Predictive analysis of cardiac diseases with machine learning techniques

Dakamari Jagadeep, Student, B. Sc. (AIML), Dept. of Computer Science, P.B. Siddhartha College of Arts & Science, AI Intern at Codegnan, Vijayawada, AP, India
B. V. Mikhil Pathro, Student, B. Sc. (AIML), Dept. of Computer Science, P.B. Siddhartha College of Arts & Science, AI Intern at Codegnan, Vijayawada, AP, India
Yedla Manoj Kumar, Student, B. Sc. (AIML), Dept. of Computer Science, P.B. Siddhartha College of Arts & Science, AI Intern at Codegnan, Vijayawada, AP, India
Jitendra Chautharia, M.Tech (CPS, EE, IIT Jodhpur), B.Tech (EEE, RTU Kota), AI Engineer at Codegnan IT Solutions, Vijayawada, AP, India.

Abstract:

Heart disease remains one of the leading causes of mortality worldwide, necessitating the development of effective predictive models to identify at-risk individuals and implement timely interventions. Advances in machine learning and artificial intelligence have revolutionized heart disease prediction by leveraging large datasets that include patient demographics, medical histories, lifestyle choices, and genetic information. These sophisticated algorithms detect patterns and correlations that are often missed by traditional statistical methods, thus improving the accuracy of risk assessments. By enabling early detection, these models play a critical role in preventative healthcare, helping to reduce the incidence and severity of heart disease.

We used different algorithms of machine learning such as logistic regression, random forest and Naïve Bayes to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using Naïve Bayes, Random Forest Classifier and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the 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.

Introduction:

Machine learning is a powerful tool that enables computers to learn from and make decisions based on data. Its applications span various fields, transforming industries by automating processes, improving predictions, and uncovering insights from vast datasets. However, successful implementation requires careful consideration of data quality, model selection, and ethical implications. We can use that knowledge in our project of HDP as it will help a lot of people.

Cardiovascular diseases are very common these days, they describe a range of conditions that could affect your heart. Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report. GENEVA, 20 May 2023 – Deaths from cardiovascular disease (CVD) jumped globally from 12.1 million in 1990 to 20.5 million in 2021, according to a new report from the World Heart Federation (WHF).

It is the primary reason of deaths in adults. Our project can help predict the people who are likely to diagnose with a heart disease by help of their medical history. It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly. This project focuses on mainly three data mining techniques namely: (1)

Logistic regression, (2) Naïve Bayes and (3) Random Forest Classifier. The accuracy of our project is 88.5% for which is better than previous system where only one data mining technique is used. So, using more data mining techniques increased the HDPS accuracy and efficiency. Out of the above three techniques Logistic regression given us the accurate and highest result in predicting of heart diseases in patients based on their health conditions from a data set.
This study focuses on predicting the likelihood of heart disease using logistic regression. The dataset, sourced from the Kaggle datasets, includes 303 patients with 13 medical attributes. These medical attributes are trained under three algorithms: Logistic regression, Naïve Bayes and Random Forest Classifier. Most efficient of these algorithms is Logistic Regression which gives us the accuracy of 88.54%. And, finally we classify patients that are at risk of getting a heart disease or not and also this method is totally cost efficient.

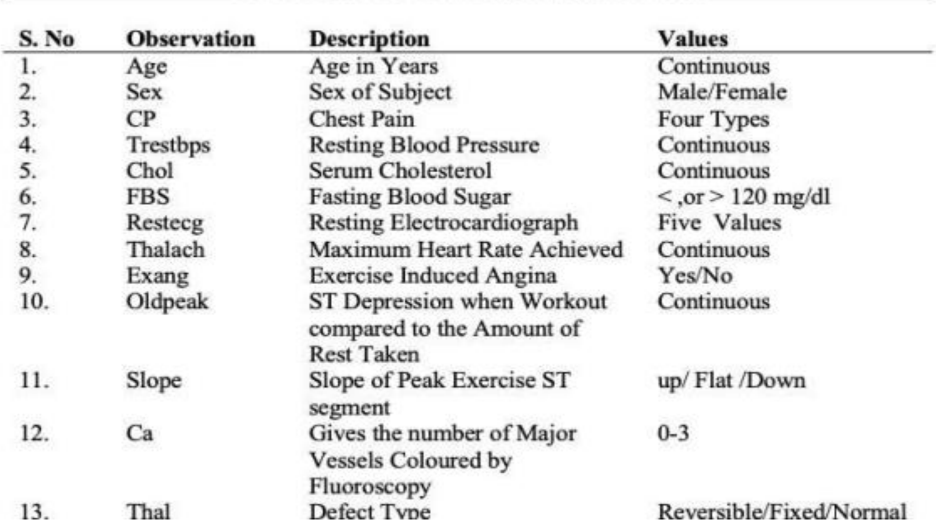
Prior Work:

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include *Logistic* Regression, Naïve Bayes and Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives.

Data Source:

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions. Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of 303 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 303 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the Kaggle datasets. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 13 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'.

Table 1. Various Attributes used are listed



S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal

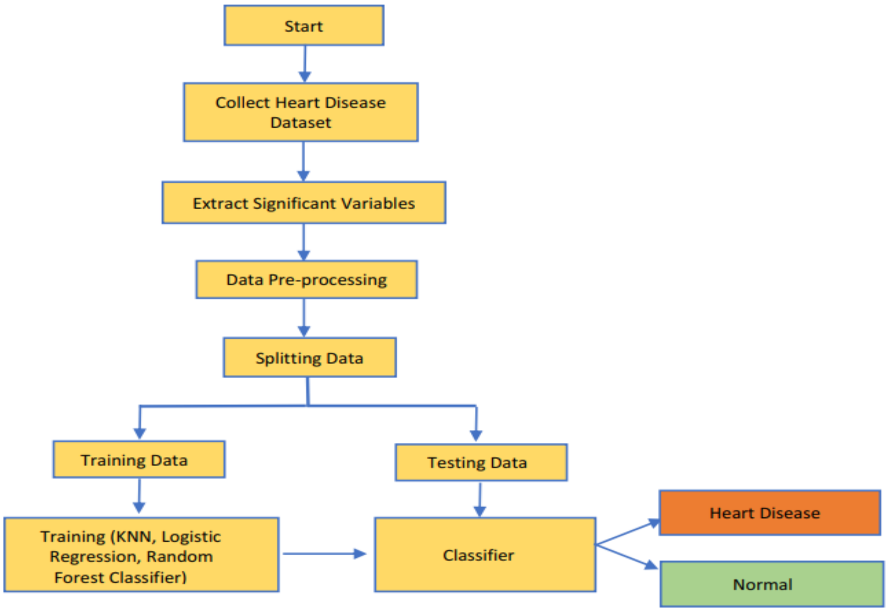
Methodology:

Machine Learning: Machine learning is like teaching a computer to learn from examples and patterns, rather than giving it explicit instructions for every task. Just like how we learn from experience, machine learning algorithms use data to recognize patterns and make predictions or decisions without being explicitly programmed for every scenario. It's like teaching a computer to think for itself by showing it lots of examples and letting it figure things out on its own.

And it has some algorithms which can perform classifications and predictions some of them are:

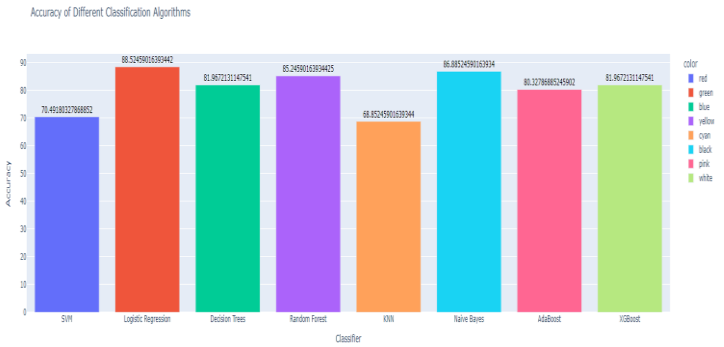
1. XGBoost is an efficient and scalable implementation of gradient boosting for classification and regression tasks. It builds a series of decision trees sequentially, where each subsequent tree corrects the errors made by the previous ones.
2. AdaBoost: AdaBoost (Adaptive Boosting) is a boosting algorithm that combines multiple weak learners (typically decision trees) to create a strong learner. It focuses more on the examples that are hard to classify and adjusts the weights of incorrectly classified examples in subsequent iterations.
3. Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features. Despite its simplicity, it is often effective for text classification and other tasks.
4. K-Nearest Neighbors (KNN): KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure, typically distance functions like Euclidean distance.
5. Random Forest: Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
6. Decision Tree: Decision Tree is a flowchart-like tree structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. It's a versatile algorithm used for classification and regression tasks.
7. Logistic Regression: Despite its name, logistic regression is a linear model for binary classification tasks. It models the probability that each input belongs to a particular class using the logistic (sigmoid) function.
8. Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm used for classification and regression tasks. It finds the optimal hyperplane that best separates the data points into different classes while maximizing the margin between classes.

Methodology gives a framework for the proposed model. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology in table 1 includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stage where we explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are Logistic Regression, Naïve Bayes and Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction.


Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Splitting Data
Training Data
Testing Data
Training (KNN, Logistic Regression, Random Forest Classifier)
Classifier
Heart Disease
Normal

Results & Discussions:

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, Logistic Regression, Random Forest Classifier and Naïve Bayes yield a better result to out rule them. The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by Logistic Regression and Naïve Bayes are about the percentage of 88.5% and 86.8% which is greater or almost equal to accuracies obtained from previous researches. So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and Naïve Bayes outperforms Random Forest Classifier in the prediction of the patient diagnosed with a Heart Disease. This proves that Logistic Regression and Naïve Bayes are better in diagnosis of a heart disease. The following bar graph shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain etc.



The above Bar graph shows the accuracy of the patients of having Heart disease or not by using machine learning algorithms. Out of those algorithms Logistic Regression, Naïve Bayes and Random Forest gives the highest rate of accuracy of having heart disease or not based on the taken data set of 303 patients for 13 different attributes. The highest accuracy resulted Logistic Regression is of 88.5%, now the second highest shown is Naïve Bayes is of 86.8% and third highest accuracy given is Random Forest is of 85.2%. Above all other machine learning algorithms these three performed accurate results for having Cardiovascular heart disease or not.

Deployment and User Interface:

To make our research findings and analytical tools accessible to a wider audience, we developed a user-friendly web application using Streamlit, a Python library for building Interactive web applications for data science and machine learning projects.

User Interface: The user interface (UI) of our web application was designed with simplicity and functionality in mind, ensuring that users can easily navigate through different features and analyses.

Functionality and features:

1 Name of the patient**.**

2.Age of the patient.

3.Gender of the patient.

4.cp: Refers to chest pain type. It's categorized into four types:
Typical angina, Atypical angina, Non-anginal pain and Asymptomatic

5.Trestbps: Denotes the resting blood pressure (in mm Hg) of the patient upon admission to the hospital.

6.chol: Stands for serum cholesterol level (in mg/dl).

7.fbs: Represents the fasting blood sugar level (> 120 mg/dl is considered as 1 and <= 120 mg/dl is considered as 0).

8.restecg: Refers to resting electrocardiographic results. It categorizes into three types: 0: Normal 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria.

9.thalach: Represents the maximum heart rate achieved.

10.exang: Indicates exercise-induced angina (1 = yes, 0 = no).

11.oldpeak: Refers to the ST depression induced by exercise relative to rest.

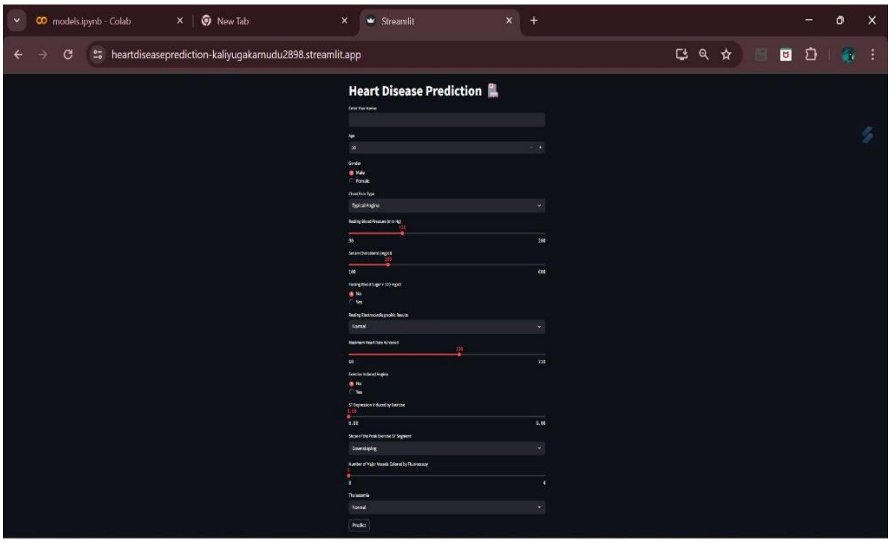
12.slope: Denotes the slope of the peak exercise ST segment. It's categorized into three types: 0: Upsloping 1: Flat 2: Downsloping

13.ca: Represents the number of major vessels (0-3) colored by fluoroscopy.

14.thal: Refers to a blood disorder called thalassemia. It's categorized into three types: 1: Normal 2: Fixed defect 3: Reversible defect

Deployment:

The web application was deployed using Streamlit's built-in functionality for deploying and sharing data science applications. The deployment process involved packaging the application code into a standalone Python script and deploying it to a cloud-based server. Users can access the application through a web browser, eliminating the need for local installation or setup.



Heart Disease Prediction

Enter Your Name:

Age
50

Gender
● Male
○ Female

Chest Pain Type
Typical Angina

Resting Blood Pressure (mm Hg)
120
90 200

Serum Cholesterol (mg/dl)
200
100 600

Fasting Blood Sugar > 120 mg/dl
● No
○ Yes

Resting Electrocardiographic Results
Normal

Maximum Heart Rate Achieved
150
60 220

Exercise Induced Angina
● No
○ Yes

ST Depression Induced by Exercise
1.00
0.00 6.00

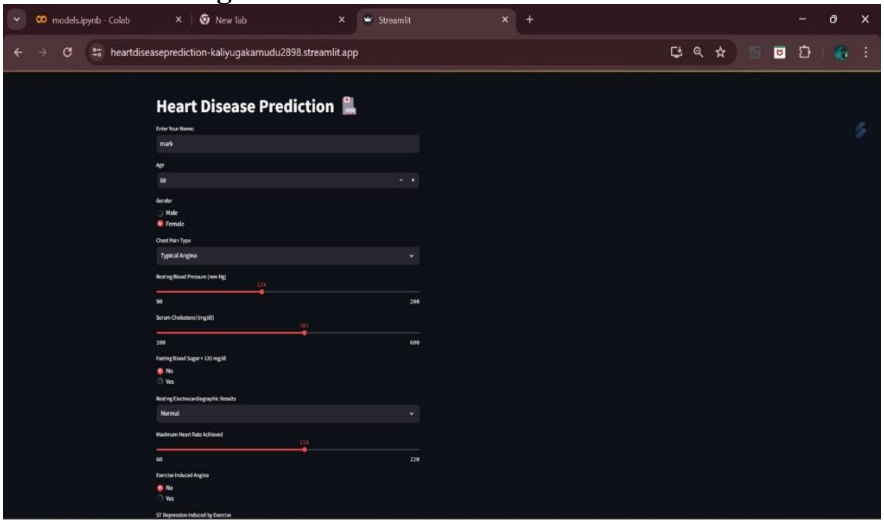
Slope of the Peak Exercise ST Segment
Downsloping

Number of Major Vessels Colored by Fluoroscopy
0
0 4

Thalassemia
Normal

Predict

Fig-1: It shows the interface of home.



Heart Disease Prediction

Enter Your Name:
mark

Age
60

Gender
○ Male
● Female

Chest Pain Type
Typical Angina

Resting Blood Pressure (mm Hg)
134
90 200

Serum Cholesterol (mg/dl)
381
100 600

Fasting Blood Sugar > 120 mg/dl
● No
○ Yes

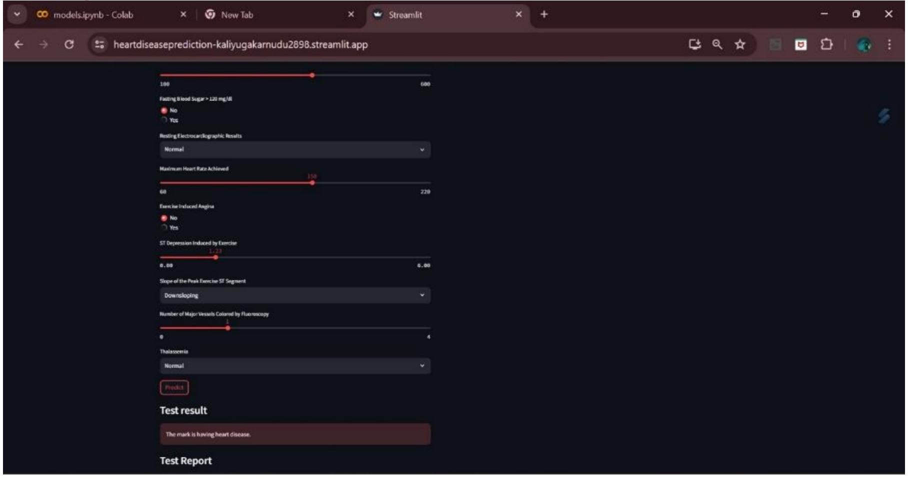
Resting Electrocardiographic Results
Normal

Maximum Heart Rate Achieved
158
60 220

Exercise Induced Angina
● No
○ Yes

ST Depression Induced by Exercise

Fig-2: For finding that patient is having heart disease or not.



100 600

Fasting Blood Sugar > 120 mg/dl
● No
○ Yes

Resting Electrocardiographic Results
Normal

Maximum Heart Rate Achieved
158
60 220

Exercise Induced Angina
● No
○ Yes

ST Depression Induced by Exercise
1.23
0.00 6.00

Slope of the Peak Exercise ST Segment
Downsloping

Number of Major Vessels Colored by Fluoroscopy
1
0 4

Thalassemia
Normal

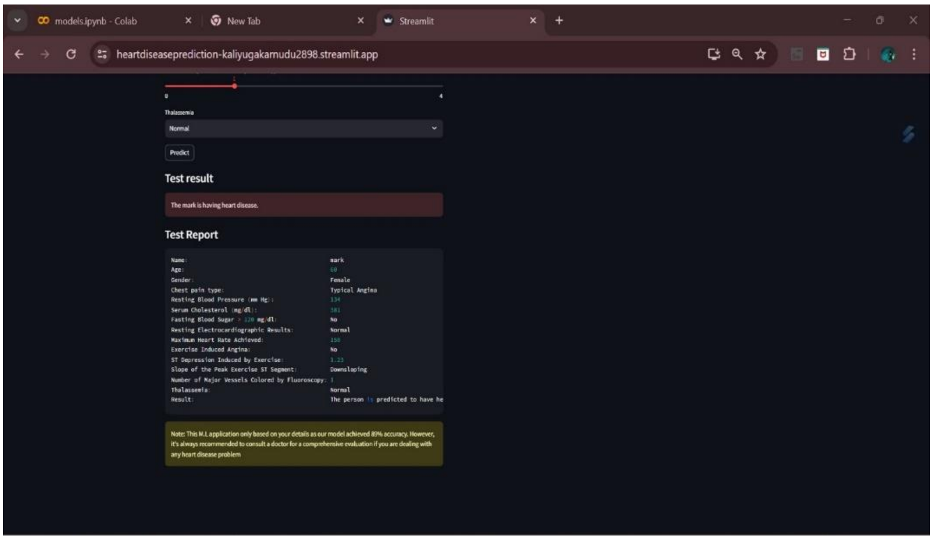
Predict

Test result

The mark is having heart disease.

Test Report

Fig-3: Entering the data of the patient.



Thalassemia

Normal

Test result

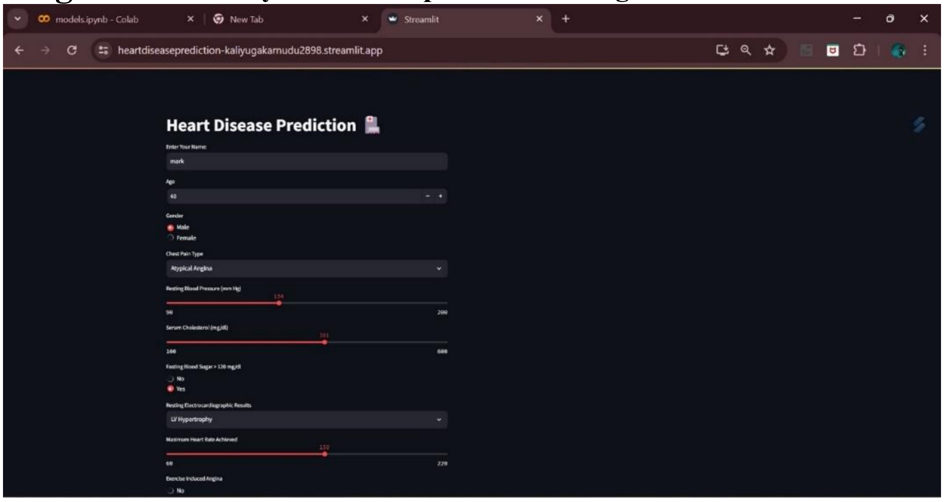
The mark is having heart disease.

Test Report

```
Name:                                          mark
Age:                                           60
Gender:                                        Female
Chest pain type:                               Typical Angina
Resting Blood Pressure (mm Hg):                134
Serum Cholesterol (mg/dl):                     381
Fasting Blood Sugar > 120 mg/dl:               No
Resting Electrocardiographic Results:          Normal
Maximum Heart Rate Achieved:                   158
Exercise Induced Angina:                       No
ST Depression Induced by Exercise:             1.23
Slope of the Peak Exercise ST Segment:         Downsloping
Number of Major Vessels Colored by Fluoroscopy: 1
Thalassemia:                                   Normal
Result:                                        The person is predicted to have he
```

Note: This M.L application only based on your details as our model achieved 83% accuracy. However, it's always recommended to consult a doctor for a comprehensive evaluation if you are dealing with any heart disease problem

Fig-4: **Successfully found that patient is having heart disease.**



Heart Disease Prediction

Enter Your Name:
mark

Age
60

Gender
● Male
○ Female

Chest Pain Type
Atypical Angina

Resting Blood Pressure (mm Hg)

Serum Cholesterol (mg/dl)

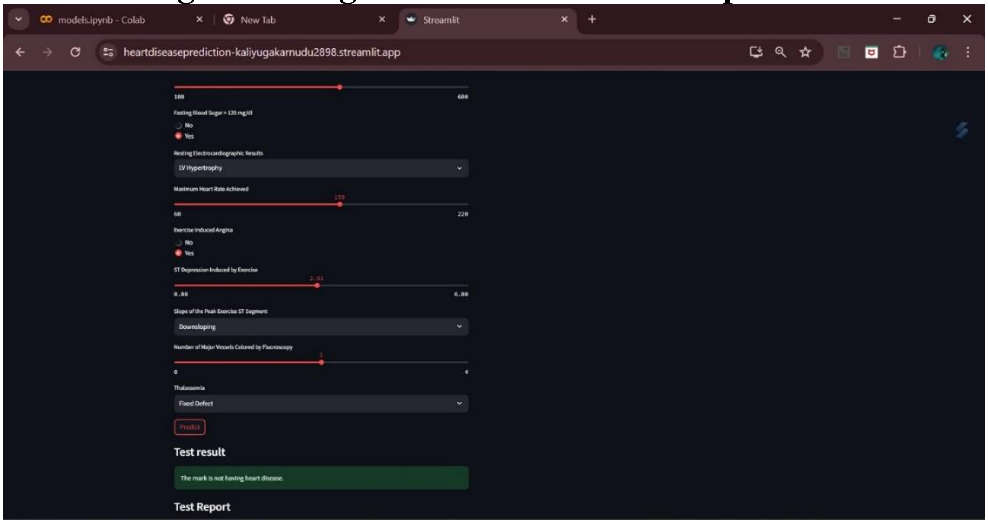
Fasting Blood Sugar > 120 mg/dl
○ No
● Yes

Resting Electrocardiographic Results
LV Hypertrophy

Maximum Heart Rate Achieved

Exercise Induced Angina
○ No

Fig-5: Taking the data of the second patient.



Fasting Blood Sugar > 120 mg/dl
○ No
● Yes

Resting Electrocardiographic Results
LV Hypertrophy

Maximum Heart Rate Achieved

Exercise Induced Angina
○ No
● Yes

ST Depression Induced by Exercise

Slope of the Peak Exercise ST Segment
Downsloping

Number of Major Vessels Colored by Fluoroscopy

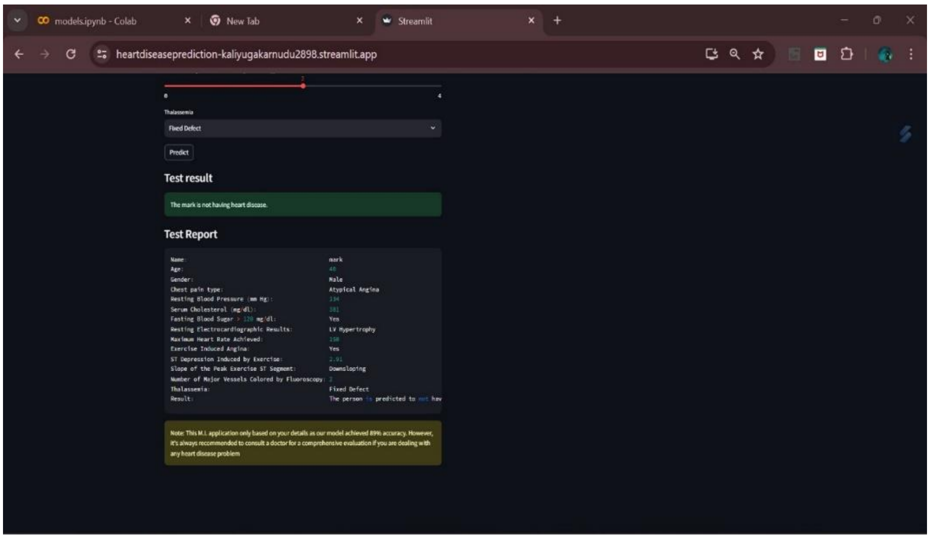
Thalassemia
Fixed Defect

Test result

The mark is not having heart disease.

Test Report

Fig-6: Entering the data of the patient.



Test result

The mark is not having heart disease.

Test Report

```
Name:                                         mark
Age:                                          40
Gender:                                       Male
Chest pain type:                              Atypical Angina
Resting Blood Pressure (mm Hg):               134
Serum Cholesterol (mg/dl):                    381
Fasting Blood Sugar > 120 mg/dl:              Yes
Resting Electrocardiographic Results:         LV Hypertrophy
Maximum Heart Rate Achieved:                  158
Exercise Induced Angina:                      Yes
ST Depression Induced by Exercise:            2.91
Slope of the Peak Exercise ST Segment:        Downsloping
Number of Major Vessels Colored by Fluoroscopy: 2
Thalassemia:                                  Fixed Defect
Result:                                       The person is predicted to not hav
```

Note: This M.L application only based on your details as our model achieved 89% accuracy. However, it's always recommended to consult a doctor for a comprehensive evaluation if you are dealing with any heart disease problem

Fig-7: Successfully found that patient is not having heart disease.

Conclusion:

A cardiovascular disease detection model has been developed using three ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and Naïve Bayes. The accuracy of our model is 88.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and Naïve Bayes to get an accuracy of an average of 88.5% and 86.2% on our model. Also, it is concluded that accuracy of Logistic Regression is highest between the three algorithms that we have used i.e. 88.52%.

References:

1. "Development and Validation of a Machine Learning Model for Predicting Risk of Cardiovascular Disease" by Krittanawong, C., Zhang, H. J., Wang, Z., Aydar, M., Kitai, T., Baber, U., Min, J. K., & Tang, W. H. (2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10577538
2. "A Comparative Study on Machine Learning Algorithms for Heart Disease Prediction" by Padma, A., & Geetha, T. V. (2017). https://www.researchgate.net/publication/348064698
3. "Heart Disease Prediction Using Machine Learning Algorithms" by Kumar, P., & Srivastava, M. (2020). https://www.researchgate.net/publication/342405070
4. "Efficient heart disease prediction system", by K Saxena, R Sharma-(2016). https://www.sciencedirect.com/science/article/pii/S187705091630638X
5. "Machine learning techniques for heart disease prediction: a comparative study and analysis", by R Katarya, SK Meena- Health and Technology, 2021. https://link.springer.com/article/10.1007/s12553-020-00505-7