CARDIOVASCULAR DISEASE PREDICTION

Author(s): Anyawuike, Ikenna Omedobi
Group: Zidio Development.
Date: 21/03/2024

Abstract:

Cardiovascular diseases (CVDs) are a major cause of morbidity and mortality worldwide, emphasizing the importance of early detection and prevention. This research explores the development of predictive models for CVDs using machine learning techniques. The study utilizes a dataset containing clinical and demographic attributes related to cardiovascular health, which undergoes comprehensive data cleaning, preprocessing, and exploratory data analysis (EDA). Three classification models—Random Forest Classifier, XGBoost Classifier, and Decision Tree Classifier—are evaluated for their ability to predict cardiovascular disease based on patient attributes. Hyperparameter tuning techniques are employed to optimize model performance. Model evaluation metrics, including accuracy, precision, recall, and F1 score, are utilized to assess the predictive capabilities of each model. The results highlight the efficacy of machine learning models in predicting CVDs, with the Random Forest Classifier exhibiting the highest performance. These predictive models have the potential to aid healthcare professionals in early detection and intervention for individuals at risk of cardiovascular disease, thereby improving patient outcomes and reducing mortality rates.

Introduction:

Cardiovascular diseases (CVD) are among the leading causes of death globally. Early detection and prediction of CVD play a crucial role in preventing adverse health outcomes. This study aims to develop predictive models for cardiovascular disease using machine learning techniques. The dataset used in this study contains various clinical and demographic features, allowing for the development of robust prediction models.

Methodology:

Data collection: Data collection is the process of gathering and measuring information on targeted variables in an established system which then enables one to answer relevant questions and evaluate outcomes.
Furthermore, the data set was provided by Zidio Development. The dataset comprises clinical and demographic attributes related to cardiovascular health. Attributes include age, gender, blood pressure, cholesterol levels, exercise-induced angina, maximum heart rate, chest pain type, and more.

Data Wrangling and Preprocessing: Data wrangling is a crucial step that cleanses and organizes messy or complex datasets for easier access and visualization during exploratory data analysis (EDA). Below are the steps observed.

- Column Renaming for Consistency: To enhance clarity and facilitate analysis, the remaining columns were renamed using a consistent mapping scheme. This involved; `heart.rename(columns={'target':'class','cp':'Chp','trestbps':'Bp','chol':'Sch','restecg':'Ecg','thalach':'Mhrt','exang':'Exain','oldpeak':'Opk','thal':'Thal','ca':'Vessel'},inplace=True)`.

- Missing Value Analysis: A comprehensive analysis of missing values is essential to identify and address any potential data quality issues. The `isnull().sum()` method was employed to identify the extent of missing attribute data in the dataset. Missing values were then carefully examined to understand the underlying causes and potential impact on the analysis. To address missing values and ensure robust analysis, a data imputation technique was implemented. In this case, missing values in the dataset were filled with the mean value for that specific attribute. This imputation strategy assumes that missing values are randomly distributed around the mean and avoids introducing undue bias.

- Redundant data filtering: We filtered out redundant rows from the dataset based on a specific criterion. Initially, a boolean mask is created based on a condition that checks whether the value in the 'class' column is equal to 2. This creates a mask where rows meeting this criterion are marked as True, while others are marked as False. Using this boolean mask, rows from the dataset (heart) are selected where the condition evaluates to True. Essentially, this step filters the dataset to retain only those rows where the 'class' column contains a value of 2.
Finally, the `.drop()` method is applied to the original dataset (heart) with the index of the rows identified in the previous step. This effectively removes all rows where the 'class' column equals 2, leaving behind a cleaned dataset without redundant rows.

- Addressing Outliers: Initially, we calculated the first quartile (Q1), third quartile (Q3), and interquartile range (IQR) for each column in the dataset (heart). This was done using the `quantile()` function, which calculates the value below which a given percentage of observations fall. With the quartile values in hand, they defined a threshold value (typically 1.5 times the IQR) to identify outliers. It then checks each data point in the dataset against this threshold to determine if it lies outside the acceptable range. The `np.where()` function was used to identify the indices of rows containing outliers based on the defined threshold. This function returns the indices where the condition evaluates to True.
Finally, the identified indices are used to index the dataset (heart) and drop the corresponding rows containing outliers. This operation effectively removes the outliers from the dataset, leaving behind a cleaned dataset without the problematic data points.

Exploratory Data Analysis (EDA)

EDA techniques were employed to gain insights into the patient demographics, distribution of clinical indicators, and their relationships with cardiovascular disease. Visualizations such as pie charts, histograms, descriptive statistics, correlation heatmaps, and violin plots were utilized to explore the dataset and identify potential patterns.

1. Patient Demographics:

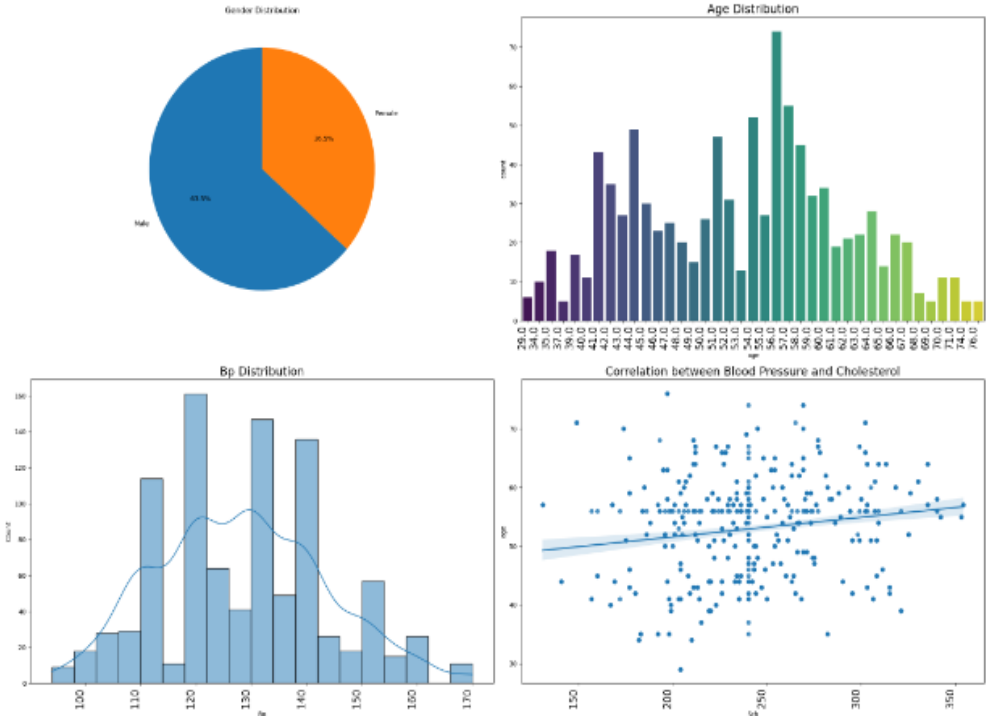
- Gender Distribution: The pie chart displayed in the top-left subplot illustrates the distribution of

genders in the dataset. Each slice of the pie represents the proportion of males and females, with corresponding percentage values displayed. This visualization helps understand the gender balance within the dataset.

- Age Distribution: The top-right subplot showcases the distribution of patient ages using a count plot. Each bar represents the count of patients falling within an age group. The x-axis labels denote the age groups, and the bars' heights depict the corresponding frequency. This visualization offers insights into the age distribution of the patients in the dataset.

- Blood Pressure (Bp) Distribution: The bottom-left subplot presents the distribution of blood pressure (Bp) values among the patients. It utilizes a histogram overlaid with a kernel density estimation (KDE) curve to represent the Bp distribution. This visualization aids in understanding the spread and central tendency of blood pressure readings in the dataset.

- Correlation between Blood Pressure and Cholesterol: The bottom-right subplot demonstrates the correlation between blood pressure (x-axis) and cholesterol levels (y-axis). It employs a scatter plot overlaid with a regression line to depict the relationship between these two variables. This visualization helps assess the correlation strength and direction between blood pressure and cholesterol, providing insights into potential relationships between these health indicators.



2. Distribution of Patient age and selected indicators:

- Gender and Heart Disease: The first subplot displays a count plot showing the distribution of heart disease among different genders. The x-axis represents the genders (male and female), while the y-axis represents the count of individuals. The bars are further differentiated by color

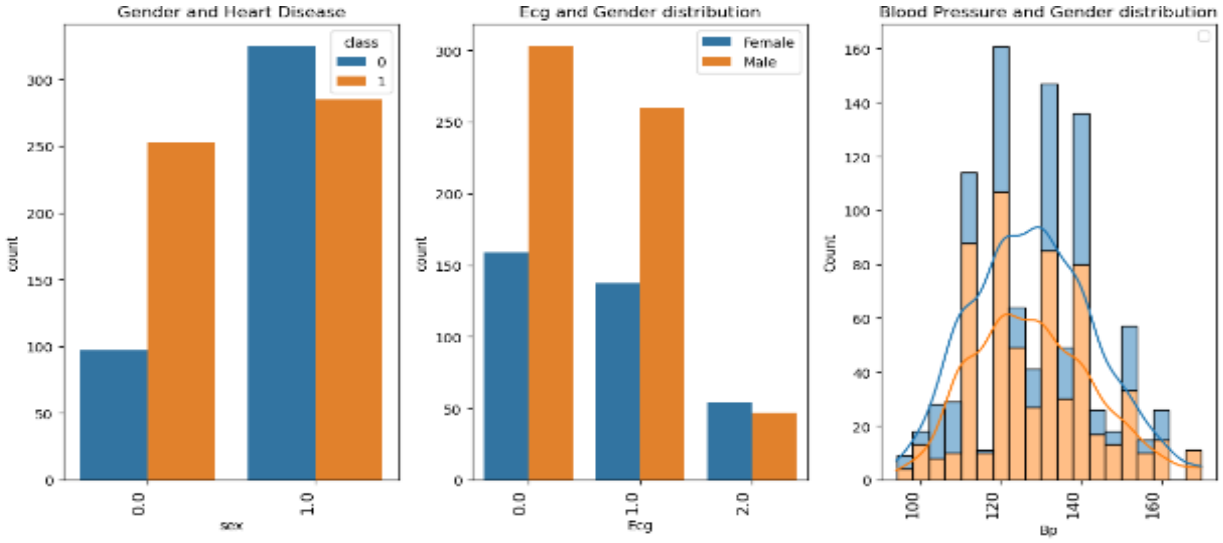
to indicate the presence or absence of heart disease. This visualization helps us understand how heart disease prevalence varies across genders.

- Ecg and Gender Distribution: The second subplot presents another count plot, this time showing the distribution of electrocardiogram (ECG) results categorized by gender. The x-axis represents the ECG categories, while the y-axis denotes the count of individuals. The bars are grouped by gender, with different colors representing males and females. This visualization aids in exploring the distribution of ECG results across genders.

- Blood Pressure (Bp) and Gender Distribution: The third subplot utilizes a histogram with KDE (kernel density estimation) overlaid to visualize the distribution of blood pressure (Bp) among different genders. The x-axis represents the blood pressure values, while the y-axis indicates the frequency of occurrence. The histogram bars are stacked and differentiated by gender, allowing for a comparison of blood pressure distributions between males and females.

3. Distribution of Exercise Induce Angina to Heart Disease and Age:

- Exercise-Induced Angina and Heart Disease: The first subplot, represented by a count plot, illustrates the distribution of heart disease based on the presence or absence of exercise-induced angina. The x-axis represents the different categories of exercise-induced angina, while the y-

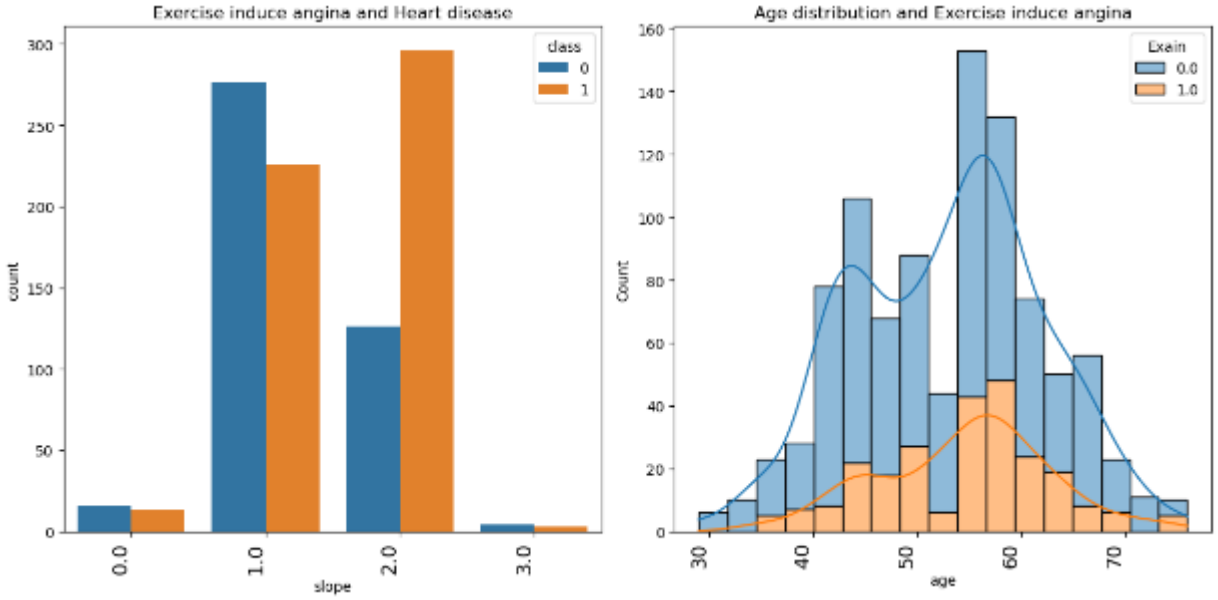


axis indicates the count of individuals. The bars are grouped by the presence or absence of heart disease, distinguished by different colors. This visualization helps analyze the relationship between exercise-induced angina and the likelihood of heart disease.

- Age Distribution and Exercise-Induced Angina: The second subplot displays a histogram overlaid with kernel density estimation (KDE), depicting the distribution of age among individuals categorized by the presence or absence of exercise-induced angina. The x-axis represents age values, while the y-axis indicates the frequency of occurrence. The histogram bars are stacked and differentiated by the presence or absence of exercise-induced angina, facilitating the comparison of age distributions between these two groups. This visualization aids in

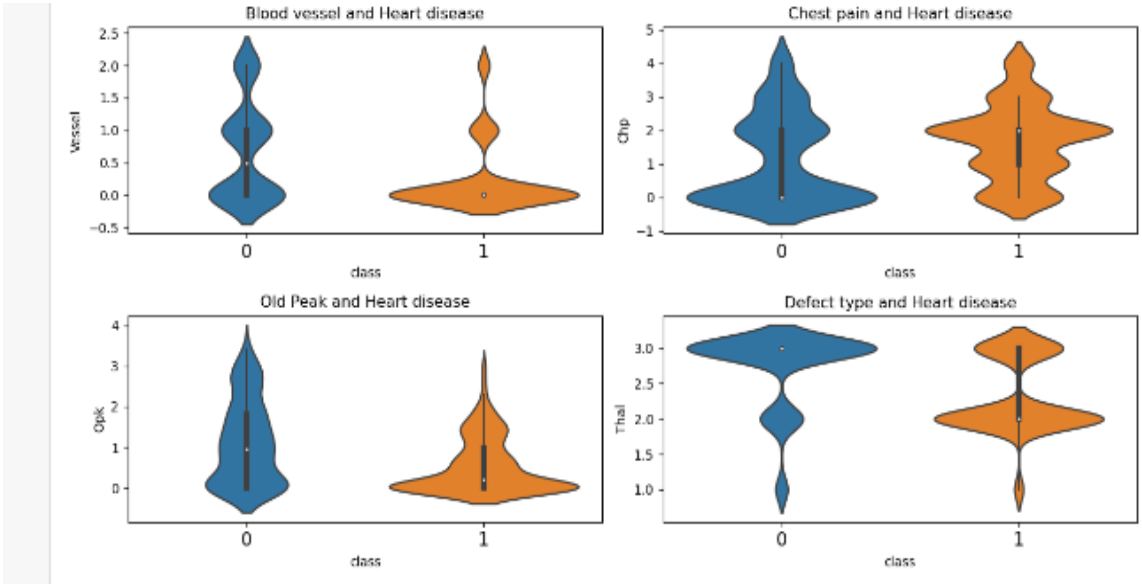
understanding how age relates to the occurrence of exercise-induced angina.

4. Relationship between selected Indicator to Heart Disease:
These violin plots offer a comprehensive visualization of how different indicators relate to heart



disease.

- Blood Vessel and Heart Disease: The first subplot depicts a violin plot with heart disease status (class) on the x-axis and the number of blood vessels (Vessel) on the y-axis. The width of the violin plot at each level of heart disease indicates the distribution of blood vessel counts,



providing insights into how this indicator varies with the presence or absence of heart disease.

- Chest Pain and Heart Disease: The second subplot displays a violin plot showing the relationship between chest pain (Chp) and heart disease status (class). Similar to the first subplot, the width of the violin plot represents the distribution of chest pain levels across different heart disease categories, aiding in understanding the association between chest pain and heart disease.

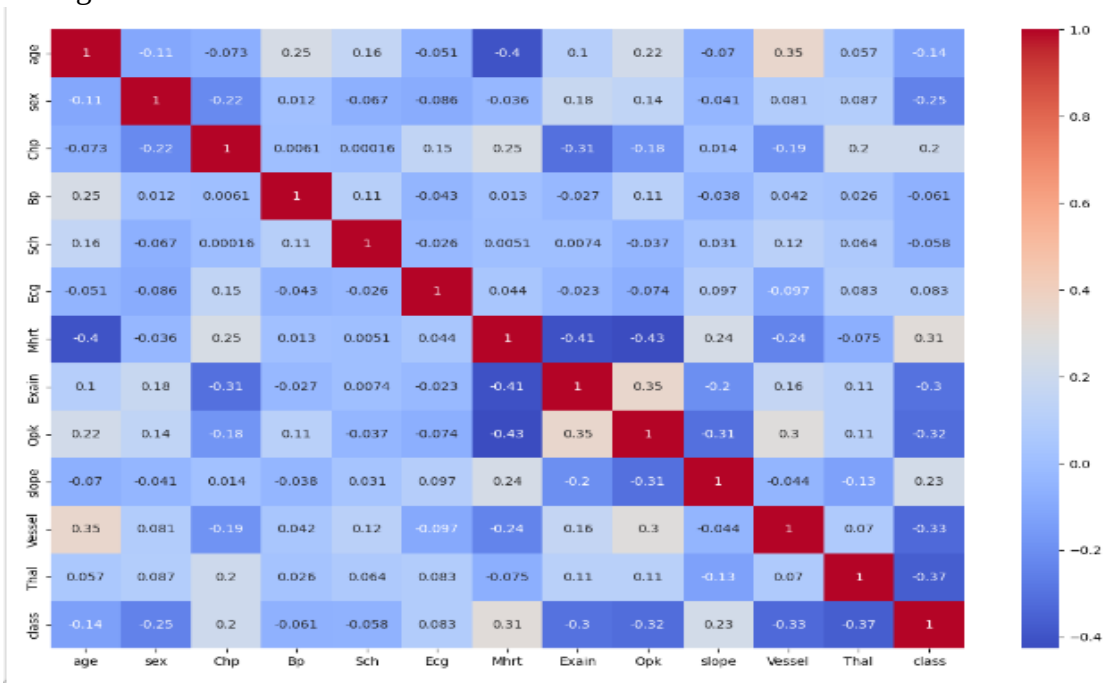
- Old Peak and Heart Disease: In the third subplot, a violin plot demonstrates the relationship between the magnitude of the ST depression induced by exercise relative to rest (Opk) and heart disease status (class). The varying widths of the violin plots illustrate the distribution of ST depression levels concerning the presence or absence of heart disease, offering insights into this indicator's relevance to heart disease diagnosis.

- Defect Type and Heart Disease: The fourth subplot presents a violin plot showcasing the relationship between the type of defect observed (Thal) and heart disease status (class). Similarly, in the previous plots, the width of the violin plots at each defect type level provides information about how defect types are distributed concerning heart disease presence or absence.

5. Correlation Heatmaps:
Correlation heatmaps are generated to visualize the pairwise correlations between attributes within the dataset (age, sex, Chp, Bp, Sch, Ecg, Mhrt, Exain, Opk, and more). These heatmaps help Identify statistically significant correlations between attributes stated formerly. It also gives insight into the direction (positive or negative) and strength of these relationships.

Predictive Analysis Methodology: The methodology outlined the machine learning algorithms, statistical methods, model evaluation metrics, cross-validation, and hyperparameter tuning techniques employed in the study.

- Building the model:



1. Feature Engineering: Feature Engineering: Feature engineering is a pivotal stage in machine learning aimed at enhancing model performance by selecting, transforming, or creating new features.
In this context, we initiated the data preparation process by implementing feature selection. This involved segregating the features (independent variables) from the target variables (dependent variables) within the dataset. Our Independent Variables encompass factors that could impact class, such as age, sex, Chp, Bp, Sch, Ecg, Mhrt, Exain, Opk, and more. While our dependent variables encompass the class.

2. Train-Test Split: We employed the train_test_split function from the sklearn.model_selection module to divide each dataset (independent and dependent variables) into training and testing sets. We set the split ratio to 80% for training and 20% for testing which was latterly adjusted to a 60:40 % ratio, with a random state of 42 to ensure reproducibility. The training set was utilized to train the machine learning model, while the testing set was employed to assess the model's performance on unseen data.

- Model Selection:

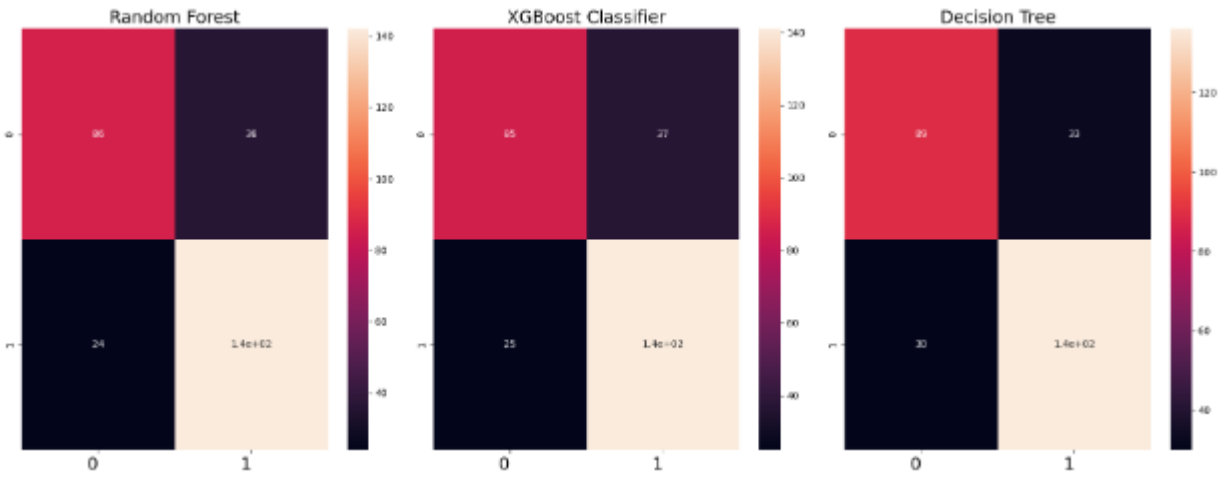
Three classification models were evaluated for predicting cardiovascular disease:
Random Forest Classifier, XGBoost Classifier, Decision Tree Classifier.

- Hyperparameter Tuning:

Hyperparameter tuning involves systematically searching for the optimal hyperparameters of a model to enhance its performance. We utilized GridSearchCV and BayesSearchCV techniques for hyperparameter tuning, allowing for an exhaustive search over a specified parameter grid.

- Model Evaluation:

The performance of each model was evaluated using metrics such as accuracy, precision, recall, and F1 score. Confusion matrices were utilized to visualize the classification results and assess model performance in predicting true positive and true negative cases.
- Results:



The predictive models achieved varying levels of performance:

Algorithm	Precision	Recall	F1 Score	Accuracy
RandomForest Classifier	0.797753	0.855422	0.825581	0.791667
XGB Classifier	0.792135	0.849398	0.819767	0.784722
Decision Tree	0.804734	0.819277	0.811940	0.781250

Conclusion:

The research demonstrates the effectiveness of machine learning models in predicting cardiovascular disease based on clinical and demographic attributes. The Random Forest Classifier exhibited the highest performance among the models evaluated. These predictive models can potentially assist healthcare professionals in early detection and intervention for individuals at risk of cardiovascular disease, thereby improving patient outcomes and reducing mortality rates.

References:

GitHub: https://github.com/Omedobi/Data-Science-Projects/tree/78de2a22ae63c2d12f84e935356425879a292ec2/Cardiovascular%20Disease%20Prediction