Heart-Attack prediction using Machine Learning

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

COMPUTER SCIENCE & ENGINEERING

Submitted By

SOMERON BAKULI	University Roll No. 1R9000390
AMAR SARKAR	University Roll No. 1R9000454
SAPTORSHE DAS	University Roll No. 1R9000058
KOYEL DAS	University Roll No. 1R9000520
SUBHRAJYOTI SAHA	University Roll No. 1R9000418
SAYANTAN DAS	University Roll No. 1R9000412

Under the guidance of

PROF. ARUNABHA TARAFDAR

Department of Computer Science & Engineering

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled **Heart-Attack prediction using Machine Learning**

submitted by **Someron Bakuli(University Roll No.1R9000390), Amar Sarkar(University Roll No. 1R9000454)**, **Saptorshe Das(University Roll No. 1R9000058), Koyel Das(University Roll No.1R9000520)** and **Subhrajyoti Saha(University Roll No. 1R9000418)** and **Sayantan Das(University Roll No. 1R9000412)**, students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfilment of requirement for the degree of Bachelor of Computer Science, is a bonafide work carried out by them under the supervision and guidance of Prof. Arunabha Tarafdar during 7^{th} and 8^{th} Semester of academic session of 2020 - 2021. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

Prof. Arunabha Tarafdar	Prof. Sukalyan Goswami
Assistant Professor	Head of the Department
Department of CSE Department of CSE	Department of CSE Department of CSE
UEM, Kolkata	UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Arunabha Tarafdar of the Department of Computer Science& Engineering, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. Sukalyan Goswami, HOD, Computer Science& Engineering, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

SOMERON BAKULI

AMAR SARKAR

SAPTORSHE DAS

KOYEL DAS

SUBHRAJYOTI SAHA

SAYANTAN DAS

TABLE OF CONTENTS

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Machine learning is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Decision tree, Logistic Regression, and Support Vector Machine algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with Support Vector Machine.

CHAPTER – 1

INTRODUCTION

Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke. The vast number of deaths is common amongst low and middle-income countries. Many predisposing factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various habitual risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are predisposing factors for heart disease. The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death.

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis. In this research paper, I'll discuss a project where I worked on predicting potential Heart Diseases in people using Machine Learning algorithms. The algorithms included Decision tree, Logistic Regression, and Support Vector Machine algorithm.

CHAPTER – 2

LITERATURE SURVEY

Mohammed Abdul Khaleel has given paper in the Survey of Techniques for mining of data on Medical Data for Finding Frequent Diseases locally. This paper focus on dissect information mining procedures which are required for medicinal information mining particularly to find locally visit illnesses, for example, heart infirmities, lung malignancy, bosom disease et cetera. Information mining is the way toward extricating information for finding inactive examples which Vembandasamy et al. performed a work, to analyze and detect heart disease. In this the algorithm used was Naive Bayes algorithm. In Naïve Bayes algorithm they used Bayes theorem. Hence Naive Bayes has a very power to make assumption independently. The used data-set is obtained from a diabetic research institutes of Chennai, Tamilnadu which is leading institute. There are more than 500 patients in the dataset. The tool used is Weka and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%.

Costas Sideris, Nabil Alshurafa, Haik Kalantarian and Mohammad Pourhomayoun have given a paper named Remote Health Monitoring Outcome Success prediction using First Month and Baseline Intervention Data. RHS systems are effective in saving costs and reducing illness. In this paper, they portray an up-graded RHM framework, Wanda- CVD that is cell phone based and intended to give remote instructing and social help to members. CVD counteractive action measures are perceived as a basic focus by social insurance associations around the world.

L.Sathish Kumar and A. Padmapriya has given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone. This paper gives a programmed and concealed way to deal with recognize designs that are covered up of coronary illness. The given framework utilize information mining methods, for example, ID3 algorithm. This proposed method helps the people not only to know about the diseases but it can also help’s to reduce the death rate and count of disease affected people.

M.A.Nishara Banu and B.Gomathy has given a paper named Disease Predicting system using data mining techniques. In this paper they talk about MAFIA (Maximal Frequent Item set algorithm) and K-Means clustering. As classification is important for prediction of a disease. The classification based on MAFIA and K-Means results in accuracy.

Wiharto and Hari Kusnanto have given a paper named Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.

D.R. PatiI and Jayshril S. Sonawane have given a paper named Prediction of Heart Disease Using Learning Vector Quantization Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.

CHAPTER – 3

PROBLEM STATEMENT

Heart disease has created a lot of serious concerned among researches; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in finding it even medical professor are not so much efficient enough in predicating the heart disease. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human. According to latest survey conducted by WHO, the medical professional able to correctly predicted only 67% of heart disease so there is a vast scope of research in area of predicating heart disease in human. With advancement in computer science has brought vast opportunities in different areas, medical science is one of the fields where the instrument of computer science can be used. In application areas of computer science varies from metrology to ocean engineering. Medical science also used some of the major available tools in computer science; in last decade artificial intelligence has gained its moment because of advancement in computation power. Machine Learning is one such tool which is widely utilized in different domains because it doesn't require different algorithm for different dataset. Reprogrammable capacities of machine learning bring a lot of strength and opens new doors of opportunities for area like medical science. In medical science heart disease is one of the major challenges; because a lot of parameters and technicality is involve for accurately predicating this disease. Machine learning could be a better choice for achieving high accuracy for predicating not only heart disease but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the heart disease, algorithms such as Decision tree, Logistic Regression, and Support Vector Machine

CHAPTER – 4

PROPOSED SOLUTION

4.1 Background

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Machine learning is a software technology that helps computers to build and classify various attributes. This research paper uses classification techniques to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions, data pre-processing, evaluation measurements and description of the dataset used in this research.

Machine Learning

Machine learning is an emerging subdivision of artificial intelligence. Its primary focus is to design systems, allow them to learn and make predictions based on the experience. It trains machine learning algorithms using a training dataset to create a model. The model uses the new input data to predict heart disease. Using machine learning, it detects hidden patterns in the input dataset to build models. It makes accurate predictions for new datasets. The dataset is cleaned and missing values are filled. The model uses the new input data to predict heart disease and then tested for accuracy. Machine learning techniques are classified as:

Supervised Learning

The model is trained on a dataset that is labelled. It has input data and its outcomes. Data are classified and split into training and test dataset. Training dataset trains our model while testing dataset functions as new data to get accuracy of the model. The dataset exists with models and its output. The classification and regression are its example.

Unsupervised Learning

Data used to train are not classified or labelled in the dataset. Aim is to find hidden patterns in the data. The model is trained to develop patterns. It can easily predict hidden patterns for any new input dataset, but upon exploring

data, it draws conclusion from datasets to describe hidden patterns. In this technique, no responses in the dataset are seen. The clustering method is an example of an unsupervised learning technique.

Reinforcement Learning

It does not use labelled dataset nor the results are associated with data, thus model learns from the experience. In this technique, the model improves its presentation based on its association with environment and figures out how to discuss its faults and to get the right outcome through assessment and testing various prospects.

Classification algorithms are commonly used supervised learning techniques to define probability of heart disease occurrence.

4.2Approach Methodology

This research aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients . To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in this research paper. This paper additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome .

Data Source

For this study, I have used dataset from Kaggle Machine learning repository. It comprises a real dataset of more than 1000 examples of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. In this research, we have used three algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.

Sr. no.	Attribute	Representative icon	Details
1	Age	Age	Patients age, in years
2	Sex	Sex	0 = female; 1 = male
3	Chest pain	Cp	4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic)
4	Rest blood pressure	Trestbps	Resting systolic blood pressure (in mm Hg on admission to the hospital)
5	Serum cholesterol	Chol	Serum cholesterol in mg/dl
6	Fasting blood sugar	Fbs	Fasting blood sugar > 120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	Restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	MaxHeart rate	Thalch	Maximum heart rate achieved
9	Exercise-induced angina	Exang	Exercise-induced angina (0—no; 1—yes)
10	ST depression	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope	slope of the peak exercise ST segment (1—upsloping; 2—flat; 3—down sloping)
12	No. of vessels	Ca	No. of major vessels (0–3) colored by fluoroscopy
13	Thalassemia	Thal	Defect types; 3—normal; 6—fixed defect; 7—reversible defect
14	Num(class attribute)	Class	diagnosis of heart disease status (0—nil risk; 1—high risk)

Data Pre-Processing

Cleaning: Data that we want to process will not be clean that is it may contain noise or it may contain values missing of we process we cant get good results so to obtain good and perfect results we need to eliminate all this, the process to eliminate all this is data cleaning. We will fill missing values and can remove noise by using some techniques like filling with most common value in missing place.

Transformation: This involves changing data format to one form to other that is making them most understandable by doing normalization, smoothing, and generalization, aggregation techniques on data.

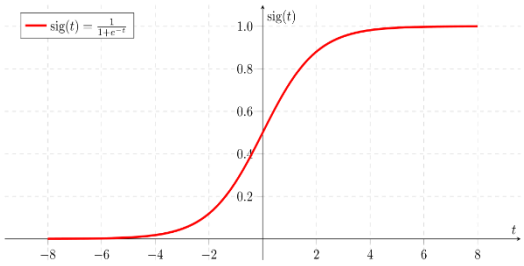
Integration: Data that we need not process may not be from a single source sometimes it can be from different sources we do not integrate them it may be a problem while processing so integration is one of important phase in data preprocessing and different issues are considered here to integrate.

Reduction: When we work on data it may be complex and it may be difficult to understand sometimes so to make them understandable to system we will reduce them to required format so that we can achieve good results.

Algorithm Used

Logistic Regression :

It's a type of statistical regression analysis method used for approximation and prediction of result of a definite dependent attributes. Dependent means it can take only some set of values for example binary values such as true or false, good or bad, on or off likewise. Logistic regression is mainly used for prediction besides that it can also be used in calculating the probability of success. Basically Logistic Regression involves fitting an equation of the form to the data:



$$Y = ß_0 + ß_{1x1} + ß_{2x2} + \ldots + ß_{nxn} \quad \text{----- eq. 1}$$

The regression coefficients are usually estimated using maximum likelihood estimation. The maximum likelihood ratio helps to determine the statistical significance of independent variables on the dependent variables. The likelihood-ratio test assesses the contribution of individual predictors (independent variables). Then the probability (p) of each case is calculated using odds ratio,

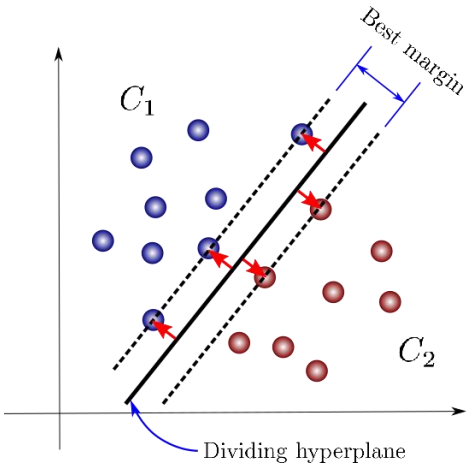
$$P/(1-P) = e^{Y} \quad \text{---- eq. 2}$$

From this p–value is found out. This gives the probability or chance for the individual to have coronary heart disease.

SVM:

SVM is a supervised machine learning model that follows classification algorithms. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value(0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin. In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.



if $y*f(x) >= 1$, $c(x, y, f(x)) = 0$

else , $c(x, y, f(x)) = 1-y*f(x)$

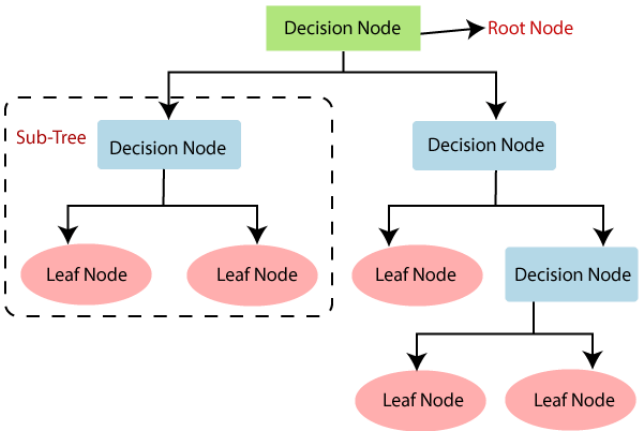
The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

We used SVM as we are working here on multiple values as well our dataset is used in a optimized way.

DECISION TREE :

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Below diagram illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), No Rain(No)).



Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this.

Often we find overfitting in decision tree. If model is overfitted it will poorly generalized to new samples. To avoid decision tree from overfitting we remove the branches that make use of features having low importance. This method is called as Pruning or post-pruning.

CAPTER – 5

Experimental Setup And Result Analysis

5.1Experimental Setup

ML libraries

Numpy

Pandas

Matplotlib

Sklearn

Seabon

Pandas Profiling

Joblib and Pickle

Front-end & Back-end

HTML

CSS

Flask

Python

System Configuration

2 GB RAM

100MB free memory

Browser

Stable internet connection

Co-lab credentials

5.2 Model Evaluation

Logistic Regression

Confusion Matrix

[[77 21]
[7 100]]

Accuracy of Logistic Regression: 86.34146341463415

	precision	recall	f1-score	support
0	0.92	0.79	0.85	98
1	0.83	0.93	0.88	107
accuracy			0.86	205
macro avg	0.87	0.86	0.86	205
weighted avg	0.87	0.86	0.86	205

Decision Tree

Confusion Matrix:

[[95 3]
[8 99]]

Accuracy of Decision Tree Classifier: 94.6341463414634

	precision	recall	f1-score	support
0	0.92	0.97	0.95	98
1	0.97	0.93	0.95	107
accuracy			0.95	205
macro avg	0.95	0.95	0.95	205
weighted avg	0.95	0.95	0.95	205

Support Vector Machine

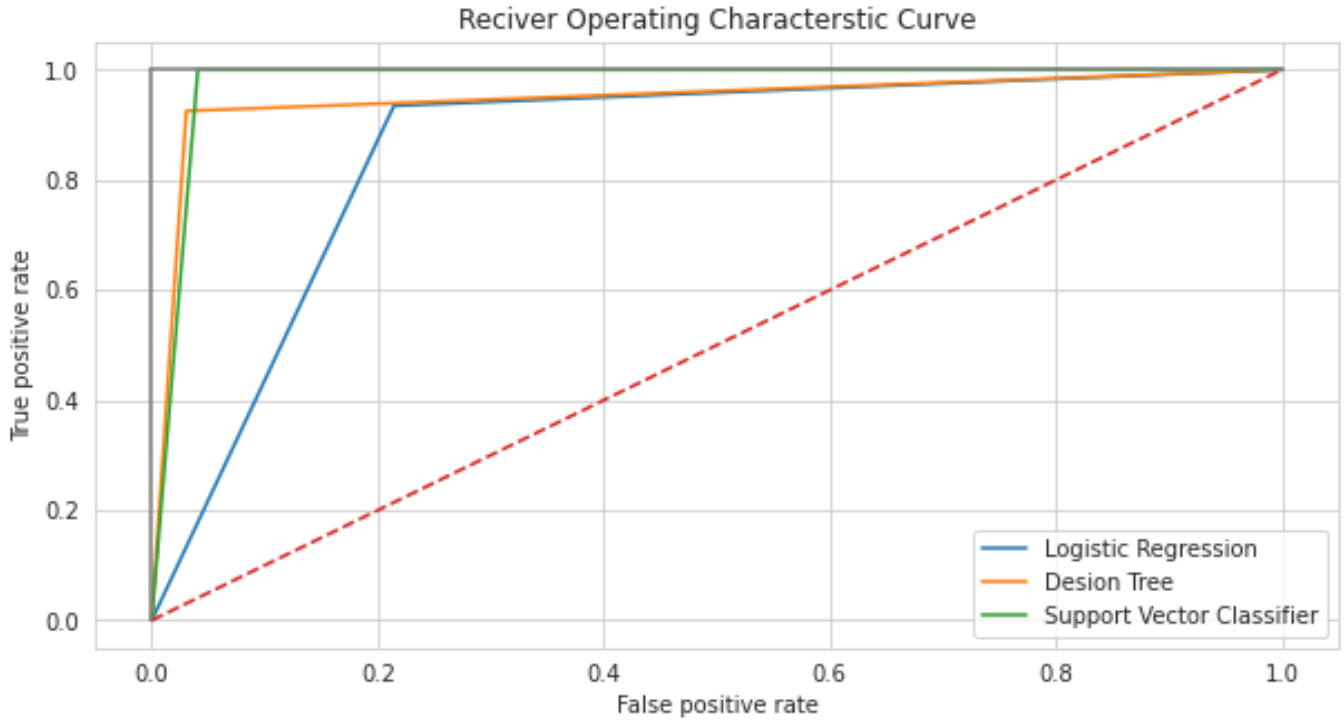
Confusion Matrix:

[[94 4]
[0 107]]

Accuracy of Support Vector Classifier: 98.04878048780488

	precision	recall	f1-score	support
0	1.00	0.96	0.98	98
1	0.96	1.00	0.98	107
accuracy			0.98	205
macro avg	0.98	0.98	0.98	205
weighted avg	0.98	0.98	0.98	205

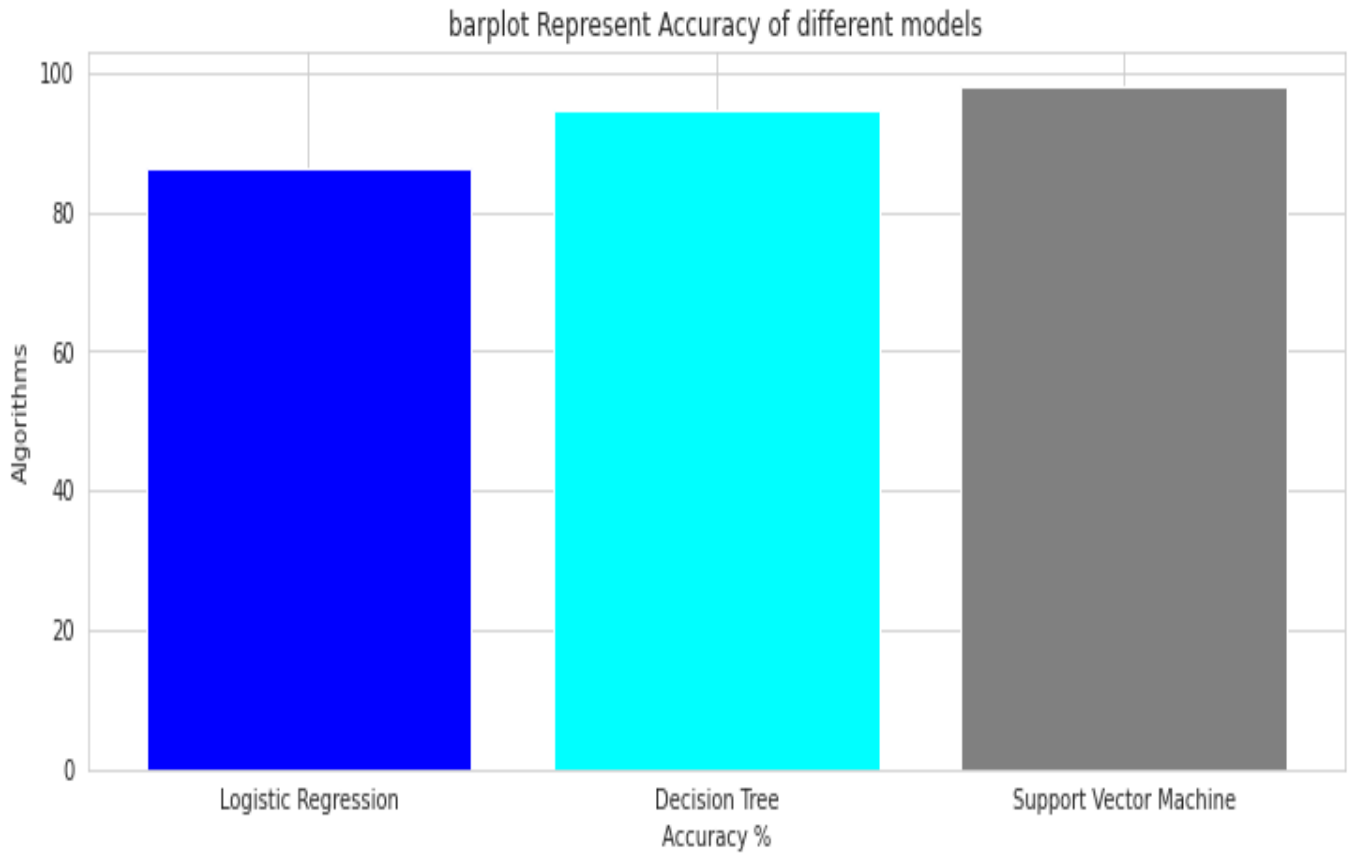
Reciver Operating Characterstic Curve



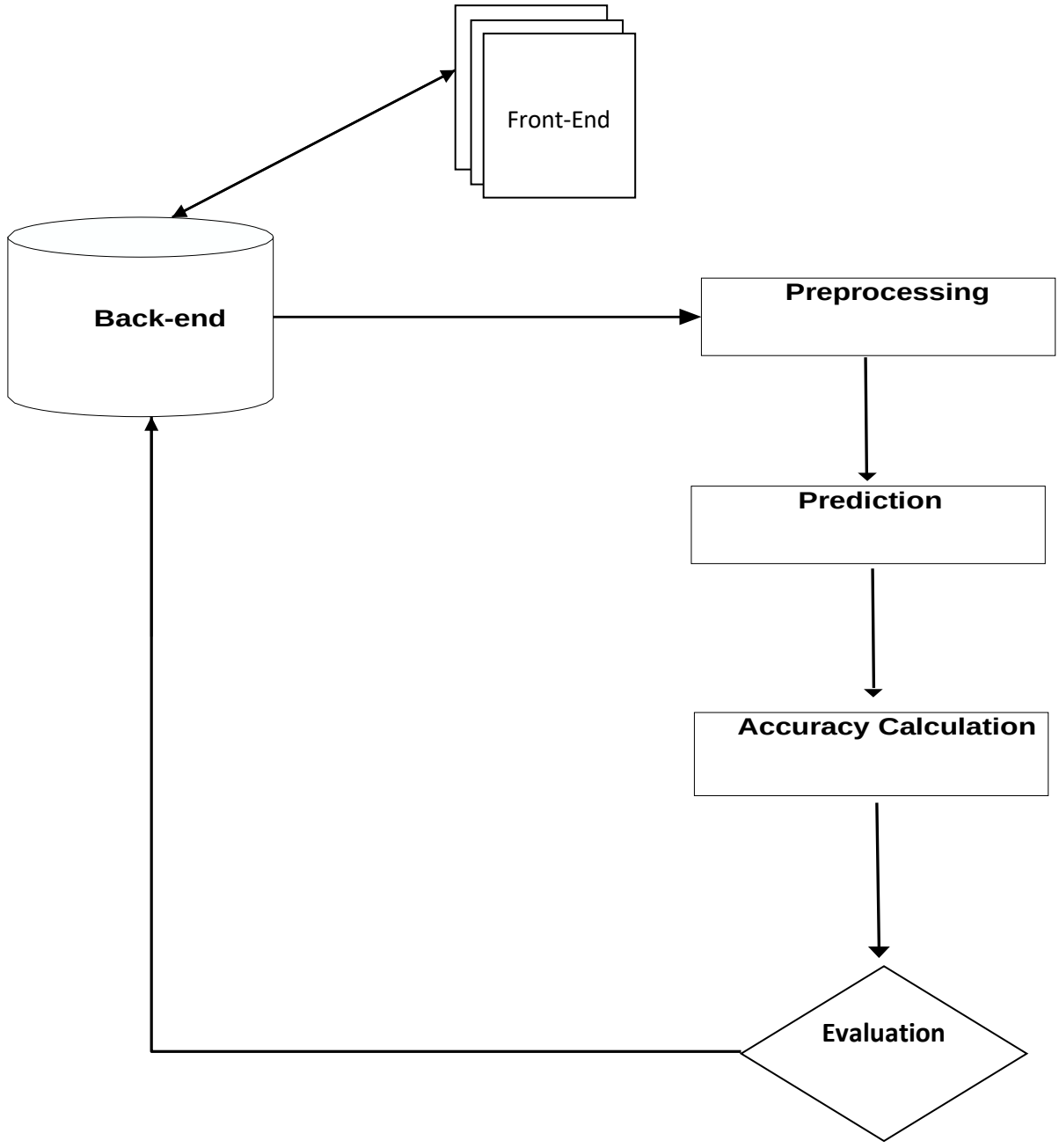
Accuracy Comparison

	Model	Accuracy
0	Logistic Regression	86.341463
1	Decision Tree	94.634146
2	Support Vector Machine	98.048780

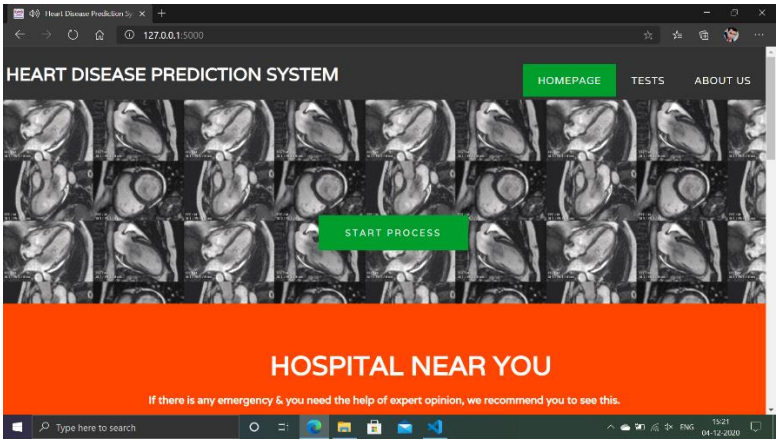
Barplot Represent Accuracy of different models

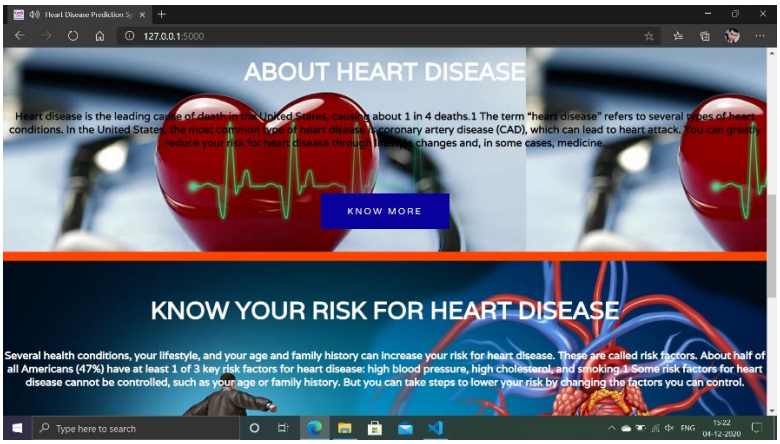


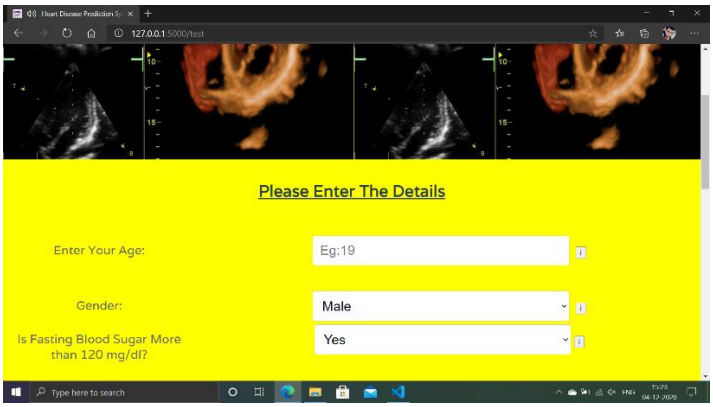
Data Flow Diagram:

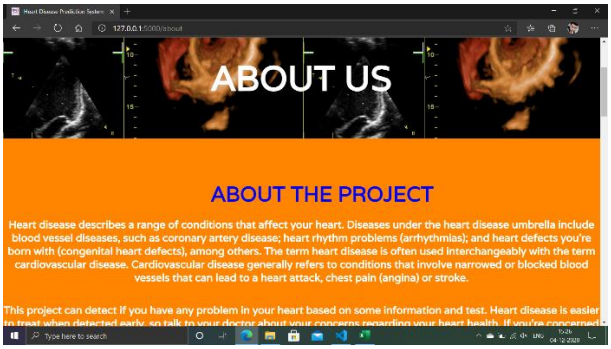


Working Demo:









CHAPTER – 6 :

CONCLUSION & FUTURE SCOPE

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 modeling 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

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

Conclusion and Future Work

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

BIBLIOGRAPHY

Dataset Source : https://www.kaggle.com/johnsmith88/heart-disease-dataset

Reference :

- **https://link.springer.com/article/10.1007/s42979-020-00365-y**
- **https://www.researchgate.net/publication/326733163_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms**
- **https://en.wikipedia.org/wiki/Decision_tree**
- **https://en.wikipedia.org/wiki/Logistic_regression**
- **https://en.wikipedia.org/wiki/Support_vector_machine**
- **https://lucdemortier.github.io/projects/3_mcnulty**