CardioRisk: A Machine Learning-Based Approach to Heart Disease Risk Prediction

Prepared by:

Omar Hussein Jankhot (202020109)

Ayham Emad Mohammad Ali (202110008)

Supervisor:

Dr. Mohammad Alhawarat

Graduate Project

Submitted to the Faculty of Information Technology / Middle East University as a partial fulfillment for BSc in Computer Science / Artificial Intelligence.

Department of Artificial Intelligence

Saturday 21st June, 2025

Declaration

إقرار ملكية

We hereby declare that the work presented in this report and the ideas based on it are our own, unless otherwise stated, properly cited in the text, and referenced at the end of the document.

Student ID	Student's Name	Signature	Date
202020109	Omar Hussein Jankhot		
202110008	Ayham Emad Mohammad Ali		

Supervisor Approval

موافقة المشرف

APPROVAL FOR SUBMISSION

We certify that the project report entitled **Predicting Heart Disease Risk Using Machine Learning Models**, prepared by **Ayham Emad Mohammad Ali, Omar Hussein Jankhot**, meets the required standard for submission as partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science / Artificial Intelligence at MEU.

Approved by:

Signature: _____

Supervisor: Dr._____

Date: _____

Acknowledgements

الشكر والعرفان

ACKNOWLEDGEMENTS

We sincerely appreciate everyone who played a role in the successful completion of this project, especially our supervisor, Dr. Mohammad Al Hawarat, whose invaluable guidance was instrumental throughout the process.

Abstract

This project aims to address the issue of cardiovascular diseases by providing an Artificial Intelligence (AI) based solution using machine learning models for the early prediction of heart disease to both decrease it's mortality rate and reduce the economical weight it puts on patients and hospitals.

In this project we made a combined dataset by joining UCI and Kaggle heart disease datasets, preprocessed and cleaned the dataset using techniques like random forest imputer, standardscaler and quantile transformer then trained 8 machine learning models on it (Logistic Regression, K-nearest neighbor (KNN), Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, and Extreme Gradient Boosting (XGBoost)) used hyperparameter tuning and cross validation to get the best possible results of each model and built a simple user interface to provide ease of use to medical professionals.

Our best model our best performing model was Extreme Gradient Boosting (XGBoost) with a Cross-Validation Accuracy score of: 88%, accuracy score of : 92%, f1 score of: 96% and roc-auc scored: 98%, which indicates that Artificial Intelligence (AI) can be used as a diagnosis tool or a decision support system for medical experts to assist in making a accurate and timely diagnosis.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. 6, 8–11, 16, 28, IV, V

AUC Area Under the Curve. 13

CNN Convolutional neural network. 6

CVD Cardiovascular Disease. 5

ECG Electrocardiogram. 6, 22, VI

KNN K-nearest neighbor. 9, 16, 27, IV, V

ML Machine Learning. 8

NIH National Institutes of Health. 5

ROC Receiver Operating Characteristic. 13

SVM Support Vector Machine. 6, 9, 16, 27, IV, V

UCI University of California, Irvine. 6, 7, 18, 20

WHO World Health Organization. 5

XGBoost Extreme Gradient Boosting. 6, 11, 16, 27–29, IV, V

Introduction

1.1 Problem Statement and Purpose

Heart disease is one of the leading causes of mortality around the world. Despite the improvement along the medical field and it's equipments, The early diagnosis of heart disease remains challenging because of the high variety of symptoms and how they overlap with many other medical conditions, these challenges have a high impact on the health of the patients and they have a heavy economical load on health sectors and hospitals, we plan to develop a machine learning based approach on the prediction of heart disease as a solution to both help in the early detection of heart disease and to lessen the economical load that these diseases have on both the health sector and the patients.

1.2 Project and Design Objectives

In this project we made a machine learning model for the prediction of heart disease, it starts by acquiring the right data from hospitals on heart patients then preprocessing the data using the right techniques, after having a fully developed and true to patient dataset we will start the process of building the machine learning models themselves using multiple machine learning models, and then comparing all the models using the appropriate evaluation methods for each of the models to see which model achieves the best results, and we will finish up the project as a package by developing a user interface to be able to easily enter the patients data.

1.3 Intended Outcomes

By the end of this project we expect to have a fully documented and step by step process on building a machine learning model on the prediction of heart disease, and to have a fully operational machine learning model with high and precise outcomes on the prediction of heart disease, and a comparison between multiple machine learning models and their ability to accurately predict the patient's status, which will greatly help to improve patient outcomes and contribute to health sectors.

1.4 Motivations

The motivation for this project comes from the severity of impact that cardiovascular diseases have on both patients and health sectors, heart diseases have a huge impact on their patients lives because by the time they start experiencing the symptoms they would be classified as severe cases in example we have the case of heart attacks which carry little to no symptoms before the heart attack event itself occurs which puts it's patient in a severely dangerous state, on the other side of the problem we have hospitals and health sectors that suffer a huge a huge economical toll, based on studies like [3] it was shown that the expected economical toll on just the united states health sector from cardiovascular diseases would exceed 1$ trillion dollars, our motivation is to help both medical professionals and patients on the mission of carrying out early diagnosis's of heart diseases and to be able to do a heart disease risk assessment before seeing the symptoms, and to decrease the economical toll of the heart disease diagnosis process and the economic toll of treating severe heart disease cases that would have been prevented by early detection.

1.5 Outline of Report

This report is structured in a step by step manner that will guide the reader through the processes of making a machine learning model for the prediction of heart disease at first we will address the subject of data collection and data analysis, then we will show the process of preprocessing the dataset then building multiple machine learning models while using measures like hyper parameter tuning to achieve the best results for each model and cross validation to make sure that the models are in their best shape, then we will show insights on which is the best performing model by evaluating the models and comparing them all using the appropriate evaluation method, after that we will build a simple user interface for the practicality of using the models for prediction on new patients.

Literature Review

The problem of cardiovascular disease has been brought to light in the past years by many studies that offer solutions to the problem that it develops for the health of it's patients and the effect it has on the health sector we will start by looking at studies on the issue of cardiovascular disease .

2.1 Cardiovascular Diseases

As stated by (World Health Organization [WHO], 2021)[4] An estimated 17.9 million people died from Cardiovascular Disease(CVD) in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke and It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin leading to the decrease of the severity of cases.As mentioned by (Gaziano, Thomas and Reddy)[3] for the cases of Middle East and North Africa the increase economic wealth in the Middle East and North Africa has been characteristically accompanied by urbanization. The rate of CVD has been increasing rapidly and is now the leading cause of death, accounting for 25 to 45 percent of total deaths. There are many different types of CVDs but As stated by (National Institutes of Health [NIH], 2022)[5] four of the main types are (Coronary heart disease) which occurs when the flow of oxygen-rich blood to the heart muscle is blocked or reduced,(Strokes and TIAs)A stroke is where the blood supply to part of the brain is cut off, which can cause brain damage and possibly death. A transient ischaemic attack (also called a TIA or "mini-stroke") is similar, but the blood flow to the brain is only temporarily disrupted, (Peripheral arterial disease) which occurs when there's a blockage in the arteries to the limbs, usually the legs, (Aortic disease)which are a group of conditions affecting the

aorta. This is the largest blood vessel in the body, which carries blood from the heart to the rest of the body. It is mentioned it the article (Forecasting the Economic Burden of Cardiovascular Disease and Stroke in the United States Through 2050: A Presidential Advisory From the American Heart Association)[6] One in 3 US adults received care for a cardiovascular risk factor or condition in 2020. Annual inflation-adjusted (2022 US dollars) health care costs of cardiovascular risk factors are projected to triple between 2020 and 2050, from $400 billion to $1344 billion. For cardiovascular conditions, annual health care costs are projected to almost quadruple, from $393 billion to $1490 billion, and productivity losses are projected to increase by 54%, from $234 billion to $361 billion. Stroke is projected to account for the largest absolute increase in costs.

2.2 AI in Cardiovascular Disease prediction

We looked into multiple studies on the subject of heart disease prediction starting with Monitoring Cardiovascular Problems in Heart Patients Using Machine Learning [7] That used University of California, Irvine (UCI) Dataset that comprises of 13 features and 303 instances per hospital to build a predictive model for heart disease prediction using multiple models like Extreme Gradient Boosting (XGBoost), Decision tree, Random Forest, Support Vector Machine (SVM), and other models. Then we looked into the research titled Heart Disease Detection Using Machine Learning Methods – A Comprehensive Review [8] their research worked using multiple datasets(X-ray images, Electrocardiogram(ECG), Blood Tests, Patient history) each of these datasets were trained with the Artificial Intelligence(AI) that fits it like using Convolutional neural network (CNN) to process X-ray images and other machine learning techniques to handle numerical and categorical values, they got astonishing results with their predictive model but it was at the cost of demanding many tests from patients that are costly and time consuming, after that we took great insights from the research titled Machine Learning-Based Approach to the Diagnosis of Cardiovascular Disease Using a Combined Dataset [9] in this research they focused on building a dataset using three different datasets (University of California, Irvine(UCI)dataset[2],IEEE dataset[10], Kaggle dataset[11]) after finishing

the combined dataset they developed multiple machine learning models and compared them using evaluation metrics to establish the model with the best results and in their case in was the random forest, and finally we looked at the study titled Prediction of Heart Disease Using Data Mining Techniques – A Case Study [12] in this study they used the University of California, Irvine (UCI) dataset and they used preprocessing techniques like filling out values using mean and median which lowered their accuracy rate, after gaining insight from the past work done in this subject our improvement will be mainly focused on data preprocessing and evaluating which is the best processing techniques to use and model creation.

Background

In this chapter we will be giving a technical explanation of our project and explaining all of the methods we used in a simple to understand yet informative structure we will start by defining Artificial Intelligence (AI)

3.1 Artificial Intelligence (AI)

. AI is a field of science that works on building machines that are able to mimic human logic and thinking gaining the ability to solve problems, AI has a subset called Machine Learning (ML) that focuses on making systems that learn from datasets and make prediction and decisions accordingly, many machine learning models were developed across the years to enhance the computers ability to make predictions and decisions, from these models we will be using

3.1.1 Logistic Regression

Logistic Regression is a statistical model that works on binary classification that predicts the target value (1 or 0) for multi classification it uses a method called one vs all which creates multiple models each predicting one of the target values and picks the one with the highest probability, logistic regression most commonly uses sigmoid function represented as

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Where:

- $f(x)$ is the output of the sigmoid function.
- x is the input to the function.

- e is Euler's number.

And the logistic regression as a whole is represented by:

$$y = \frac{e^{(b_0+b_1X)}}{1+e^{(b_0+b_1X)}} \tag{3.2}$$

Where:

- y is the predicted probability that the output is class 1 (positive class).
- b_0 is the intercept (bias term).
- b_1 is the coefficient (weight) for the feature X.
- X is the input feature (e.g., a variable like age, income, etc.).

3.1.2 K-nearest neighbor (KNN)

KNN is a clustering algorithm that is used for both classification and regression and it works by creating clusters (groups) from the dataset by similarity and when given a new point it classifies it based on it's neighbors, represented with this equation:

$$\hat{y}_i = \text{majority vote}\left(\{y_1, y_2, \ldots, y_K\}\right) \tag{3.3}$$

Where:

- $\hat{y}_i$ is the predicted class label for the i-th sample.
- $y_1, y_2, \ldots, y_K$ are the class labels of the K nearest neighbors of the i-th sample.
- K is the number of nearest neighbors considered.

3.1.3 Support Vector Machine (SVM)

SVM is a machine learning model that works on solving both classification and regression problems by finding the optimal way to separate the data into different classes by creating a plane using this equation

$$f(x) = \mathbf{w}^T\mathbf{x} + b \tag{3.4}$$

Where:

- $f(x)$ is the decision function, which gives the predicted value.
- $\mathbf{w}$ is the weight vector that defines the orientation of the hyperplane.
- $\mathbf{x}$ is the feature vector of the input data. - b is the bias term that shifts the hyperplane.

3.1.4 Decision Tree

Decision Tree is a machine learning model that has tree like structure to make predictions and decisions, it works by splitting the data based on feature values into subsets once for each feature leading to the final prediction on the final node (leaf node).

3.1.5 Random Forest

Random Forest is machine learning model for both classification and regression derived from the Decision Tree model, that works by making multiple decision tree model's and training them on random subsets of the dataset, after training the prediction of the class is made by the random forest by averaging the prediction of all trees which is represented by

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} h_t(x_i) \tag{3.5}$$

Where:

- $\hat{y}_i$ is the predicted output for the i-th sample.
- x_i is the feature vector of the i-th sample.
- T is the total number of trees in the forest.
- $h_t(x_i)$ is the prediction of the t-th decision tree for the i-th sample.

3.1.6 Gradient Boosting

Gradient Boosting is a machine learning model that works by building multiple weak prediction models sequentially with each model correcting the errors of the model before

it resulting in a strong and accurate model that makes the final prediction represented with this equation

$$\hat{y}_i = F(x_i) = \sum_{m=1}^{M} \alpha_m h_m(x_i) \tag{3.6}$$

Where:

- $\hat{y}_i$ is the predicted output for the i-th sample.
- x_i is the feature vector of the i-th sample.
- $F(x_i)$ is the overall model prediction for the input x_i.
- M is the total number of iterations (or trees) in the boosting process.
- α_m is the weight or contribution of the m-th model.

3.1.7 Extreme Gradient Boosting (XGBoost)

XGBoost is a machine learning model that is used for both classification and regression, it is derived from gradient boosting, it works by building a gradient boosting model's and optimizing the models using multiple techniques like parallel processing, regularization and many other techniques the final is represented by this equation

$$\hat{y}_i = F(x_i) = \sum_{m=1}^{M} \alpha_m h_m(x_i) \tag{3.7}$$

Where:

- $\hat{y}_i$ is the predicted output for the i-th sample.
- x_i is the feature vector of the i-th sample.
- $F(x_i)$ is the final prediction from the ensemble of M models.
- M is the number of boosting rounds (trees) in the ensemble.
- α_m is the weight or contribution of the m-th base model.
- $h_m(x_i)$ is the m-th base learner applied to input x_i.

3.1.8 AdaBoost

AdaBoost is machine learning model that works by combining multiple weak learners and training them iteratively to make a strong more accurate prediction rather than

just randomly guessing and the final prediction is represented by

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \tag{3.8}$$

Where:

- $H(x)$ is the final strong classifier.
- T is the total number of weak learners.
- $h_t(x)$ is the prediction of the t-th weak learner.
- α_t is the weight of the t-th weak learner based on its performance.
- sign$(\cdot)$ returns +1 or -1 depending on the sign of the sum.

3.2 Model Evaluation Metrics

After creating these models we need to evaluate them to figure out how accurate the models are as problem solvers and in comparison to each other, the evaluation metrics we are using are

3.2.1 Accuracy

Accuracy is the percent of correct guesses among all guesses which is represented by

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

Where:

- TP = True Positives.
- TN = True Negatives.
- FP = False Positives.
- FN = False Negatives.

3.2.2 Precision

Precision is the percent of correct positive guesses among all positive guesses represented by

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.10}$$

Where:

- TP: True Positive.

- FP: False Positive.

3.2.3 Recall

Recall measures the percentage of positives guesses that were correctly made, represented by

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.11}$$

Where:

- TP: True Positive.

- FN: False Negative.

3.2.4 F1 Score

F1 Score represent a balanced measure using both recall and precision, represented by

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.12}$$

3.2.5 ROC-AUC Curve

Receiver Operating Characteristic (ROC) is a plot that evaluates classification models by visualizing true positive rates and false positive rates Area Under the Curve (AUC) is a metric for evaluating classification models by distinguishing between two classes. ROC-AUC Curve combines both evaluation metrics to create a more concise evaluation metric for classification models as shown in figure

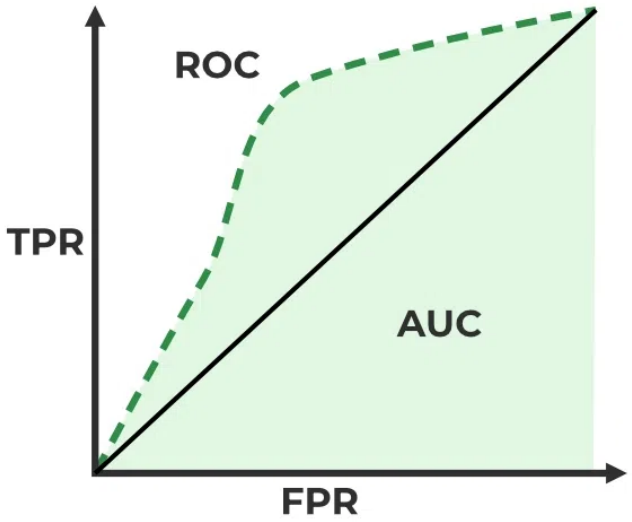

Figure 3.1: ROC-AUC Curve[1]

3.3 Preprocessing techniques

before creating the model and evaluating it we need to make sure that the learning dataset is provided in a balanced and clean structure to get the best performance and accuracy from the models, to achieve that we used the following techniques

3.3.1 Random Forest Imputer

Random Forest Imputer is a preprocessing technique that uses random forest regressor to predict missing values based on the prediction accuracy of the machine learning model.

3.3.2 Label Encoder

Label Encoder is used to convert numerical values to categorical values.

3.3.3 StandardScaler

StandardScaler is used to change the values of numerical data into a range of (0-1) while keeping the same weights to the values within the same feature.

3.3.4 Quantile Transformer

Quantile Transformer is used to transform the features in our dataset following specific distributions.

Objectives

In this project we aim to use Artificial Intelligence (AI) to develop an accurate model for heart disease prediction our objectives are:

4.1 Developing a Predictive Model

To develop machine learning models that are able to predict heart disease risk, the models we aim to develop are Logistic Regression, K-nearest neighbor (KNN), Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, and Extreme Gradient Boosting (XGBoost). then we will run evaluation metrics on all models to recognize the best model for heart disease detection.

4.2 Leveraging AI for Early Disease Risk Prediction

To use Artificial Intelligence (AI) as diagnosis tool and a decision support system for medical professionals increasing the ability to identify heart disease in ti's early stages and carrying our the right treatment plan for patients, these machine learning models carry the ability to identify the risk of heart disease before the appearance of symptoms on patients.

4.3 Decreasing the Economical Toll on the Health Sector

To reduce the financial toll on the healthcare sector by providing a powerful diagnosis tool and a decision making support system that reduces the diagnosis cost on both hospitals and patients, and to decrease the need for patient care and expensive patient treatment by increasing the chances of early disease detection and improving patient outcomes.

4.4 Providing a Clear Base for Future Work

To write a well structured thoroughly explained report, clarifying each stage that the project went through and all the insights we gained into the dataset, and our evaluation of each model used in this project, aiming to help any future work on this field of study.

Dataset

5.1 Data Selection

In this project we used two datasets and combined them into one dataset, the datasets we used were kaggle and UCI,we selected the UCI Cleveland Heart Disease Dataset [2], which is one of the most widely used datasets in machine learning research for cardiovascular disease prediction. It was chosen based on the High clinical relevancy it offers with it having the 14 most important features in the subject of cardiovascular disease prediction, and because of its public availability in the field, and the the other dataset we used was (Heart Disease Dataset)[11] from kaggle due to the variety and high number of patients and it's compatibility with our base dataset .

5.2 Dataset Description

The University of California, Irvine (UCI) data set contains 76 features but not all features have a connection to heart disease prediction and it contains features such as

Figure 5.1: UCI Data card [2]

null values and unrelated features like earlobe type so we mainly focus on the selected 13 key features of the dataset as shown in the data card in figure (4.1), and in the kaggle dataset we also used the exact same 13 key features to establish one large combined dataset .

Table 5.1: Key Features of the combined Disease Dataset

No.	**Feature**	**Description**	**Data Type**
1	Age	Patient's age in years (28-77)	Numeric
2	Sex	1 = Male, 0 = Female (79% males, 21% females)	Categorical
3	Chest Pain Type (cp)	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic	Categorical
4	Resting Blood Pressure (trestbps)	Resting blood pressure (80-200 mmHg)	Numeric
5	Serum Cholesterol (chol)	Serum cholesterol levels (120-420 mg/dL)	Numeric
6	Fasting Blood Sugar (fbs)	>120 mg/dL: 1 = True, 0 = False	Categorical
7	Resting ECG Results (restecg)	0 = Normal, 1 = ST-T abnormality, 2 = Left ventricular hypertrophy	Categorical
8	Maximum Heart Rate (thalach)	Maximum heart rate achieved (80-202 bpm)	Numeric
9	Exercise-Induced Angina (exang)	1 = Yes, 0 = No	Categorical
10	ST Depression (oldpeak)	ST depression induced by exercise (0.84-4.44)	Numeric
11	Slope of Peak ST Segment (slope)	1 = Upsloping, 2 = Flat, 3 = Downsloping	Categorical
12	Major Vessels (ca)	Number of major vessels colored by fluoroscopy (0-3)	Numeric
13	Thalassemia Test (thal)	3 = Normal, 6 = Fixed defect, 7 = Reversible defect	Categorical

5.3 Data Collection

The University of California, Irvine (UCI) Dataset was originally collected from the Cleveland Clinic and was donated to the UCI Machine Learning Repository on June 30, 1988. The dataset is available in CSV format, making it easy to use for machine learning applications, and the data includes both numerical and categorical variables that are essential for training models, and the kaggle dataset was collected from multiple sources and worked on and validated by many machine learning experts and is highly compatible with our base dataset.

5.4 Data Analysis

We thoroughly explored our dataset to identify the most common occurrences and made multiple plots to visually identify patterns and to better understand the relationships between the dataset features:

5.4.1 Age

In this plot you can see the age with the number of cases:

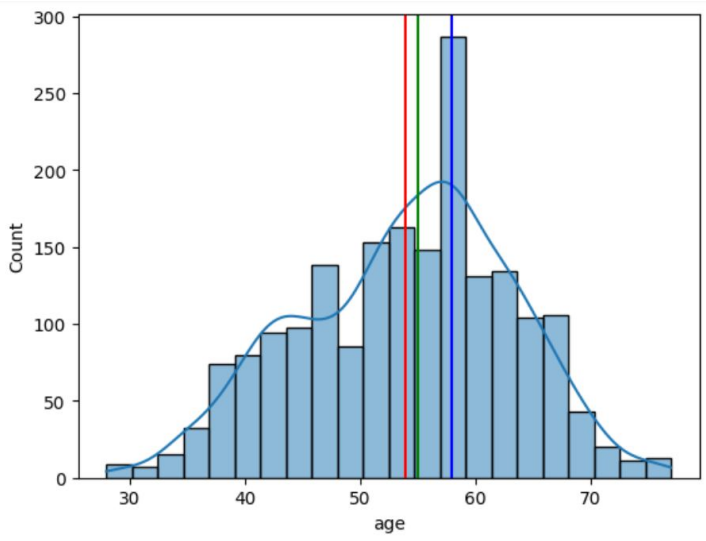

Figure 5.2: Age-Count plot

As seen the most common ages to test for heart disease are ages between 50 and 60 with a mean of 54, a median of 55 and the mode being 58.

5.4.2 Gender

The dataset displayed a huge difference between genders where the male percentage of our records is 78.91% and the female percentage is 21.09%, this shows that the male percentage is 274.23% more than female percentage.

5.4.3 Chest Pain

The dataset we are using contains four types of chest pain:

- asymptomatic(red): chest pain where the patient doesn't experience any pain and is not showing any symptoms.
- typical angina(blue): chest pain that is caused by physical or emotional stress.
- atypical angina(green): chest pain that is usually describes as pressure, tightness, or squeezing in the chest.
- non-anginal(purple): chest pain that is described as stabbing which is caused mainly by disease.

in this figure we visualize the pain type in (age-count) plot

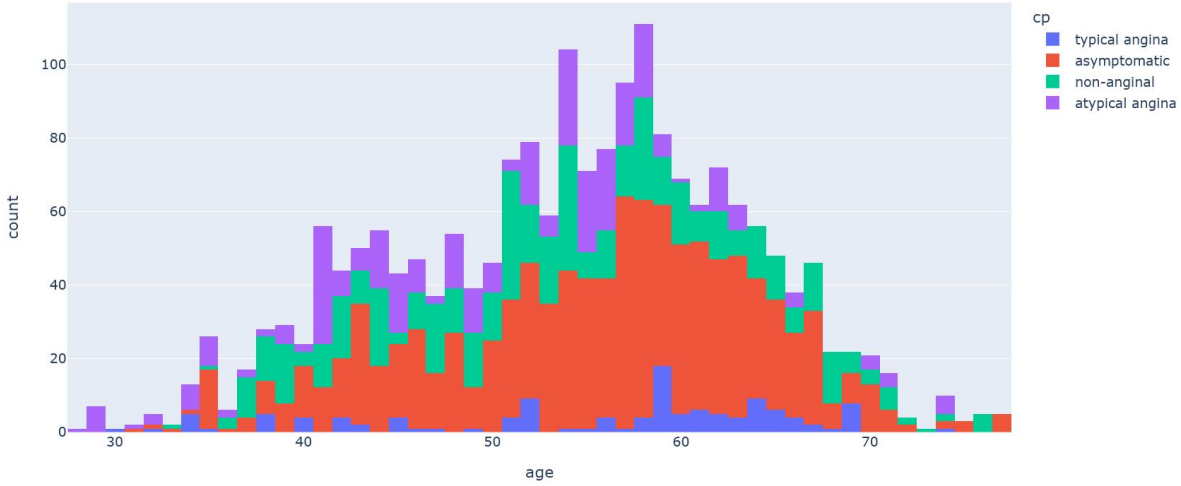

Figure 5.3: Chest Pain plot

5.4.4 Resting Electrocardiogram (ECG)

Resting ECG has three values (normal, lv-hypertrophy, st-t-abnormality) as shown in the figure

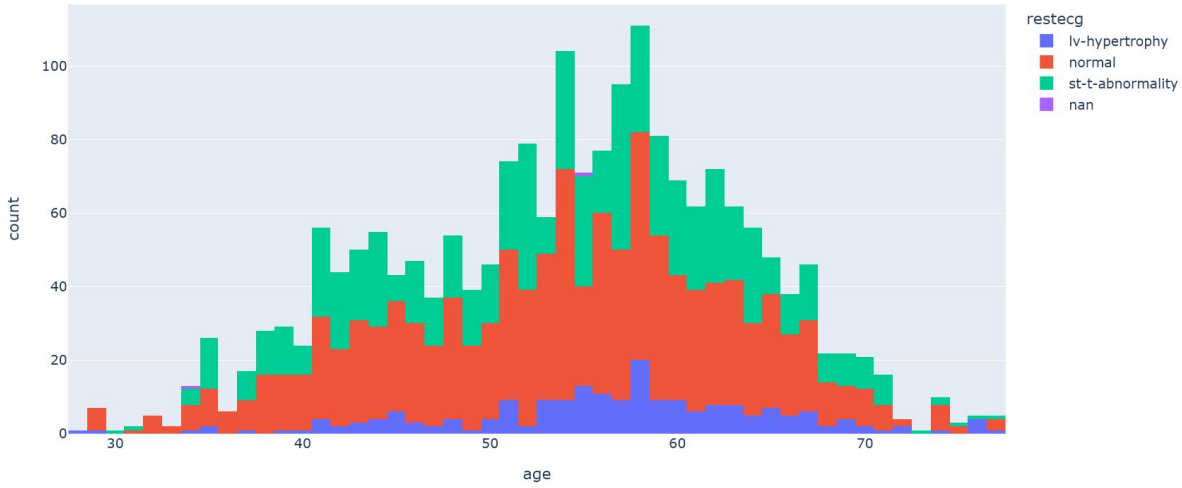

Figure 5.4: ECG plot

Where normal is the mast common ECG reading, after that comes st-t-abnormality which indicates a sign of an underlying heart condition.

5.4.5 ECG Slope

ECG Slope has 3 different reading (downsloping, flat, upsloping) and a null value (NaN) in our dataset as shown in the figure

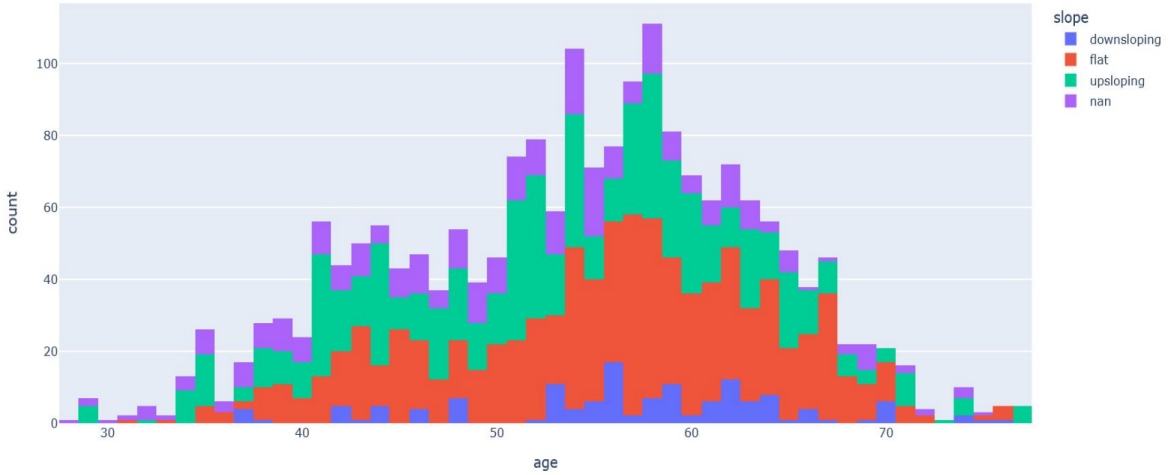

Figure 5.5: ECG Slope plot

Anything other than normal may indicate the existence of an underlying heart condition.

5.4.6 Fasting Blood Sugar

Fasting Blood Sugar is indicated as (True, False) in the dataset by giving True values to numbers above 120 as shown in the figure

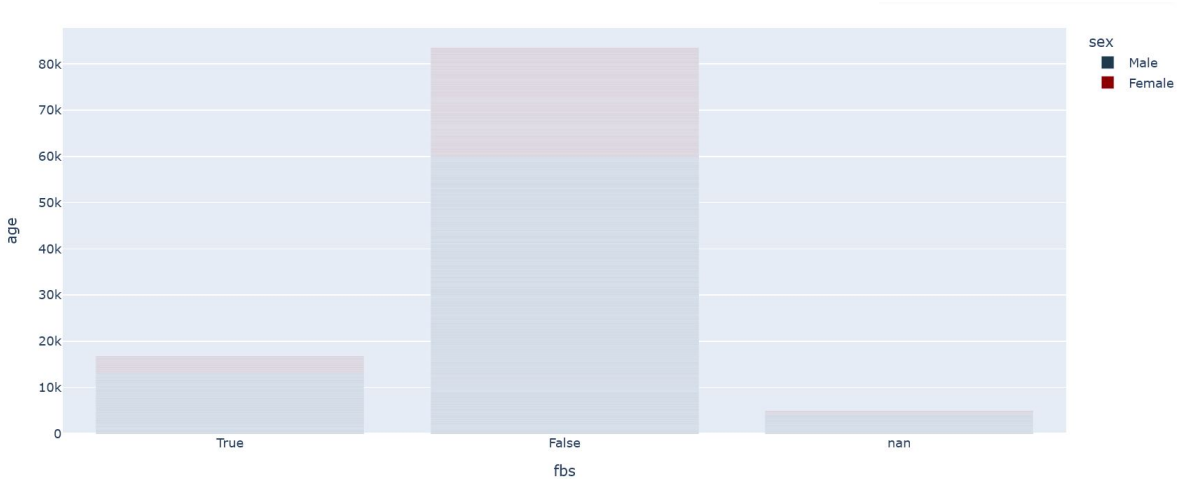

Figure 5.6: Fasting Blood Sugar Plot

where True values indicate the likelihood of having a heart condition

5.4.7 Heart disease (target)

This is our prediction target were it has 4 values ,(0) meaning no heart disease,(1) meaning mild heart disease, (2) meaning moderate heart disease,(3) meaning severe heart disease,(4) critical-state heart disease.

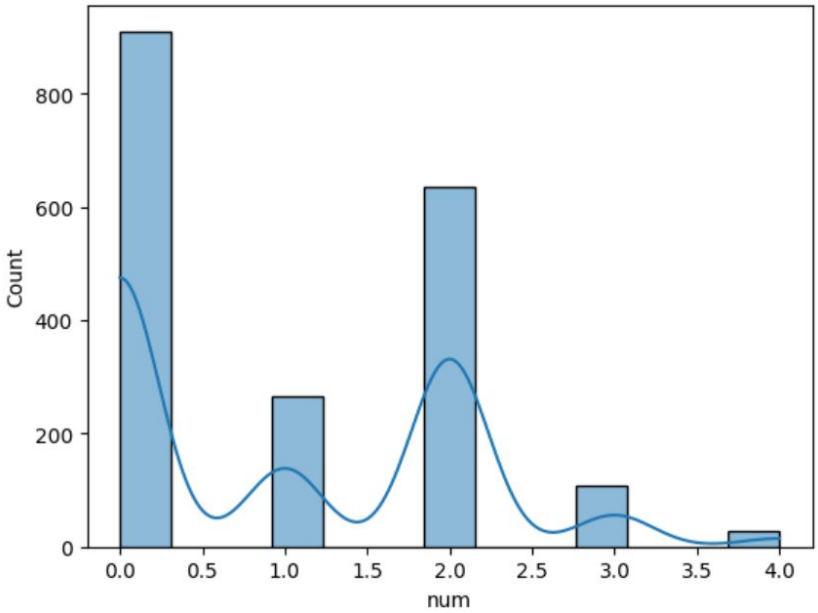

Figure 5.7: Heart Disease plot

As seen in the figure the two most common occurrences is (0) meaning most cases don't suffer from heart disease and this is due to the normal heart disease ratio in the population and after it comes (2) becuase this is when patients start experiencing heart disease symptoms and want to get medically checked.

Methodology

This chapter will provide a step by step explanation for the development of this project and the results of all the used models and techniques.

6.1 Dataset Preprocessing

In this section we will explain and Provide the techniques used to preprocess and join our datasets and their results

6.1.1 Joining Datasets

We joined both UCI and Kaggle dataset into one large combined dataset, to do this we had to rename the feature values of the Kaggle dataset minimally without changing it's accrual value like changing categorical and numerical value types and changing the actual string values in example from (true, false) to (True, False).

6.1.2 Filling Missing Values

To fill out missing values we used random forest imputer which provides a more accurate value in comparison to (mean, median, mode) and here is the accuracy scores it achieved

Feature	Imputation Accuracy (%)
fbs	86.38
restecg	77.38
exang	87.40
slope	72.24
thal	78.92

Table 6.1: Imputation accuracy for various features

6.1.3 Handling Outliers

The dataset was collected by medical professionals and all unrealistic values where left missing so the dataset only showed one outlier through visualizing each feature and it was dropped and filled by the imputer.

6.1.4 Standardization and Normalization

For standardization we used Standardscaler which is provided from sklearn.
And for Normalization i used Quantile Transformer which is also provided by sklearn.

6.1.5 Train-Test Split

We will try three different ratios of training and testing data splits discussing thier results later in the report

- 70%training, 30% testing
- 75%training, 25% testing
- 80%training, 20% testing

6.2 Model Development

This section will address all the steps in the model development

6.2.1 Model Creation

Based on Past Work in the subject we chose 8 different machine learning models, the machine learning models we chose were (Logistic Regression, K-nearest neighbor (KNN), Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, and Extreme Gradient Boosting (XGBoost)).

6.2.2 Model Optimization

All models but logistic regression were subject to hyperparameter tuning each model tuned with it's own parameters to achieve the best results on all models.

6.2.3 Cross Validation

Cross Validation was used on all models to best minimize the chance of overfitting the models to the dataset.

6.3 Model Evaluation

All models were evaluated using accuracy and f1-score and cross validation score

Model	**Cross-Validation Accuracy**	**Test Accuracy**	**F1 Score**
SVM	0.7999	0.8432	0.6715
XGBoost Classifier	0.8835	0.9254	0.9603
AdaBoost Classifier	0.6564	0.6735	0.5416
KNN	0.7330	0.7584	0.4958
Logistic Regression	0.7632	0.7918	0.7811
Random Forest Classifier	0.8777	0.9254	0.8735
Decision Tree Classifier	0.8642	0.9177	0.9563
Gradient Boosting Classifier	0.8790	0.9100	0.9517

Table 6.2: Model Performance Comparison

As seen in the table our best performing model was Extreme Gradient Boosting (XGBoost) with a Cross-Validation Accuracy score of: 88%, accuracy score of : 92%, f1 score of: 96% and we ran roc-auc evaluation for it and it scored: 98% out performing all other models in problem solving the heart disease prediction problem specifically, using 20% testing data

In other training-testing splits Extreme Gradient Boosting (XGBoost) achieved an accuracy of 89%(30-70 split),and an accuracy of 90%(25-75 split).

6.4 User Interface Development

After fully developing our Artificial Intelligence (AI) model we started building a user interface by exporting our best performing model and integrating it into out user interface that is built using python and streamlit specifically to host a local web interface for ease of use and prototyping a fully functional interface for the prediction and diagnosis of heart disease

Figure 6.1: User Interface

Discussion

Our first test was by looking into the past work and findings in this subject then we decided to carry out a mimicking process for one of the past studies named Prediction of heart disease using data mining techniques: A Case study [12] after doing it step by step we were able to replicate their finding but to further look into the process we decided to change on one of their preprocessing techniques by changing their way of filling out missing values from using mean and median to using the random forest imputer method which fills out the missing values using random forest regression on each patient, through this change we were able to achieve higher results their best performing model that was (Random forest) which achieved the results of 60% accuracy we improved upon that result by just changing the techniques used for filling out missing values to increase the accuracy of the best performing model (Random forest) to 72% accuracy. After that we started working on building our development pipeline from scratch by making our own custom combined dataset preprocessing it and developing our machine learning models and evaluating them, our best performing model was Extreme Gradient Boosting (XGBoost) with an accuracy score of : 92% and an f1 score of: 96% which out performs most studies and researches in the field but one study(Machine learning-based approach to the diagnosis of cardiovascular disease using a combined dataset)[9] where they achieved 99.12% accuracy using decision tree machine learning model but without clear clarification on the techniques used in the dataset or the measures taken to increase the models performance.

our results gave us great hope in further accomplishments in the subject and the ability to greatly help in the detection of heart disease in it's early stages to improve patients outcomes and lessen the economical burden on both the patients and the health sector.

Conclusion

In this project we developed a machine learning model that is capable of acting as a diagnosis tool or a decision support system for medical experts helping in making accurate and timely diagnosis and detecting heart disease in it's early stages to start the right treatment for patients helping in decreasing the mortality rate of heart diseases and decreasing the economical toll of heavy treatment and diagnosis on both hospitals and patients as a whole, and opening up the space for AI to be used as a tool in the healthcare field to provide for better outcomes

8.1 Limitations

This project's limitations are the difficulty in predicting rare cases of heart disease that are not documented commonly, and the difficulty in detecting heart diseases that don't show any indications on the data of patients, regarding the technological side of it we have the limitation of having models that demand high computational power that will not be able to run on low level devices like watches and electronic accessories.

8.2 Future Work

We suggest implementing the model into the health sector by either providing it to healthcare experts to use as a decision support system to help make more accurate and timely interventions, or connecting the model to a patient data base to act as an early detection diagnosis tool,this work can be used as base to be worked on for further improvement and research like integrating a real time updating dataset from multi regional hospitals to further improve the model.

References

[1] GeeksforGeeks, "Auc roc curve in machine learning," 07 Feb, 2025.

[2] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease." UCI Machine Learning Repository, 1989. DOI: `https://doi.org/10.24432/C52P4X`.

[3] T. Gaziano, K. S. Reddy, F. Paccaud, S. Horton, and V. Chaturvedi, "Cardiovascular disease," *Disease Control Priorities in Developing Countries. 2nd edition*, 2006.

[4] W. H. Organization, "Cardiovascular diseases (cvds)," 11 June, 2021.

[5] N. I. of Health, "Heart and vascular diseases," 17 Jun, 2022.

[6] D. S. Kazi, M. S. Elkind, A. Deutsch, W. N. Dowd, P. Heidenreich, O. Khavjou, D. Mark, M. E. Mussolino, B. Ovbiagele, S. S. Patel, R. Poudel, B. Weittenhiller, T. M. Powell-Wiley, K. E. J. Maddox, and on behalf of the American Heart Association, "Forecasting the economic burden of cardiovascular disease and stroke in the united states through 2050: A presidential advisory from the american heart association," *Circulation*, vol. 150, no. 4, pp. e89–e101, 2024.

[7] A. Al Ahdal, M. Rakhra, R. R. Rajendran, F. Arslan, M. A. Khder, B. Patel, B. R. Rajagopal, and R. Jain, "Monitoring cardiovascular problems in heart patients using machine learning," *Journal of healthcare engineering*, vol. 2023, no. 1, p. 9738123, 2023.

[8] M. Hajiarbabi, "Heart disease detection using machine learning methods: a comprehensive narrative review," *Journal of Medical Artificial Intelligence*, vol. 7, no. 0, 2024.

[9] K. M. Mohi Uddin, R. Ripa, N. Yeasmin, N. Biswas, and S. K. Dey, "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset," *Intelligence-Based Medicine*, vol. 7, p. 100100, 2023.

[10] M. Siddhartha, "Heart disease dataset (comprehensive)," 2020.

[11] J. Smith, "Heart disease dataset." `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`, 2020.

[12] M. R. Rafi, S. Shadid, A. Shafkat, M. Yasmeen, and S. Shibli, "Prediction of heart disease using data mining techniques: A case study," *ResearchGate*, 12 2019.