KONKAN GYANPEETH COLLEGE OF ENGINEERING, KARJAT
Affiliated to University of Mumbai, Approved by AICTE, New Delhi.

Predicting Heart Disease using Machine Learning

Presented by :

- Shubham Ingle(Roll No. 20)
- Sarthak Jadhav(Roll No. 23)
- Sushant Kumar Mishra(Roll No. 40)

Guided by :

Prof. Poonam B. Lad

Index

1. Abstract
2. Introduction
3. Literature Survey
4. Problem statement
5. Implemented system
6. Work Flow
7. Process Flow
8. Objectives
9. Requirements
10. Parameters Of Dataset
11. User Interface Design
12. Output
13. Scope
14. Conclusion
15. Refrences

Abstract

- Predicting and detection of heart disease has always been a critical and challenging task for healthcare practitioners. Hospitals and other clinics are offering expensive therapies and operations to treat heart diseases. So, predicting heart disease at the early stages will be useful to the people around the world so that they will take necessary actions before getting severe. Over the years, machine learning shows effective results in making decisions and predictions from the broad set of data produced by the health care industry. Some of the supervised machine learning techniques used in this prediction of heart disease are artificial neural network (ANN), decision tree (DT), random forest (RF),support vector machine (SVM), naïve Bayes) (NB) and knearest neighbour algorithm. Furthermore, the performances of these algorithms are summarized.

Introduction

- Heart disease is the leading cause of death among all others diseases ,even cancers.
- One in 4 deaths in India are now because of CVDs with ischemic heart disease and stroke.
- The diagnosis is often made, based on doctor's intutions & experience ,this may lead to an unwanted result & excessive medical cost.
- Therefore , an automic medical daignosis system would exceedingly benificial.
- This is where data mining and machine learning technniques has recognised to be extremely effective.

Literature Survey

- Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

AuthorJK	Title	Year	Method	Description
Liaqat Ali et al.	"An Automated Diagnostic System for Heart Disease Prediction Based on χ2 Statistical Model and Optimally Configured Deep Neural Network,"	2019	X^2 statistical model, deep neural network	• LIAQAT ALI et al. recommended a model which is consists of two methods one is X^2 statistical and second is deep neural network(DNN). • Feature refinement is done by X^2 statistical model and classification is done by a deep neural network(DNN). • In their study, they have used the Cleveland dataset. There are 303 instances in that dataset, among them,297 have no missing data, and the remaining 6 have missing data. • Among 297, 207 instances are used for training data, and the remaining 90 are used as testing data. • This model gives better results compared to conventional ANN models which are present earlier . • As a result of this using this proposed model, they have got 93.33% classification accuracy using DNN. It is 3.33% more than that of the conventional ANN model.
Dr. Kanak Saxena et.al	"Efficient Heart Disease Prediction System,"	2016	Decision tree	• Dr. kanak Saxena et al. developed a data mining model to predict heart disease efficiently. • It mainly helps the medical practitioners to make efficient decisions way based on the given parameters. • The author has used Cleveland dataset from UCI, and they have used age, sex, resting blood pressure, chest pain, serum cholesterol, fasting blood sugar , etc. as attributes. • Furthermore, they have divided the data sets into two parts one is for testing, and the other one is for training. They have used a 10-fold method to find accuracy.

Author	Title	Year	Method	Description
Awais Nimat et al.	"An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure,"	2019	Support vector machine, Hybrid grid search algorithm (HGSA)	• AWAIS NIMAT et al. proposed an expert system based on two support vector machines(SVM) to predict heart disease efficiently. • These two SVM's have their purpose; first, one is used to remove the unnecessary features, and the second one is used for prediction. • Moreover, they have used the HGSA (hybrid gird search algorithm) to optimize the two methods. • By using this model, they have achieved 3.3% better accuracy than the conventional SVM models that are present earlier.

Author	Title	Year	Method	Description
Ashir Javeed et al.	"An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection,"	2019	Random search algorithm (RSA), Random forest.	• Ashir Javeed et al. developed a model to improve the prediction of heart disease by overcoming the problem of overfitting; overfitting means the proposed model performs and gives better accuracy on testing data and gives unfortunate accuracy result for training data while predicting the heart disease. • To solve this problem, they have developed a model that will give the best accuracy on both training and testing data. • That model consists of two algorithms one is RSA(Random search algorithm) other one is a random forest algorithm that is used to predict the model. • This proposed model gave them better results in training data as well as testing data.

Problem statement

- Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive. The overall objective of our work will be to predict the presence of heart disease accurately with the help of tests and attributes . Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

Implemented system

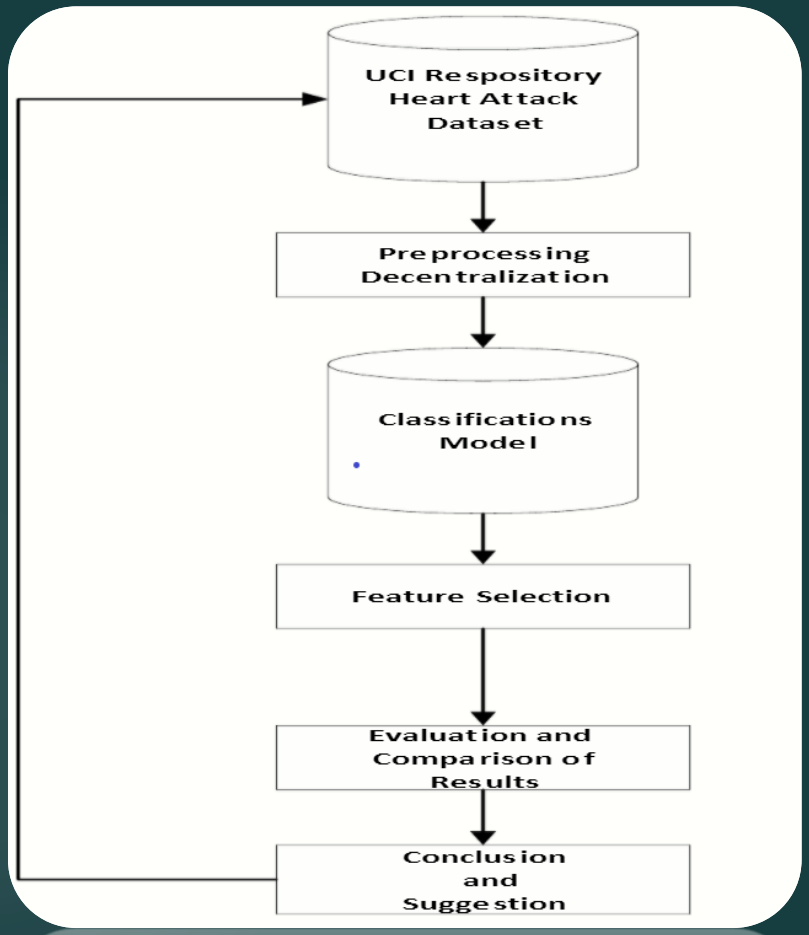
- In this system we are implementing effective heart disease prediction system . We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input. After accessing data set the operation is performed and effective heart disease level is produced.
- The implemented system will add some more parameters significant to heart disease with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart disease prediction system designed to help the identify different risk levels of heart disease like normal, low or high and also giving the prescription details with related to the predicted result.

Work Flow

We'll approach the problem with the following machine learning modelling framework

- **Exploratory data analysis (EDA)** - the process of going through a dataset and finding out more about it.
- **Model training** - create model(s) to learn to predict a target variable based on other variables.
- **Model evaluation** - evaluating a models predictions using problem-specific evaluation metrics.
- **Model comparison** - comparing several different models to find the best one.
- **Model fine-tuning** - once we've found a good model, how can we improve it?
- **Feature importance** - since we're predicting the presence of heart disease, are there some things which are more important for prediction?
- **Cross-validation** - if we do build a good model, can we be sure it will work on unseen data?
- **Reporting what we've found** - if we had to present our work, what would we show someone?

Process Flow:



objectives

- The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set.
- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- Reduce the cost of medical tests.

Requirements

- Hardware Requirements :

  Ram 1GB

  Hard disk 500GB

- Software Requirements :

  Anaconda IDE

  Python 3.8 with libraries for ML Learning.

  Operating System: Windows or any equivalent OS.

Parameters of dataset

Features are different parts of the data.

One of the most common ways to do this, is to create a **data dictionary**.

Heart Disease Data Dictionary:A data dictionary describes the data we're dealing with. Not all datasets come with them so this is where we may have to do your research or ask a **subject matter expert** for more.

The following are the features we'll use to predict our target variable (heart disease or no heart disease)

1.age - age in years
2.sex - (1 = male; 0 = female)
3.cp - chest pain type
- 0: Typical angina: chest pain related decrease blood supply to the heart
- 1: Atypical angina: chest pain not related to heart
- 2: Non-anginal pain: typically esophageal spasms (non heart related)
- 3: Asymptomatic: chest pain not showing signs of disease

4.trestbps - resting blood pressure (in mm Hg on admission to the hospital)
1. anything above 130-140 is typically cause for concern

5.chol - serum cholesterol in mg/dl
- serum = LDL + HDL + .2 * triglycerides
- above 200 is cause for concern

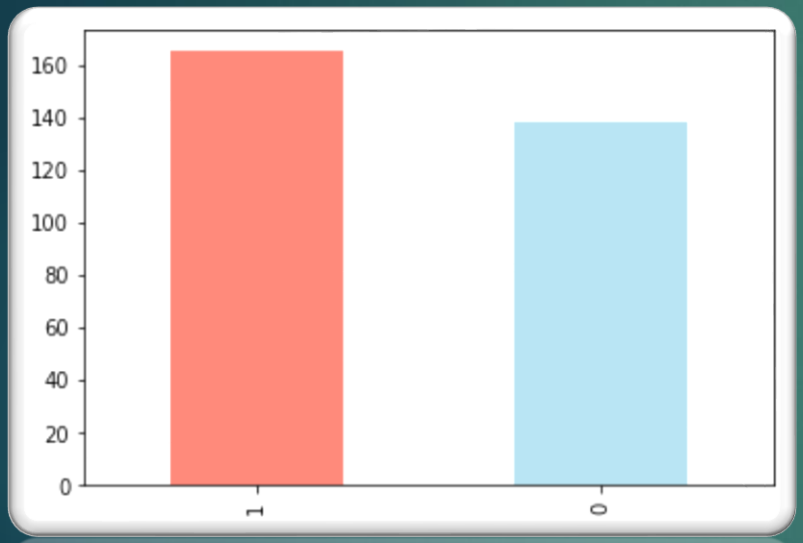
6.fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
1. '>126' mg/dL signals diabetes

7.restecg - resting electrocardiographic results
- 0: Nothing to note
- 1: ST-T Wave abnormality
  - can range from mild symptoms to severe problems
  - signals non-normal heart beat
  - 2: Possible or definite left ventricular hypertrophy
  - Enlarged heart's main pumping chamber

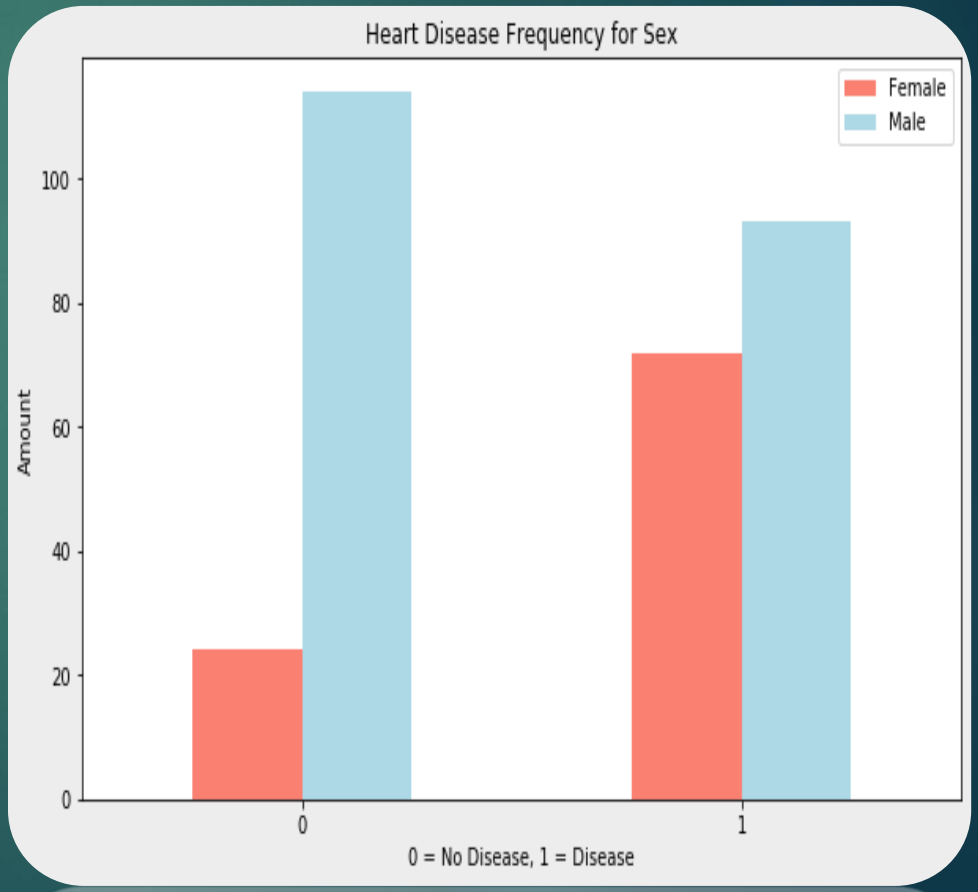
8. exang - exercise induced angina (1 = yes; 0 = no)
9.oldpeak - ST depression induced by exercise relative to rest
10.thalach - maximum heart rate achieved
    - looks at stress of heart during exercise
    - unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with exercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by fluoroscopy
    - colored vessel means the doctor can see the blood passing through
    - the more blood movement the better (no clots)
13. thal - thallium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversable defect: no proper blood movement when exercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

EDA(Exploratory data analysis)

Positive and negative samples

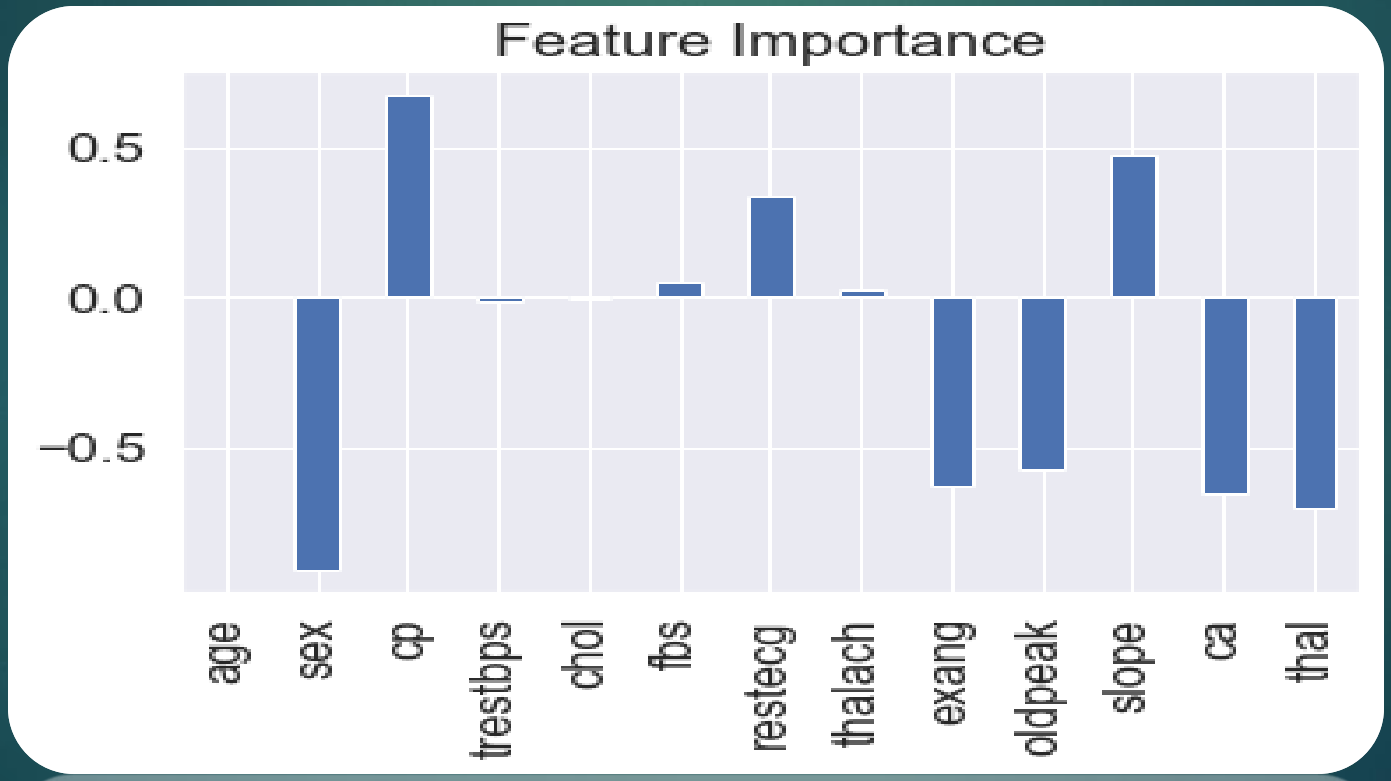


Heart disease frequency according to sex



Feature Importance

“features contributing most to the outcomes of the model”

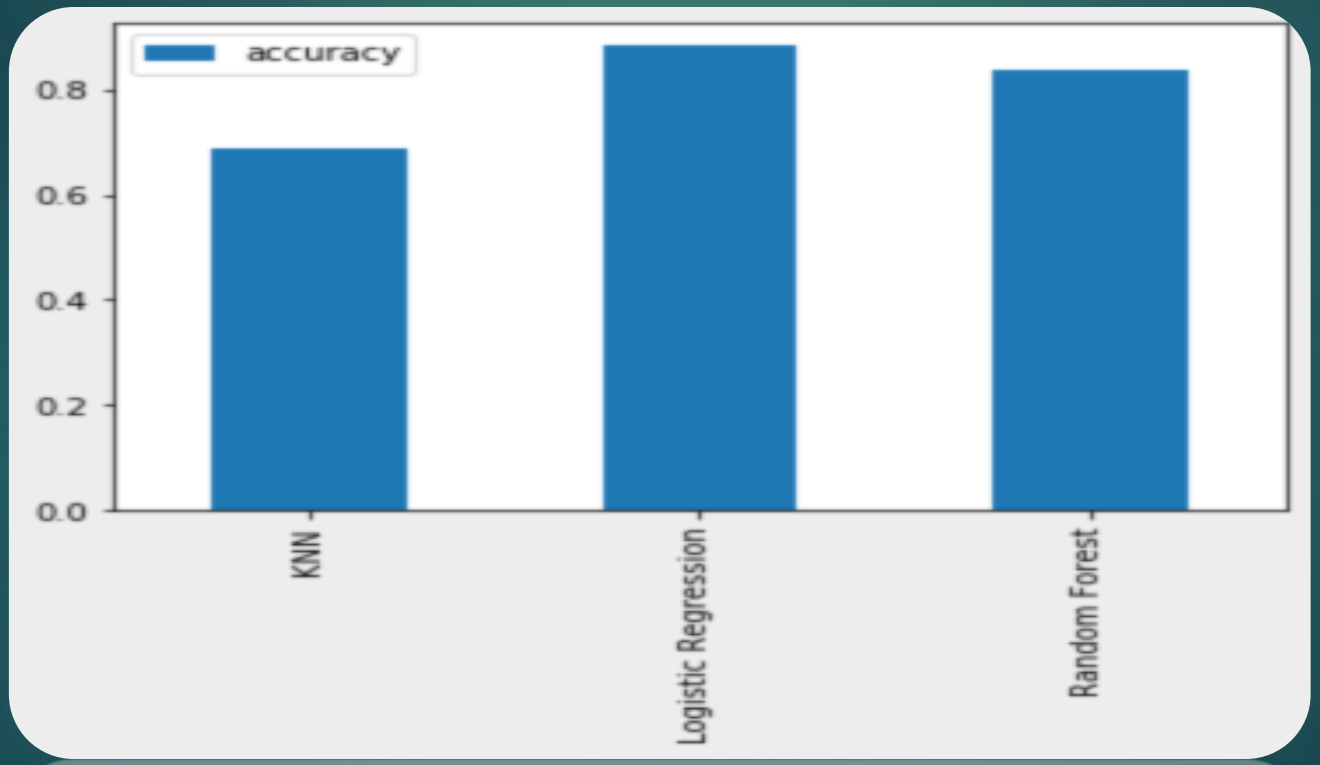


MACHINE LEARNING ALGORITHMS

- Logistic Regression
- K-Nearest Neighbors Classifier
- Random Forest Classifier

Models comparison

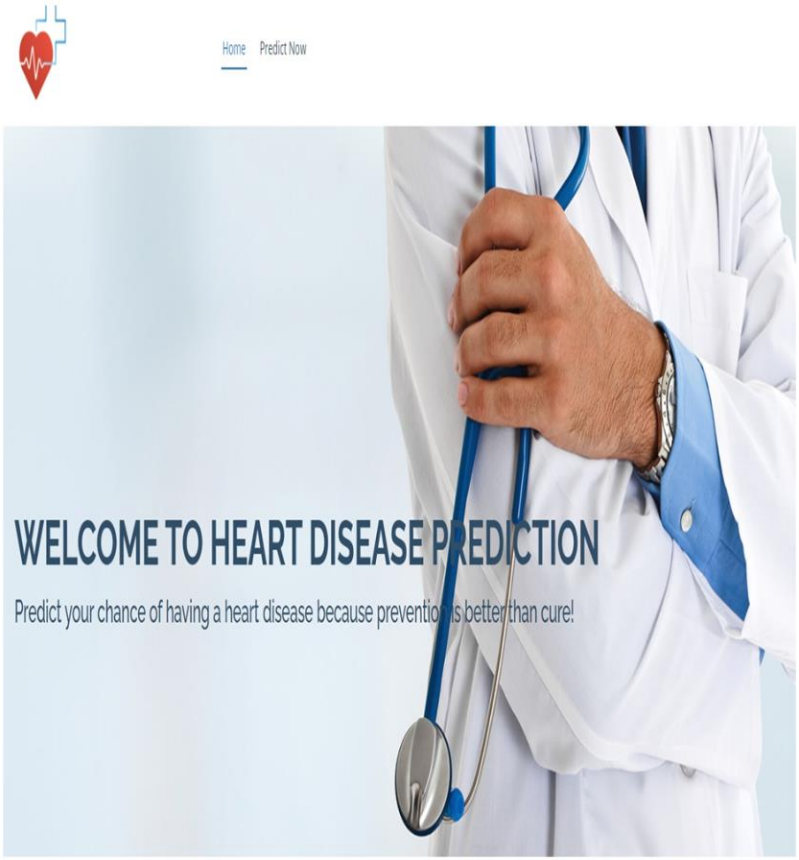
All of the algorithms in the Scikit-Learn library use the same functions, for training a model, model.fit(X_train, y_train) and for scoring a model model.score(X_test, y_test). score() returns the ratio of correct predictions (1.0 = 100% correct)

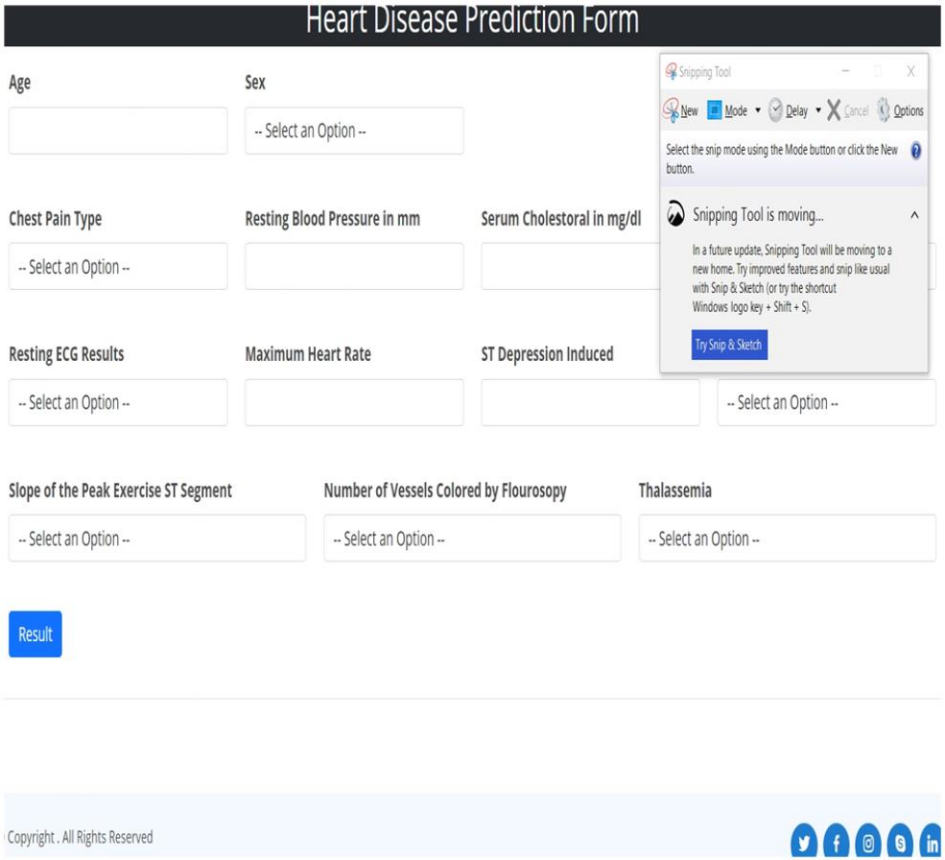


User interface design

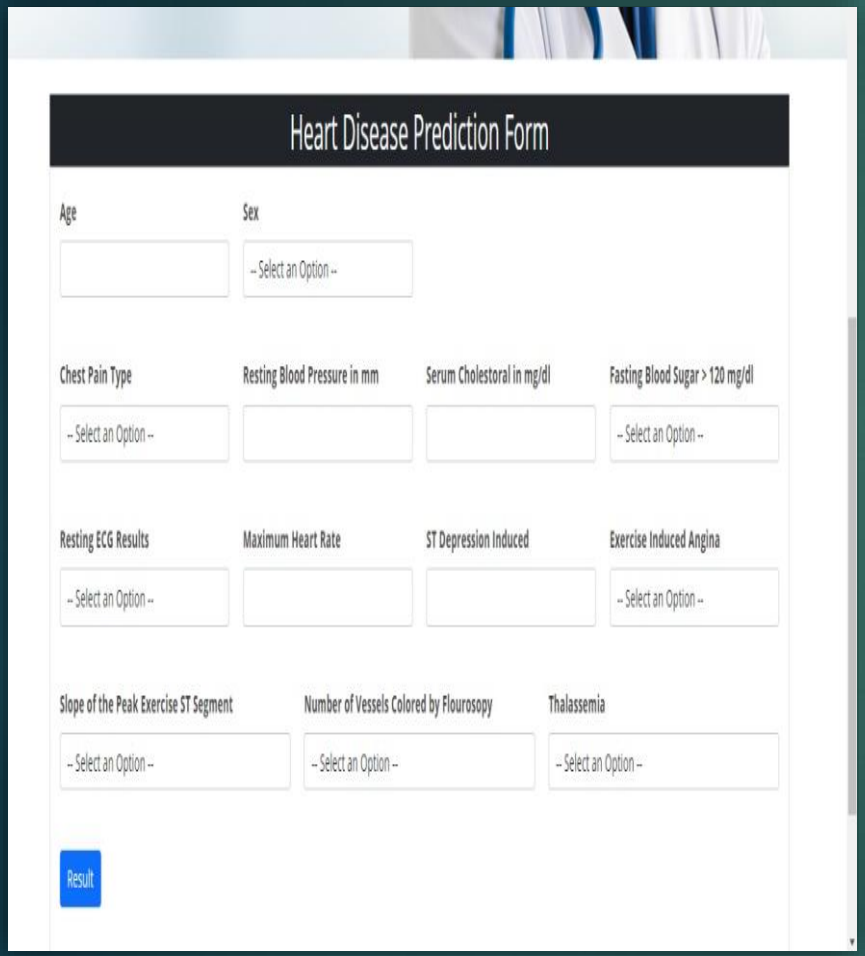
we have made a simple user interface design for users who want to know if they have chances of getting a heart disease or not for this user needs to perform some basic laboratory tests to fill the necessary input in the heart disease prediction form. The form has 13 inputs for the 13 features and a button. The button sends POST request to the/predict endpoint with the input data. In the form tag, the action attribute calls predict function when the form is submitted. Finally, the HTML page presents the stored result in the result parameter for users.

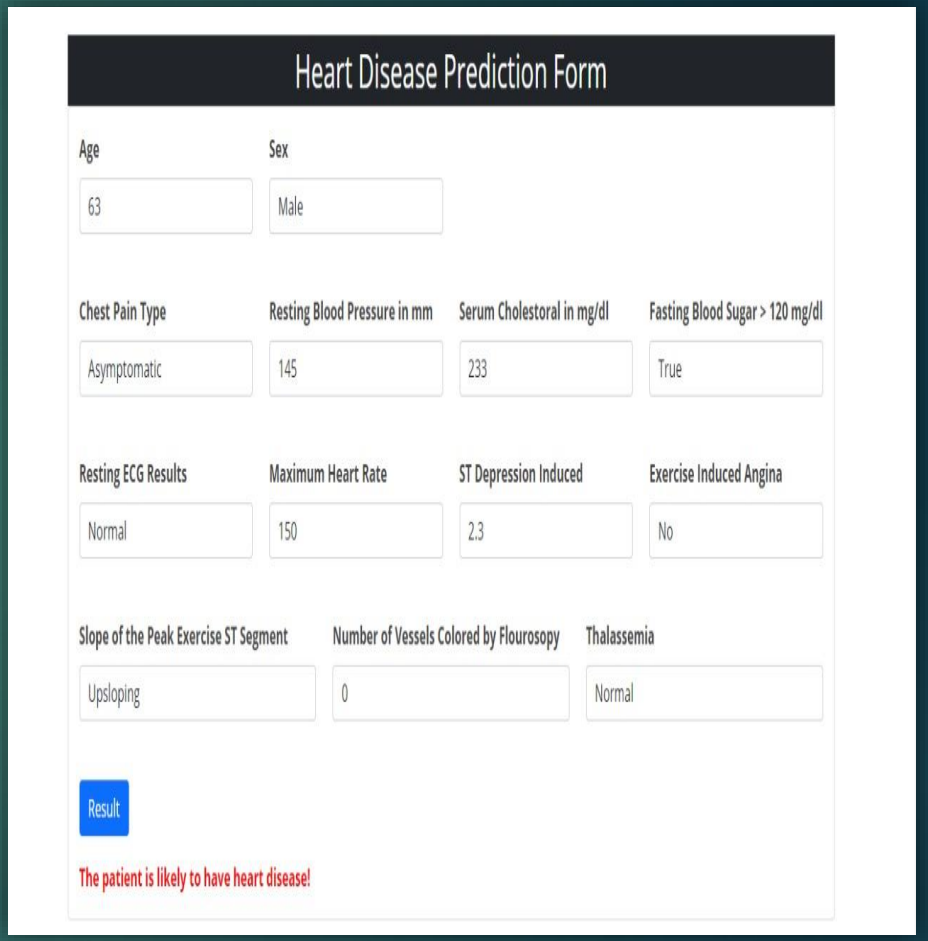
USER INTRFACE





Output


Heart Disease Prediction Form
Age
Sex
-- Select an Option --
Chest Pain Type
Resting Blood Pressure in mm
Serum Cholestoral in mg/dl
Fasting Blood Sugar > 120 mg/dl
-- Select an Option --
-- Select an Option --
Resting ECG Results
Maximum Heart Rate
ST Depression Induced
Exercise Induced Angina
-- Select an Option --
-- Select an Option --
Slope of the Peak Exercise ST Segment
Number of Vessels Colored by Flourosopy
Thalassemia
-- Select an Option --
-- Select an Option --
-- Select an Option --
Result


Heart Disease Prediction Form
Age
63
Sex
Male
Chest Pain Type
Asymptomatic
Resting Blood Pressure in mm
145
Serum Cholestoral in mg/dl
233
Fasting Blood Sugar > 120 mg/dl
True
Resting ECG Results
Normal
Maximum Heart Rate
150
ST Depression Induced
2.3
Exercise Induced Angina
No
Slope of the Peak Exercise ST Segment
Upsloping
Number of Vessels Colored by Flourosopy
0
Thalassemia
Normal
Result
The patient is likely to have heart disease!

scope

- Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome.
- This suggestion is promising as data modelling and analysis tools, e.g. data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Conclusion

The overall objective of the system is to predicts the presence of heart disease accurately with less number of tests and attributes. In this project, fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span. Three data mining classification techniques were applied namely K-Nearest Neighbor, Random Forest and Logistic Regression. It is shown that Logistic Regression has better accuracy of 89% which is better than the other techniques and then we successfully deployed our best model into web api.

References

1. L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automat ed Diagnostic System for Heart Disease Prediction Based on χ2 Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938–34945, 2019,doi: 10.1109/ACCESS.2019.2904800.
2. Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Predict ion System," Procedia Comput. Sci., vol. 85, pp. 962–969,2016, doi: 10.1016/j.procs.2016.05.288.
3. L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Predict ion of Heart Failure," IEEE Access, vol. 7, pp. 54007–54014, 2019, doi:10.1109/ACCESS.2019.2909969.
4. K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," Proc. 2016 2nd Int. Conf. Appl. Theor.Comput. Commun. Technol. iCATccT 2016, no. January 2016, pp.381–386, 2017, doi: 10.1109/ICATCCT.2016.7912028.
5. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour ,"An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235–180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

Thank you