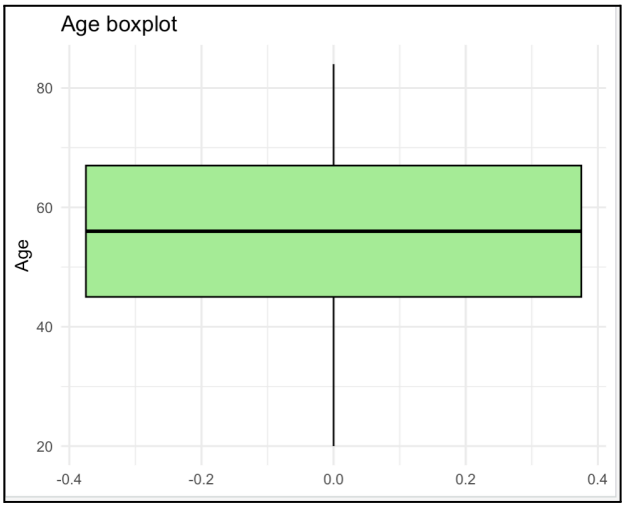
As mentioned previously, the dataset used for this analysis is synthetic, created with artificial intelligence for machine learning purposes. As a result, the frequencies of the binary variables are almost equally distributed, with approximately 50% of the entries labeled as "Yes" and 50% labeled as "No." This balance is a characteristic of the synthetic nature of the dataset, which was designed to ensure an equal representation of both categories. The equal distribution of values in these binary variables makes the dataset particularly useful for training machine learning models, as it avoids bias toward one category over the other.



In addition to the binary variables, the dataset also includes one continuous variable: Age. The descriptive statistics for the variable "Age" are as follows: the minimum value is 20 years, and the maximum value is 84 years. The first quartile of the dataset is 45 years, while the median age is 56 years. The mean age is approximately 54.43 years, and the third quartile is 67 years. This shows that the dataset contains individuals from a wide range of ages, with the majority falling between 45 and 67 years.

The standard deviation for the "Age" variable is 16.40, which indicates that there is a relatively wide spread of ages in the dataset. Although the mean age is around 54 years, the relatively high standard deviation suggests that some individuals are significantly older or younger than this average. This variability in age could be an important factor in machine learning models, as age may play a key role in predicting health risks or other factors related to the dataset.

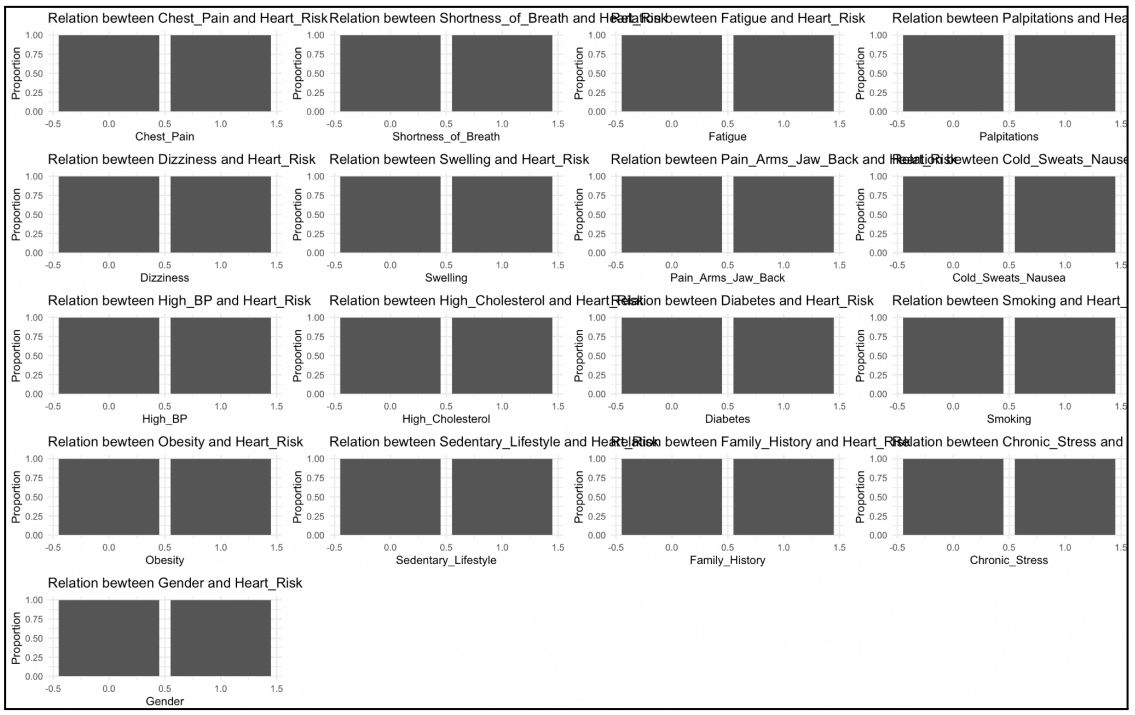




Bivariate descriptive analysis

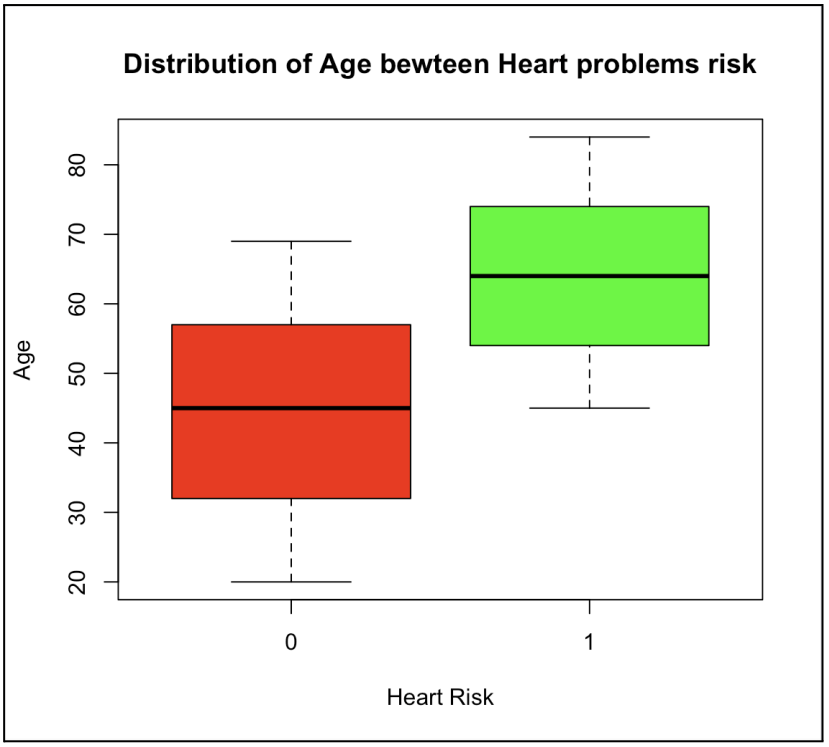
In the bivariate analysis, we examine the relationship between the binary variables and the target variable. Each binary variable has been plotted against the target variable, with the intention of identifying potential patterns or correlations. However, upon reviewing the relationships between the binary variables and the target variable, it becomes clear that there is not a strong or obvious association between them.

This lack of strong correlation can be attributed to the synthetic nature of the dataset. Since the binary variables were artificially generated for machine learning purposes, they may not reflect real-world relationships or conditions. Synthetic data often lacks the complexities and nuances of real data, which may explain the absence of significant associations between these variables and the target variable. As a result, while these variables may be useful in model training, their relevance to the target variable in this specific context is limited.



However, when analyzing the continuous variable Age, the bivariate relationship with the target variable presents a more promising insight. The boxplot of age against the target variable reveals a noticeable pattern: individuals with lower ages tend to have a lower risk while those with higher ages show a higher prevalence of risk.

This suggests that age may play an important role in predicting the likelihood of the condition in the target variable. The boxplot visually demonstrates that as age increases, the proportion of individuals with the condition (represented by the target variable) also tends to increase. This relationship could indicate that age is a significant factor in assessing health risk, and may therefore be an important predictor in machine learning models aimed at forecasting health outcomes.


Distribution of Age bewteen Heart problems risk
Age
20
30
40
50
60
70
80
0
1
Heart Risk

LOGISTIC REGRESSION

Logistic regression is a widely used statistical method for modeling the relationship between a categorical dependent variable and one or more independent variables. Unlike standard linear regression, which assumes a continuous outcome, logistic regression is specifically designed to analyze qualitative outcome variables. The primary objective of this analysis is to understand the factors that contribute to the likelihood of an individual being classified as high-risk for cardiovascular disease. By estimating the effects of predictor variables on the dependent variable, this model identifies which risk factors have the most significant impact on heart disease risk. Furthermore, logistic regression enables us to make probabilistic predictions and assess the model's accuracy by comparing predicted classifications with actual outcomes. This approach provides valuable insights for early detection and prevention of cardiovascular disease, ultimately aiding in more effective risk assessment and medical decision-making.

In this study, the target variable selected for the logistic regression model is Heart_Risk, as the primary objective is to predict an individual's likelihood of developing heart disease. This binary variable indicates whether a patient is classified as high-risk or not, allowing for a clear distinction between those who may require early intervention and those at lower risk. By using Heart_Risk as the dependent variable, the model can analyze the influence of various risk factors, such as smoking, high blood pressure, diabetes, and lifestyle habits, on the probability of developing cardiovascular disease. This predictive approach not only helps in identifying key determinants of heart disease but also supports healthcare professionals in implementing targeted prevention strategies.

The results of the model indicate that all the predictors have a statistically significant impact on heart disease risk, as shown by their p-values being < 2e-16, which confirm strong associations with the target variable. Furthermore, the Akaike Information Criterion (AIC) value of 2219.5 provides a measure of model quality, with lower values indicating better fit. These results suggest that the logistic regression model is effective in identifying key risk factors and making accurate predictions for heart disease risk.

```
Call:
glm(formula = Heart_Risk ~ ., family = binomial, data = trainData)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -25.347729   0.642809  -39.43   <2e-16 ***
Chest_Pain          2.819284   0.126646   22.26   <2e-16 ***
Shortness_of_Breath 2.694002   0.124389   21.66   <2e-16 ***
Fatigue             2.827752   0.127051   22.26   <2e-16 ***
Palpitations        2.686339   0.124262   21.62   <2e-16 ***
Dizziness           2.788672   0.126972   21.96   <2e-16 ***
Swelling            2.749614   0.125618   21.89   <2e-16 ***
Pain_Arms_Jaw_Back  2.790098   0.125804   22.18   <2e-16 ***
Cold_Sweats_Nausea  2.746303   0.125262   21.92   <2e-16 ***
High_BP             1.676416   0.119789   13.99   <2e-16 ***
High_Cholesterol    1.674086   0.119223   14.04   <2e-16 ***
Diabetes            1.614912   0.119183   13.55   <2e-16 ***
Smoking             1.653288   0.119249   13.86   <2e-16 ***
Obesity             1.666866   0.118896   14.02   <2e-16 ***
Sedentary_Lifestyle 1.733484   0.119480   14.51   <2e-16 ***
Family_History      1.685709   0.119963   14.05   <2e-16 ***
Chronic_Stress      1.660653   0.119505   13.90   <2e-16 ***
Gender              1.298884   0.118534   10.96   <2e-16 ***
Age                 0.125170   0.005741   21.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63596.1  on 45874  degrees of freedom
Residual deviance:  2181.5  on 45856  degrees of freedom
AIC: 2219.5

Number of Fisher Scoring iterations: 10
```

The signs of the estimated coefficients provide insight into the direction of the relationship between each predictor and the probability of having heart disease. A positive estimate means that as the predictor increases the probability of heart disease increases. Conversely, a negative estimate would indicate that the predictor is associated with a decrease in the probability of heart disease. In this model, all predictors have positive estimates, meaning that the presence of these symptoms or risk factors increases the likelihood of heart disease. However, while the sign tells us the direction of the effect, it does not tell us the magnitude of the effect.

To quantify the impact of each predictor, I will use the odds ratios. The odds ratio represents the multiplicative change in the odds of heart disease occurring when the predictor variable changes. For example, the odds ratio for Chest_Pain is 16.76, meaning that individuals experiencing chest pain are 16.76 times more likely to have heart disease compared to those who do not. Similarly, symptoms like Fatigue (16.90), Dizziness (16.25), and Pain_Arms_Jaw_Back (16.28) are also strongly associated with an increased risk.

Other significant risk factors include High Blood Pressure (5.35), High Cholesterol (5.33), Diabetes (5.03), Smoking (5.22), and Obesity (5.30). This suggests that individuals with these conditions have approximately 5 times higher odds of developing heart disease. The odds ratio for Age (1.13) indicates that for every one-year increase in age, the odds of heart disease increase by 13%.

The confidence intervals (CI) provide additional insights into the reliability of these estimates. For all variables in this model, the confidence intervals indicate a strong association with heart disease risk, reinforcing the validity of the findings.

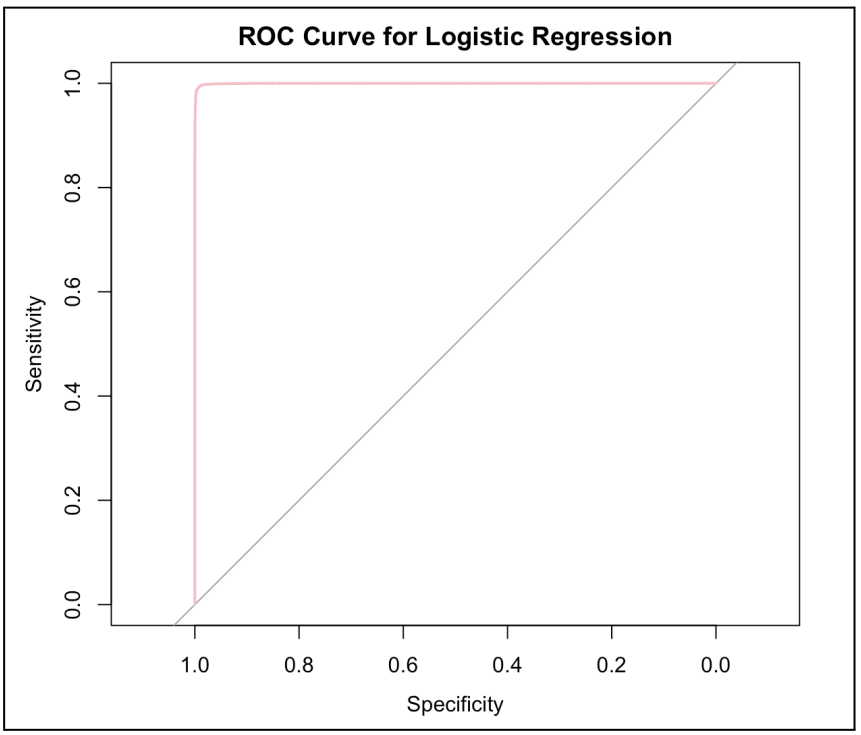
	Variable	Odds_Ratio	CI_Lower	CI_Upper
(Intercept)	(Intercept)	9.808916e-12	2.674626e-12	3.328185e-11
Chest_Pain	Chest_Pain	1.676485e+01	1.312912e+01	2.157573e+01
Shortness_of_Breath	Shortness_of_Breath	1.479075e+01	1.163056e+01	1.894447e+01
Fatigue	Fatigue	1.690741e+01	1.323104e+01	2.177777e+01
Palpitations	Palpitations	1.467784e+01	1.154447e+01	1.879487e+01
Dizziness	Dizziness	1.625942e+01	1.272548e+01	2.093917e+01
Swelling	Swelling	1.563660e+01	1.226835e+01	2.007987e+01
Pain_Arms_Jaw_Back	Pain_Arms_Jaw_Back	1.628261e+01	1.277116e+01	2.091822e+01
Cold_Sweats_Nausea	Cold_Sweats_Nausea	1.558491e+01	1.223575e+01	1.999866e+01
High_BP	High_BP	5.346359e+00	4.236855e+00	6.777805e+00
High_Cholesterol	High_Cholesterol	5.333916e+00	4.231323e+00	6.753919e+00
Diabetes	Diabetes	5.027445e+00	3.988508e+00	6.365350e+00
Smoking	Smoking	5.224129e+00	4.144131e+00	6.615395e+00
Obesity	Obesity	5.295546e+00	4.203541e+00	6.700964e+00
Sedentary_Lifestyle	Sedentary_Lifestyle	5.660341e+00	4.488535e+00	7.171681e+00
Family_History	Family_History	5.396276e+00	4.274989e+00	6.843453e+00
Chronic_Stress	Chronic_Stress	5.262748e+00	4.172668e+00	6.667668e+00
Gender	Gender	3.665204e+00	2.910349e+00	4.632858e+00
Age	Age	1.133341e+00	1.120939e+00	1.146465e+00

This logistic regression model's performance is evaluated using different accuracy measures. First, the confusion matrix provides a breakdown of the predictions compared to the actual classifications. The model correctly predicted 9767 cases as no heart risk (true negatives) and 9741 cases as heart risk (true positives). However, it misclassified 83 cases as having heart risk when they did not (false positives) and 69 cases as not having heart risk when they did (false negatives).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9767   69
         1   83 9741

               Accuracy : 0.9923
                 95% CI : (0.9909, 0.9934)
    No Information Rate : 0.501
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9845

 Mcnemar's Test P-Value : 0.2917

            Sensitivity : 0.9916
            Specificity : 0.9930
         Pos Pred Value : 0.9930
         Neg Pred Value : 0.9916
             Prevalence : 0.5010
         Detection Rate : 0.4968
   Detection Prevalence : 0.5003
      Balanced Accuracy : 0.9923

       'Positive' Class : 0
```

From these values, the accuracy of the model is calculated as 99.23%, meaning the model correctly classifies the vast majority of cases. The sensitivity (99.16%) indicates that the model correctly identifies most actual positive cases, while the specificity (99.30%) shows that it accurately detects negative cases.

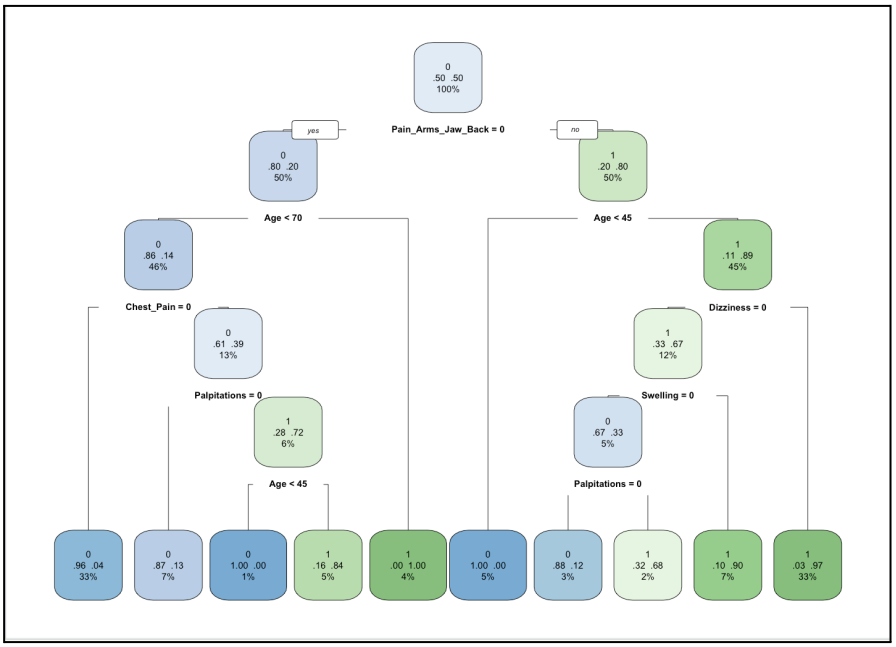
To further assess the model's effectiveness, the Area Under the Curve (AUC) is examined. The model achieves an AUC of 0.9996, meaning it is almost perfect in distinguishing between individuals with and without heart risk. A value close to 1.0 confirms the high predictive power of the logistic regression model.



DECISION TREE

For the decision tree, the target variable used to predict heart risk is also Heart_Risk.

The decision tree was constructed and the following tree diagram illustrates the decision-making process. In the tree, the variables are split at various decision points based on their importance and contribution to predicting heart risk. The leaf nodes of the tree represent the final decisions (predictions) made for each observation. Each segment, or leaf node, corresponds to a subgroup of observations that are classified in the same category based on the values of the predictors. For instance, some leaf nodes may indicate a higher likelihood of heart risk, while others suggest lower likelihood. Below is the decision tree diagram:



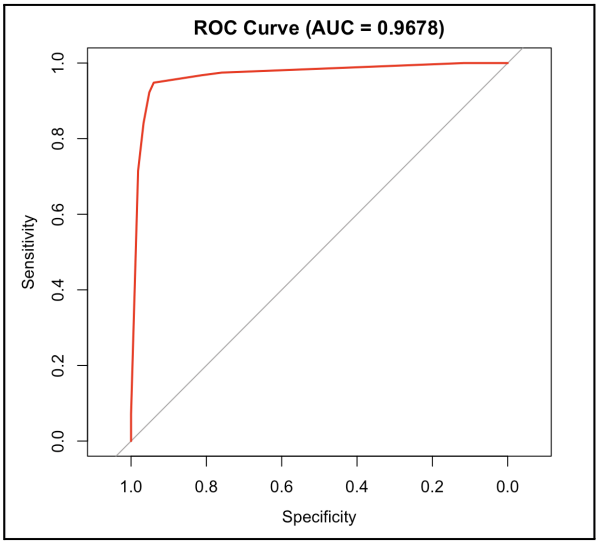
Each path through the tree leads to a specific outcome based on the splits on the predictor variables. By following these decision paths, we can classify individuals into high-risk or low-risk categories.

To evaluate the performance of the decision tree model, a ¡confusion matrix was implemented which provides important metrics for classification models, including accuracy. The confusion matrix below shows the comparison between the predicted and actual values for the target variable Heart_Risk:

```
          Reference
Prediction    0    1
         0 9204  518
         1  646 9292

               Accuracy : 0.9408
                 95% CI : (0.9374, 0.9441)
    No Information Rate : 0.501
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8816

 Mcnemar's Test P-Value : 0.0001973

            Sensitivity : 0.9344
            Specificity : 0.9472
         Pos Pred Value : 0.9467
         Neg Pred Value : 0.9350
             Prevalence : 0.5010
         Detection Rate : 0.4682
   Detection Prevalence : 0.4945
      Balanced Accuracy : 0.9408

       'Positive' Class : 0
```

The overall accuracy of the model is 94.08%, meaning the model correctly classified the risk level of 94.08% of the cases. With a 95% confidence interval will result in (93.74%, 94.41%).

An Area Under the Curve (AUC) value of 0.9678 indicates excellent model performance. This high AUC suggests that the decision tree model has a strong ability to correctly identify both the presence and absence of heart disease, making it a reliable model for predicting heart disease risk. In general, an AUC above 0.9 is considered to indicate very good model performance, and the result here shows that the model is performing well above expectations.



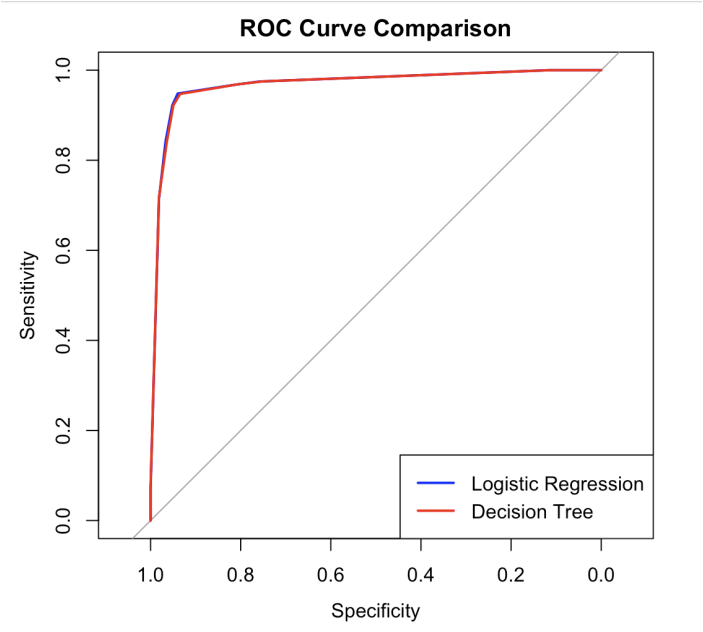
MODEL COMPARISON

The purpose of this model comparison is to evaluate which model performs better for predicting heart disease risk. To achieve this, we compare both the Logistic Regression and the Decision Tree models using two key metrics: Accuracy and Area Under the Curve (AUC).

Accuracy is a direct measure of how well the model predicts the correct outcome. The Logistic Regression model achieved an accuracy of 0.9923, while the Decision Tree model achieved an accuracy of 0.9408. These values suggest that Logistic Regression performs slightly better in terms of overall prediction accuracy.

For a more nuanced evaluation, we also examine the AUC, which assesses the model's ability to distinguish between classes. The Logistic Regression model produced an AUC of 0.9678, and the Decision Tree model generated an AUC of 0.9673. Although the AUC values for both models are very similar, with just a slight edge for Logistic Regression, both models show a strong ability to distinguish between the two classes, with AUC values approaching 1.

To visually compare the performance of both models, the following graphs illustrate the AUC curves for Logistic Regression and the Decision Tree, further confirming that both models perform well in classification tasks.



CONCLUSION

In conclusion, this project has provided a comprehensive analysis of the prediction of heart disease risk using both Logistic Regression and the Decision Tree model. The dataset was thoroughly examined and prepared, and the models were built and evaluated based on key metrics such as accuracy and AUC.

The Logistic Regression model was presented as the top performer in terms of accuracy, achieving a figure of 99.23% and an AUC of 0.9678. However, the Decision Tree model, while slightly lower in accuracy at 94.08%, still demonstrated strong performance with an AUC of 0.9673, making it a reliable model for classification tasks. Both models showed excellent discrimination ability, as evidenced by their high scores.

Overall, both models are highly capable for predicting heart disease risk, with Logistic Regression showing a slight advantage in overall accuracy and Decision Tree offering a visually interpretable classification approach. Depending on the specific application needs, either model could be suitable, but Logistic Regression appears to be marginally better for this particular dataset.

REFERENCES

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). *Can machine-learning improve cardiovascular risk prediction using routine clinical data?* PLOS ONE, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

World Health Organization. (2024, May 15). *Cardiovascular diseases kill 10,000 people in the WHO European Region every day, with men dying more frequently than women.* https://www.who.int/europe/news-room/15-05-2024-cardiovascular-diseases-kill-10-000-people-in-the-who-european-region-every-day--with-men-dying-more-frequently-than-women

World Health Organization. (2024, May 15). *Cardiovascular diseases kill 10,000 people in the WHO European Region every day, with men dying more frequently than women.* https://www.who.int/europe/news-room/15-05-2024-cardiovascular-diseases-kill-10-000-people-in-the-who-european-region-every-day--with-men-dying-more-frequently-than-women