Project Report

on

Heart Disease Detection using Machine Learning

Submitted to

Sant Gadge Baba Amravati University

In partial Fulfillment of the Requirement

For the Degree of

Bachelor of Engineering in

Computer Science and Engineering

Submitted by:

M Areeb Ozair
Suryakant Ingle
Apeksha Mundhada
Rutika Dharangaonkar

Under the Guidance of

Prof. S. B. Pagrut

Department of Computer Science and Engineering

Shri Sant Gajanan Maharaj College of Engineering,

Shegaon – 444 203 (M.S.)

2022-23

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON – 444 203 (M.S.)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **Mr. M Areeb Ozair, Mr. Suryakant Ingle** and **Ms. Apeksha Mundhada** and **Ms. Rutika Dharangaonkar** students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute has completed the project work entitled **"Heart Disease Detection using Machine Learning"** based on syllabus and has submitted a satisfactory account of his work in this report which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

Prof. S. B. Pagrut
Project Guide

Dr. S. B. Patil
Head of Department

Dr. S. B. Somani
Principal

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON – 444 203 (M.S.)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project work entitled "**Heart Disease Detection using Machine Learning**" submitted by **Mr. M Areeb Ozair, Mr. Suryakant Ingle** and **Ms. Apeksha Mundhada** and **Ms. Rutika Dharangaonkar** students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute, is a satisfactory account of his work based on syllabus which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

Internal Examiner **External Examiner**

Date: **Date:**

Abstract

In this digital world, data is an asset, and enormous data was generated in all the fields. Data in the healthcare industry consists of all the information related to patients. Now day's cardiovascular diseases are growing rapidly by busy and stress full life. All type of age groups is under the threat of these chronic diseases so there is a need of detection of these diseases by using symptoms or reports. Early identification and treatment are the best available option for the infected people. Main objective behind to develop a system is to help the doctors to cross verify their diagnosed results which gives promising solution over existing death rates.

The recent development in medical supportive technologies based on data mining, deep learning plays an important role in detecting cardiovascular diseases that are caused by taking many factors in consideration such as age, type of chest pain, blood pressure, cholesterol levels, etc.

By using our proposed work try to implement a promising solution for detection of heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease.

Acknowledgement

The real spirit of achieving a goal is through the way of excellence and lustrous discipline. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various personalities.

I would like to take this opportunity to express my heartfelt thanks to my guide ***Prof. S.B.Pagrut,*** *for his esteemed guidance and encouragement, especially through difficult times. His suggestions broaden my vision and guided me to succeed in this work. I am also very grateful forhis guidance and comments while studying part of my seminar and learnt many things under his leadership.*

I extend my thanks to ***Dr. S.B. Patil*** *Head of Computer Science & Engineering Department, Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support that made me consistent performer.*

I also extend my thanks to ***Dr. S. B. Somani*** *,Principal Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support.*

Also, we would like to thanks to all teaching and non-teaching staff of the department for their encouragement, cooperation and help. My greatest thanks are to all who wished me success especially my parents, my friends whose support and care makes me stay on earth.

1. **Mr. M Areeb Ozair**
2. **Mr.Suryakant Ingle**
3. **Ms.Apeksha Mundhada**
4. **Ms.Rutika Dharangaonkar**

Final Year B. E. CSE
Session 2022-23

Contents

List of Figures and Tables

Abbreviations

ML	Machine Learning
EDM	Educational Data Mining
LA	Learning Analytics
AI	Artificial Intelligence
GR	Gain Ratio
DT	Decision Tree
LR	Logistic Regression

CHAPTER 1
INTRODUCTION

1. Introduction

1.1. PREFACE

Machine Learning is a powerful tool that enables us to extract valuable information from data that was previously unknown or implicit. The domain of machine learning is extensive and multifaceted., and it encompasses various classifiers such as supervised, unsupervised, and ensemble learning, that can be employed to forecast and assess the precision of a particular dataset. The implementation of machine learning is increasing day by day, and it has the potential to revolutionize many fields, including healthcare. Cardiovascular disease (CVD) is an area in healthcare that can significantly gain from the implementation of machine learning techniques. With 17.9 million fatalities globally, as per the World Health Organization, CVD is currently the primary cause of death in adults. To help address this problem, our project aims to predict which patients are likely to be diagnosed with CVD based on their medical history. By recognizing patients who exhibit symptoms for example, chest pain or elevated blood pressure, we can help diagnose the illness with fewer medical examinations and provide more efficient treatments. Our project focuses on three data mining techniques: XGBoost, KNN, and Random Forest Classifier. By using these techniques in combination, we are able to achieve an accuracy rate of above 95%, which is better than previous systems that relied on only one data mining technique. The objective of our project is to classify by examining their medical characteristics, such as age, gender, fasting sugar levels, chest pain, and more, it is possible to predict whether a person is likely to have heart disease or not.

To accomplish this, we selected a dataset from the Kaggle repository this dataset was created by combining different datasets already available independently but not combined before, that contains medical history and characteristics of the patient. We trained our algorithms using the 12 medical attributes of each patient and used XGBoost, Random Forest and KNN to classify the patients based on their medical history. We found that XGBoost was the most efficient algorithm, and it provided us with an accuracy rate of above 95%. Our project has the

potential to significantly improve the diagnosis and treatment of CVD by identifying patients who are at risk of developing the disease. By using multiple data mining techniques, we were able to achieve a higher accuracy rate and provide a more cost-efficient method for predicting CVD.

1.2. MOTIVATION

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using classification algorithms namely, XGBoost, Logistic Regression, KNN, and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, these algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better way to detect heart disease.

1.3 PROBLEM STATEMENT

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1.4 OBJECTIVES

1. Data collection from different sources and pre-processing
2. To develop machine learning model to predict future possibility of heart disease by implementing Logistic Regression, KNN, XGboost, Decision Tree
3. To determine significant risk factors based on medical data set which may lead to heart disease.
4. To analyse feature selection methods and understand their working principle

1.5 SCOPE AND LIMITATIONS

1.5.1. Scope

1. The system will help identify important factors that lead to a heart disease.
2. The main scope of the project is to detect heart disease.
3. It will help the patients to obtain results quick and diagnose as early as possible

1.6 ORGANIZATION OF PROJECT

The project is organized as follows:

Chapter 1 gives introduction to the project.

Chapter 2 provides literature survey of the project.

Chapter 3 explains materials and methods required to complete the project.

Chapter 4 provides implementation of project.

Chapter 5 provides deployment phase of the project.

Chapter 6 gives conclusion of the project

Chapter 7 discuss future scope of the project.

CHAPTER 2
LITERATURE SURVEY

2. LITERATURE SURVEY

Ijaz Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling" model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analysing the results, we tend to reveal the importance of respecting the sequential nature of clinical records [1].

Aakash Chauhan et al. (2018) presented "Heart Disease Prediction using Evolutionary Rule Learning". This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient's dataset. This will facilitate (help) in decreasing the number of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness [2].

Ashir Javeed, Shijie Zhou et al. (2017) designed "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program. Two forms of experiments are used for cardiovascular disease prediction. In the first form, only random forest model is developed and within the second experiment the proposed Random Search Algorithm based random forest model is developed. This methodology is efficient and less complex than conventional random forest model. Comparing to conventional random forest it produces 3.3% higher accuracy. The proposed learning system can help the physicians to improve the quality of heart failure detection.

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposed by Senthilkumar Mohan, Chandrasegar Thirumalai et al. (2019) was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes was chosen from the pre-processed knowledge(data) set of cardiovascular disease. After prep-processing, the hybrid techniques were applied and disgnosis the cardiovascular disease [4].

K. Prasanna Lakshmi, Dr. C.R.K.Reddy (2015) designed "Fast Rule-Based Heart Disease Prediction using Associative Classification Mining". In the proposed Stream Associative Classification Heart Disease Prediction (SACHDP), we used associative classification mining over landmark window of data streams. This paper contains two phases: one is generating rules from associative classification mining and next one is pruning the rules using chi-square testing and arranging the rules in an order to form a classifier. Using these phases to predict the heart disease easily[5].

M.Satish, et al. (2015) used different Data Mining techniques like Rule based, Decision Tree, Naive Bayes, and Artificial Neural Network. An efficient approach called pruning classification association rule (PCAR) was used to generate association rules from cardiovascular disease warehouse for prediction of heart disease. Heart attack data warehouse was used for pre-processing for mining. All the above discussed data mining technique were described [6].

Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik (2015) "An Intelligent Decision Support System for Cardiac Disease Detection", designed a cost-efficient model by using genetic algorithm optimizer technique. The weights were optimized and fed as an input to the given network. The accuracy achieved was 90% by using the hybrid technique of GA and neural networks [7].

"Prediction and Diagnosis of Heart Disease by Data Mining Techniques" designed by Boshra Bahrami, Mirsaeid Hosseini Shirvani. This paper uses various classification methodology for diagnosing cardiovascular disease. Classifiers like

KNN, SVM classifier and Decision Tree are used to divide the datasets. Once the classification and performance evaluation the Decision tree is examined as the best one for cardiovascular disease prediction from the dataset [8].

Mamatha Alex P and Shaicy P Shaji (2019) designed "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique". This paper uses techniques of Artificial Neural Network, KNN, Random Forest and Support Vector Machine. Comparing with the above-mentioned classification techniques in data mining to predict the higher accuracy for diagnosing the heart disease is Artificial Neural Network [9].

CHAPTER 3
METHODOLOGY

3. METHODOLOGY

3.1 SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system. The working of this system is shown below:

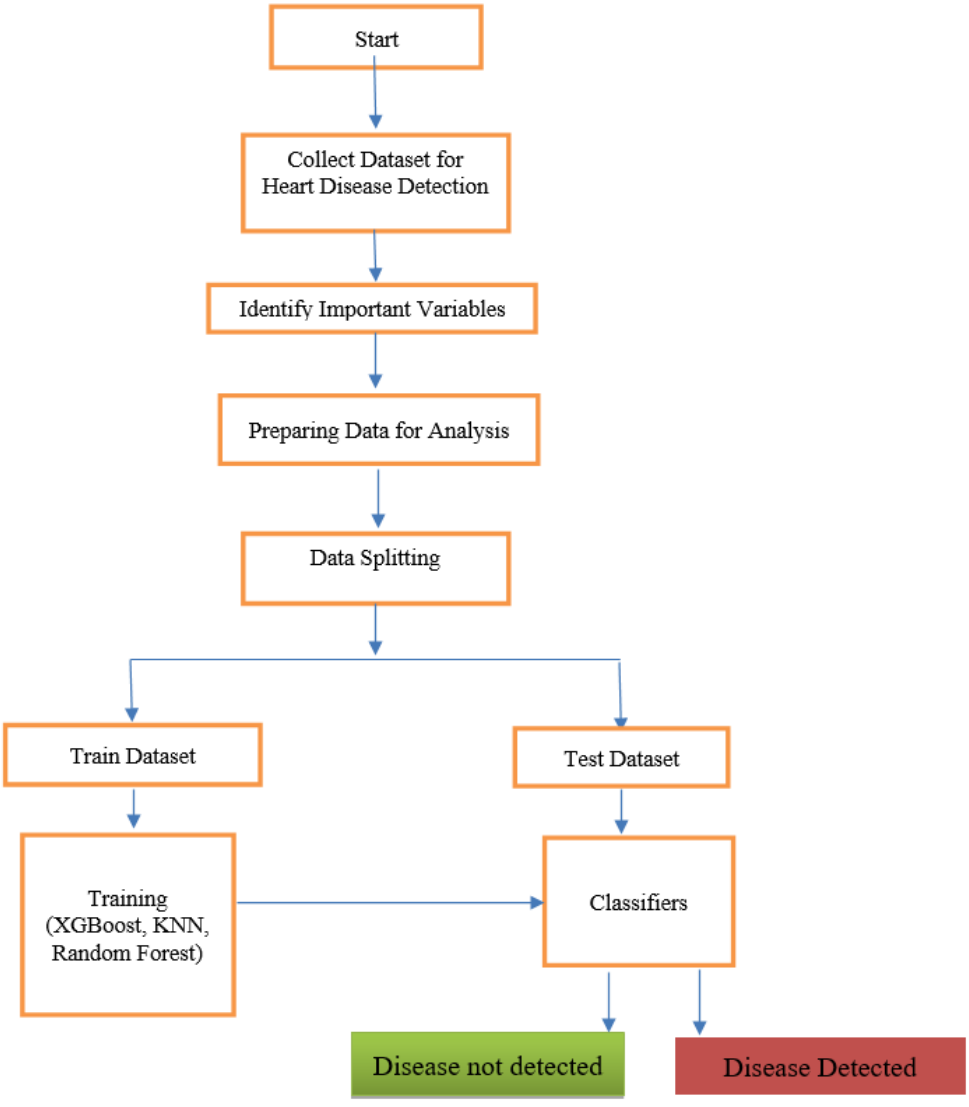

Figure 1: Proposed System

3.2 DATASET DETAILS

Dataset Link: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Dataset Attributes

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

3.3 MACHINE LEARNING

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

3.3.1 SUPERVISED MACHINE LEARNING

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the

output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset. The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).

Categories of Supervised Machine Learning:

Supervised machine learning can be classified into two types of problems, which are given below:

a) Classification

b) Regression

a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset.

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- Simple Linear Regression Algorithm
- Multivariate Regression Algorithm
- Decision Tree Algorithm
- Lasso Regression

3.3.2 UNSUPERVISED MACHINE LEARNING

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabelled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision. The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

3.4 SUPERVISED ALGORITHMS

3.4.1 RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output.

Therefore, below are two assumptions for a better Random Forest classifier:

1. There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
2. The predictions from each tree must have very low correlations.

Advantages:

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages:

- Although Random Forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

3.4.2 K-NEAREST NEIGHBOUR

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

3.4.3 LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

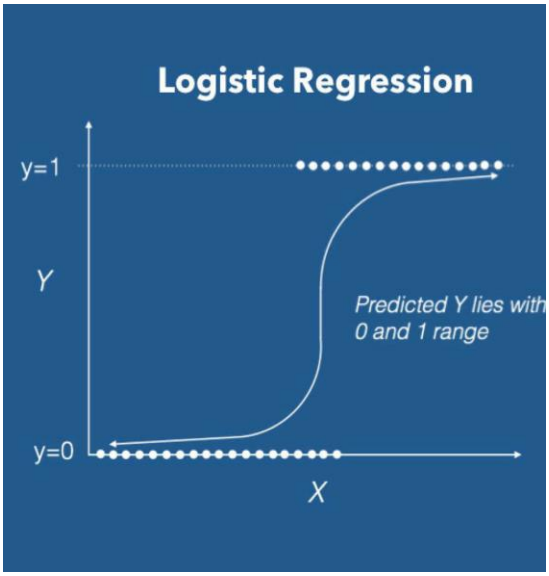

Figure 2: Logistic Regression

3.4.4 XGBOOST

XGBoost is an efficient implementation of Gradient Boosted decision trees that is designed to improve both the speed and performance of models. The algorithm creates decision trees sequentially, with weights assigned to all independent variables that are then fed into the tree for prediction. The weight of variables that are predicted incorrectly is increased, and those variables are fed into a second decision tree. The individual classifiers or predictors are then combined to create a strong and precise model that can work on regression, classification, ranking, and user-defined prediction tasks.

One of the strengths of XGBoost is its built-in L1 and L2 regularization, which helps prevent overfitting and makes it a regularized form of GBM. When using the Scikit Learn library, alpha and lambda hyper-parameters related to regularization are passed to XGBoost. Alpha is used for L1 regularization, while lambda is used for L2 regularization.

Another strength of XGBoost is its ability to leverage parallel processing to execute models much faster than GBM. When using the Scikit Learn library, the nthread hyper-parameter is used for parallel processing, representing the number of CPU cores to be used. If you want to use all available cores, don't specify a value for nthread, and the algorithm will detect them automatically.

XGBoost also has built-in capabilities to handle missing values. When the algorithm encounters a missing value at a node, it tries both left and right-hand splits and learns the way that leads to a higher loss for each node. It then does the same when working on testing data.

Cross-validation is another feature of XGBoost that allows the user to run a cross-validation at each iteration of the boosting process, making it easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM, where a grid search must be run, and only a limited number of values can be tested.

Finally, XGBoost uses effective tree pruning to remove splits beyond which there is no positive gain.

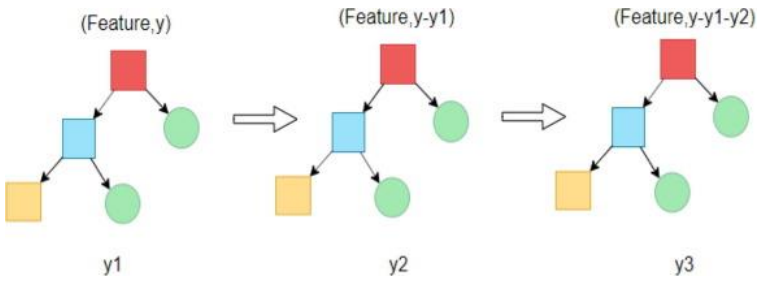

Figure 3: XGBoost

CHAPTER 4
IMPLEMENTATION

4. IMPLEMENTATION

4.1 EXISTING SYSTEM

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

4.2 PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1. Data Collection
2. Data Pre-Processing
3. Feature Selection
4. Model Selection

4.2.1 DATA COLLECTION

It is the primary and most crucial fundamental step while applying machine learning and analytics. The data required in this project is the patient's medical data. We have collected the dataset from Kaggle which includes all the required information for prediction. The features that the dataset includes are medical information like age, sex, chest paint type, resting blood pressure, cholesterol, fasting blood sugar, old peak etc. The dataset consists of 918 observations having 14 attributes.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	52	1	0	125	212	0	1	168	0	1	2	2	3	0
3	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
4	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
5	61	1	0	148	203	0	1	161	0	0	2	1	3	0
6	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
7	58	0	0	100	248	0	0	122	0	1	1	0	2	1
8	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
9	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
10	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
11	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
12	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
13	43	0	0	132	341	1	0	136	1	3	1	0	3	0
14	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
15	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
16	52	1	0	128	204	1	1	156	1	1	1	0	0	0
17	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
18	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
19	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
20	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
21	58	1	2	140	211	1	0	165	0	0	2	0	2	1
22	60	1	2	140	185	0	0	155	0	3	1	0	2	0
23	67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
24	45	1	0	104	208	0	0	148	1	3	1	0	2	1
25	63	0	2	135	252	0	0	172	0	0	2	0	2	1

Figure 4: Parameters of Selected dataset

4.2.2 DATA PRE-PROCESSING

This is one of the most crucial tasks in the process of analytics. Often it is observed that more than half of the total time of analytics process is taken by pre-processing phase. It is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

4.2.3 FEATURE SELECTION

Once we have the required data, next step is featuring extraction. Many times, it happens that some features do not contribute in evaluation or have negative impact on the accuracy. Feature selection is the step where we try to reduce number of features and try to create new features from existing ones. These new features now created should summarize the information obtained from existing features. The final features to be considered while prediction can be identified using correlation matrix shown in following image:

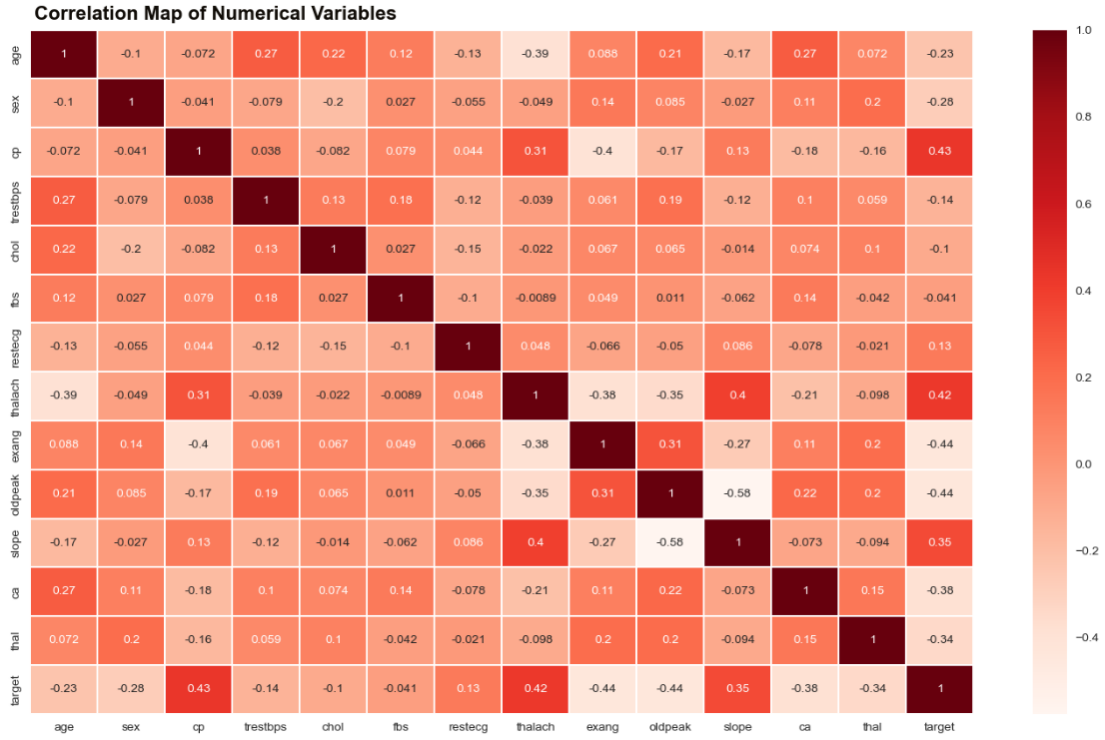

Figure 5: Correlation Matrix

4.2.4 MODEL SELECTION

It is the process to select one final algorithm for concerned purpose. It is decided by observing the accuracy by applying multiple algorithms. We can use logistic regression, XGBoost, KNN, random forest, etc. The final accuracy depends of the type of model we select.

While selecting the algorithm, we have to compare the accuracies.

Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction

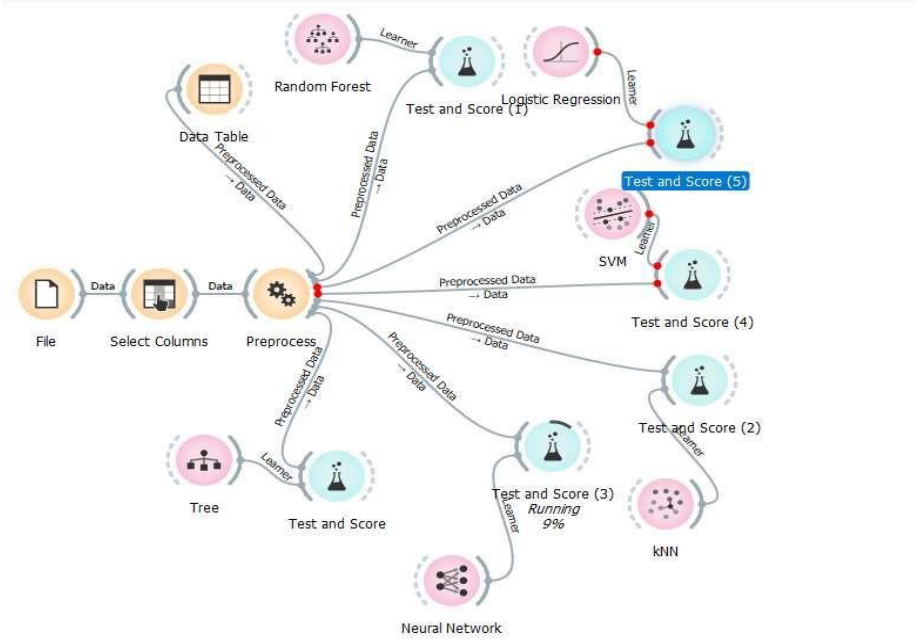

Figure 6: Connection of widgets in Orange

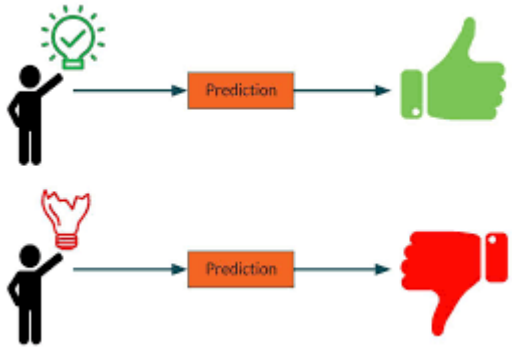

Figure 7: Prediction of Disease

In this project, we have compared following ML algorithms and obtained corresponding accuracies:

- XGBoost: 99.03 % Accuracy
- Random Forest: 96 % Accuracy
- KNN: 88.31 % Accuracy

CHAPTER 5
DEPLOYMENT

5. DEPLOYMENT

5.1 HARDWARE PLATFORM USED

The hardware requirement may serve as the basis for a contract for the implementation of the system and should therefore be complete and consistent in specification.

The hardware used for the system is mentioned below.

- PROCESSOR: Intel CORE i3 or above
- RAM: minimum 4.00GB
- HARD DISK: minimum 100GB

It should be noted that better the hardware facilities available, higher would-be response time of the system.

5.2 LIBRARIES AND SOFTWARE PLATFORM USED

The software requirement document is the specification of the system. The software requirement provides a basis for creating the software requirements specification.

OPERATING SYSTEM: Windows

SYSTEM TYPE: 64-bit, intel CORE i5

SOFTWARE: Jupyter Notebook, VS Code, Anaconda

TECHNOLOGIES: Python

LIBRARIES: Flask, pandas, NumPy, pickle, sklearn, xgboost etc

5.3 VISUALIZATION RESULTS

Based on the findings obtained from various algorithms used for identifying patients who have been diagnosed with heart disease, it is observed that KNN, Random Forest Classifier, and XGBoost have provided better results as compared to other techniques such as Logistic Regression, SVM and Decision Tree. These algorithms are not only accurate but more cost-effective and faster than the

algorithms used in previous research studies. The highest level of accuracy possible by Random Forest and XGBoost is either greater than or nearly equal to the accuracy that were obtained from earlier research studies. It can be inferred that the improvement in accuracy is due to the increased number of attributes used from the medical dataset that was used in the project.

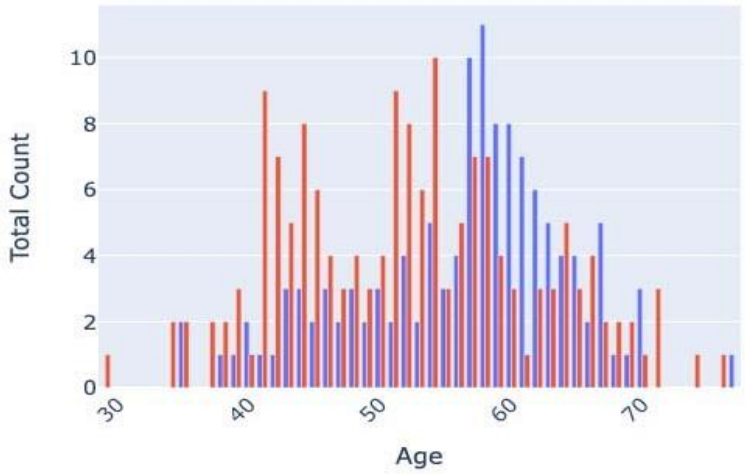

Figure 8: Shows the risk of heart attack based on age

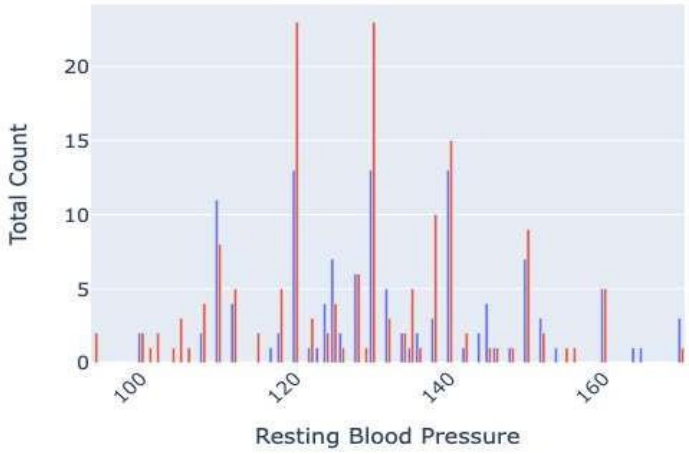

Figure 9: Shows the risk of heart attack based on restingbp

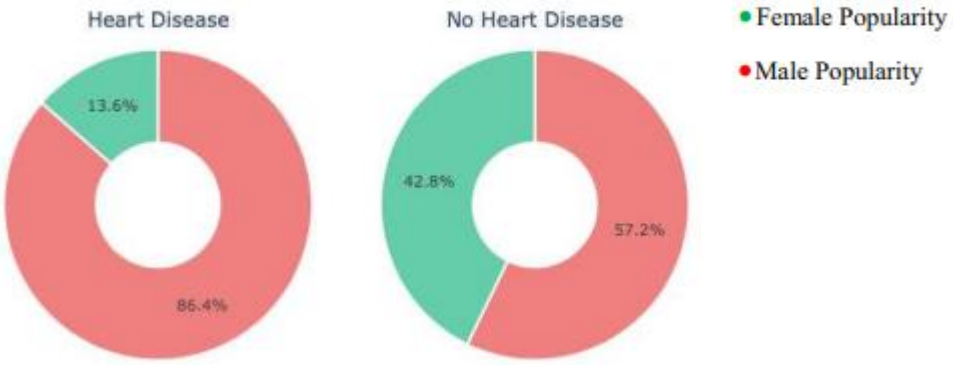

Figure 10: Shows the presence of heart attack based on Gender

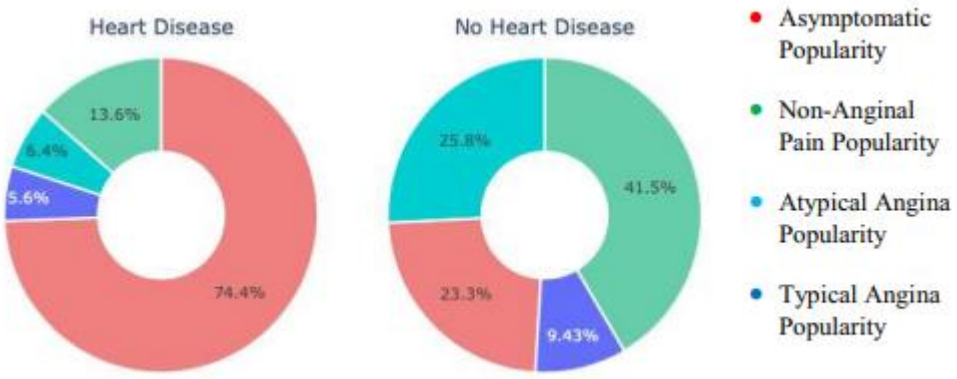

Figure 11: Shows presence of heart attack based on chestpain

Additionally, the study has revealed that XGBoost and Random Forest outperform KNN in the detection of patients who are diagnosed with the possibility of having a heart disease, indicating that XGBoost and Random Forest Classifier are more effective in diagnosing heart disease.

In this project, the data was formulated in different formulations and the model was trained using XGBoost tree algorithm with above 95% accuracy. The obtained results reveal that it is possible to achieve a high predictive accuracy.

5.4 DEPLOYMENT PROCESS

In this project, we have developed a model that can be utilized to create websites, web applications, or other customized forms based on the specific requirements of clients.

We will now proceed with building a web application using the provided dataset and incorporating the ML model. This deployment assumes that the user has a reasonable understanding of running Python code and is familiar with basic ML libraries such as Sci-kit Learn, Pandas, and NumPy.

We utilized **Flask**, a Python-based microframework, for developing small-scale websites. Flask is highly suitable for creating RESTful APIs using Python. Our current implementation includes a trained model i.e., **model.sav** capable of predicting the class of data based on various attributes. We have designed a web application where users can input attribute values, and the data will be passed to the model. Leveraging the training provided to the model, it will predict the heart disease of an individual based on the given details. We employed both the Random Forest model and Linear Regression to perform the sales prediction.

Flask script – Before starting with the coding part, we need to download flask and some other libraries. Here, we make use of virtual environment, where all the libraries are managed and makes the development job easier.

Here we import the libraries, then using app=Flask(name) we create an instance of flask. @app.route('/') is used to tell flask what URL should trigger the function index() and in the function index we use render_template('index.html') to display the script index.html in the browser.

This should run the application and launch a simple server. Open http://127.0.0.1:5000/ to see the html form.

HTML Form – In order to collect the data, we created html form which would contain all the different options to select from each attribute. Here, we have created a simple form using html.

When we click on the submit button in heartform.html, it predicts the salary for the values entered by the user, then passes on the variable `r_pred `outputted from the model and redirect to result.html template

5.5 RESULT

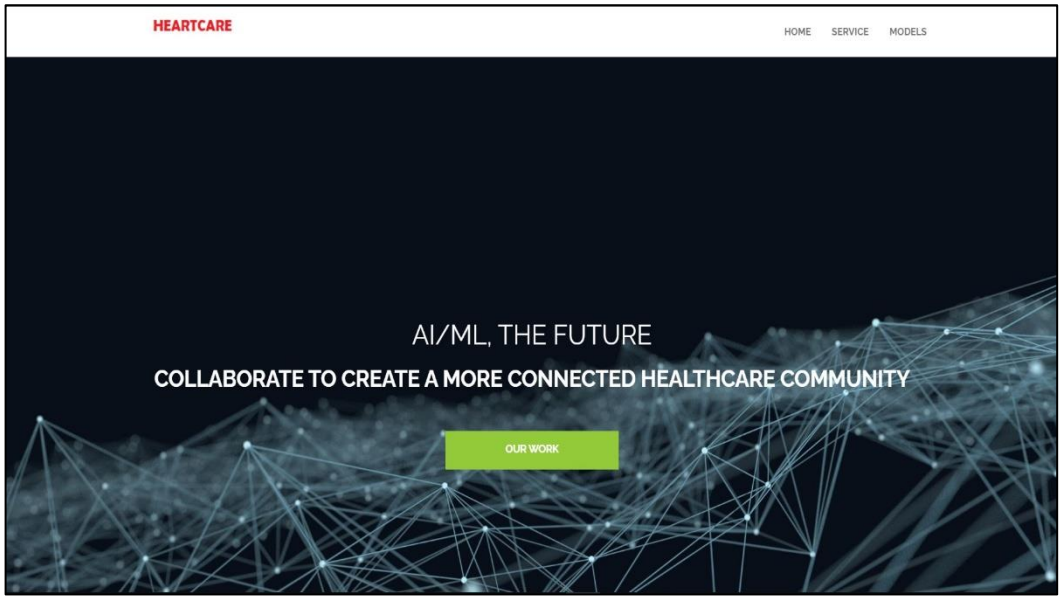

Figure 12: Homepage

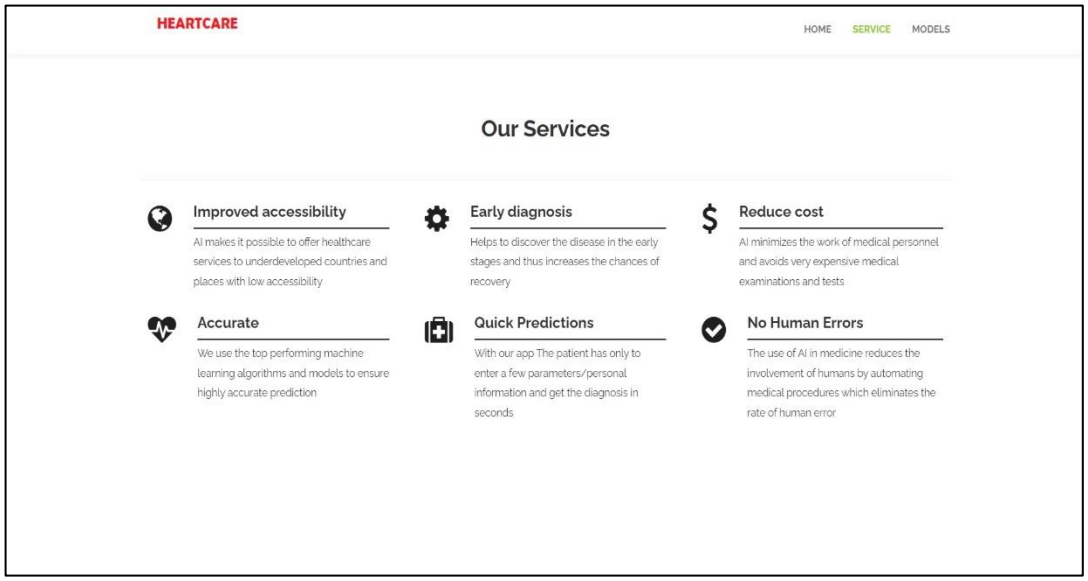

Figure 13: Our Services

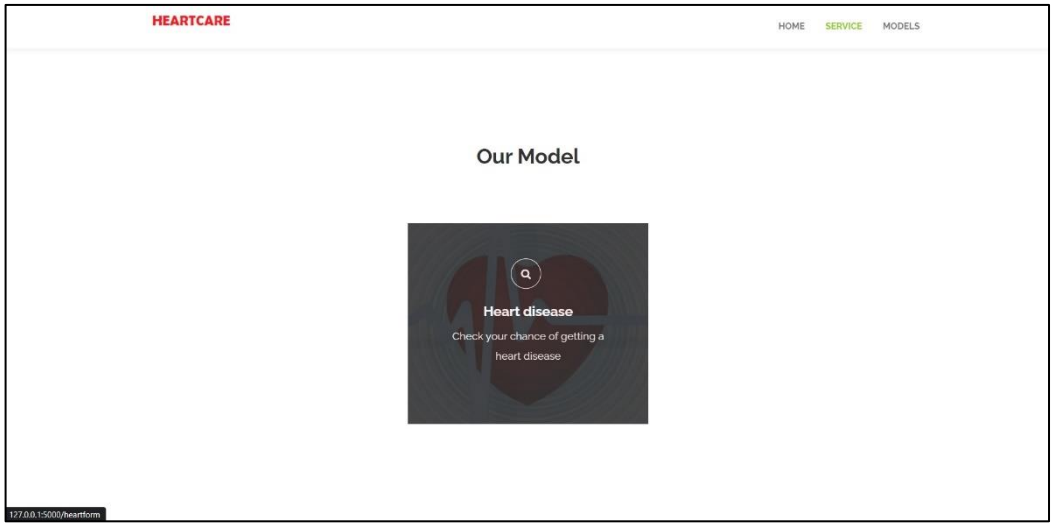

Figure 14: Heart Disease Model

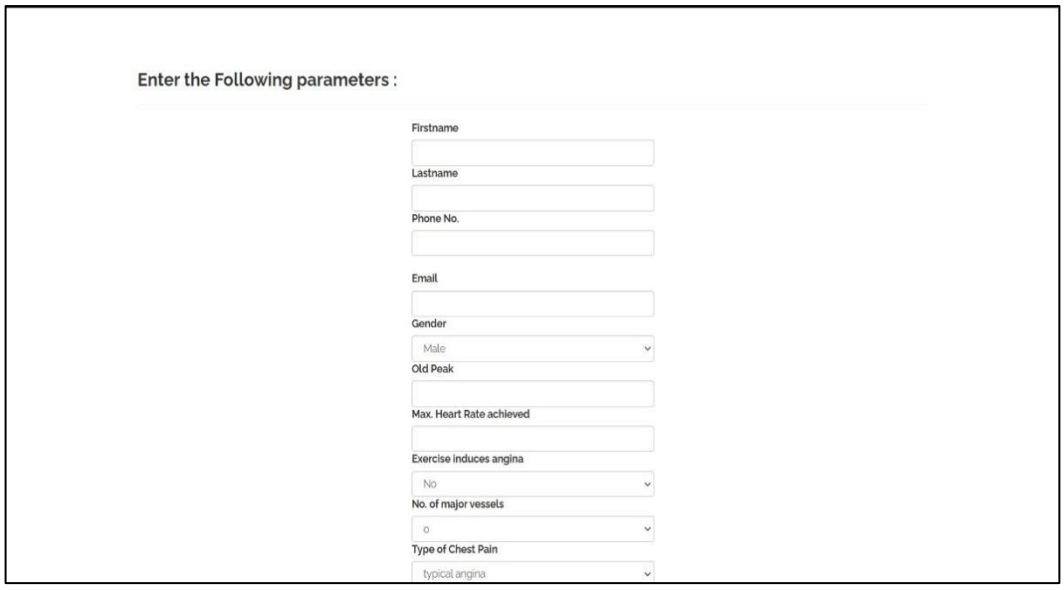

Figure 15: Heart Form

Figure 16: Detection

CHAPTER 6
CONCLUSION

6. CONCLUSION

Cardiovascular disease (CVD) is one of the leading cause of deaths happening worldwide, making early detection and intervention crucial for improving patient outcomes. To address this need, a machine learning technique were used to develop a model using patient medical history data to predict the probability of fatal heart disease. The dataset includes variables such as chest pain, sugar levels, and blood pressure, which are important indicators of heart health.

These classification algorithms - XGBoost, Random Forest Classifier, and KNN - were utilized to develop the model, which achieved an accuracy rate of over 95%. The accuracy of the model was further improved by increasing the size of the dataset, enabling the identification of more subtle patterns and risk factors.

The application of machine learning techniques in medical diagnosis has several benefits, including increased speed and accuracy of diagnoses, reduced costs, and improved patient outcomes. By analysing large amounts of data and identifying complex patterns, machine learning algorithms can provide valuable insights into patient health that may not be immediately apparent to human clinicians.

Compared to previous models, the accuracy of the developed model represents a significant improvement, with an accuracy rate of 98%. The XGBoost algorithm demonstrated the highest accuracy of 96% among the three algorithms used, indicating its effectiveness in predicting heart disease. The dataset used in this project indicates that 44% of individuals suffer from heart disease, highlighting the importance of early detection and intervention. The developed model offers a reliable and efficient method for identifying individuals who are at risk of heart disease, potentially benefiting both patients and healthcare providers.

CHAPTER 7
FUTURE SCOPE

7. FUTURE SCOPE

While the field of Educational Data Mining and Learning Analytics has already made great strides in improving education outcomes, there is still a lot of potential for further advancements. One area where this could be particularly impactful is in the detection and prevention of heart disease in students.

Heart disease is a leading cause of death worldwide, and early detection is key to successful treatment. By analysing data from student health records, as well as their physical activity levels and other relevant metrics, we can identify those who may be at risk for developing heart disease. Machine learning algorithms can be trained to recognize patterns and risk factors, and can provide early warnings to health professionals. Furthermore, with the rise of wearable technology and other health monitoring devices, we can collect even more data on students' health and activity levels. This data can be integrated with educational data, such as quiz scores and attendance records, to provide a more comprehensive picture of each student's overall health and wellbeing.

By using advanced analytics techniques to detect early warning signs of heart disease, educational institutions can play a vital role in promoting better health outcomes for their students. This will not only improve individual student outcomes, but can also have broader societal benefits by reducing healthcare costs and improving overall public health.

REFERENCES

1. Bo Jin , Chao Che,Zhen Liu ,Shulong Zhang ,Xiaomeng Yin And Xiaopeng Wei "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling". IEEE Access 2018.

2. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", "International Conference on "Computational Intelligence and Communication Technology" (CICT 2018).

3. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor. "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". IEEE Access (Volume: 7) 2019

4. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques". IEEE Access (Volume: 7) 2019.

5. K. Prasanna Lakshmi, Dr. C.R.K.Reddy. "Fast Rule-Based Heart Disease Prediction using Associative Classification Mining". International Conference on Computer, Communication and Control (IC4) 2015

6. M.Satish, D Sridhar, "Prediction of Heart Disease in Data Mining Technique", International Journal of Computer Trends & Technology (IJCTT), 2015.

7. Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.

8. Boshra Bahrami, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.

9. Mamatha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication and Signal Processing 2019.

10. Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

11. Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

12.Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

13. Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H(2016). Wireless

14.Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

15.Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office.

16.Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

DISSEMINATION OF WORK

PUBLICATION DETAILS

PAPER TITLE	CONFERENCE NAME	CONFERENCE DURATION	ISBN NUMBER
Heartcare: Heart Disease Detection using Machine Learning	International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)	April 23, 2023	2581-9429


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
M Areeb Ozair
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heartcare: Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 5, APRIL 2023
Certificate No: 042023-A1428
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9352
www.doi.org
www.crossref.org
7.301
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Suryakant Ingle
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heartcare: Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 5, APRIL 2023
Certificate No: 042023-A1429
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9352
www.doi.org
www.crossref.org
7.301
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Apeksha Mundhada
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heartcare: Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 5, APRIL 2023
Certificate No: 042023-A1431
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9352
www.doi.org
www.crossref.org
7.301
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Rutika Dharangaonkar
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Heartcare: Heart Disease Detection using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 5, APRIL 2023
Certificate No: 042023-A1430
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9352
www.doi.org
www.crossref.org
7.301
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief

PROJECT GROUP MEMBERS

Name: M Areeb Ozair
Address: Teachers Colony, Tayade Colony Road, Khamgaon
Email id: areebozair7@gmail.com
Mobile no: 8999964337

Name: Suryakant Ingle
Address: Gadge Nagar Hiwarkhed
Email id: suryakantingle04@gmail.com
Mobile no: 9545434038

Name: Rutika Dharangaonkar
Address: Civil Line, Khamgaon
Email id: ruteekadharangaonkar@gmail.com
Mobile no: 8766924091

Name: Apeksha Mundhada
Address: Civil Line, Kela Nagar, Khamgaon
Email id: apekshamundhada1234@gmail.com
Mobile no: 9370924376