
Heart Disease Prediction using ML

A Project Report Submitted in Partial Fulfillment of the requirements

For the award of the degree of

Bachelor of Degree

In

Computer Science And Engineering

by

Jyoti Gupta (171500149)

Varsha Varshney (171500375)

Under the Guidance of

Dr. Manish Raj

(Asst. Professor)

Department of Computer Engineering & Applications

Institute of Engineering & Technology



GLA University

Mathura- 281406, INDIA

2020



Department of computer Engineering and Applications

GLA University, Mathura

17 km. Stone NH#2, Mathura-Delhi Road, P.O. –Chaumuha,

Mathura – 281406

Declaration

We hereby declare that the work which is being presented in the B.Tech Project “Heart Disease Prediction” is partial fulfilment of the requirement for the award of the Bachelor of Technology in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of GLA University, Mathura is the authentic record of our team work carried under the supervision of **Dr. Manish Raj (Asst. Professor)**.

Signature : _____

Name of Candidate: Jyoti Gupta

University Roll No. : 171500149

Signature : _____

Name of Candidate: Varsha Varshney

University Roll No. : 171500375

ACKNOWLEDGEMENT

It is my pleasure to acknowledge the assistance of a number of people without whose help this project would not have been possible.

This Project in itself is an acknowledgement to the inspiration, drive and technical Assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and Guidance that we have received.

First and foremost, I would like to express our gratitude to **Dr. Manish Raj (Asst. Professor)** my project guide, for providing invaluable encouragement, guidance and assistance. I would like to thank the lab staff for the operation extended to us throughout the project. After doing this project I can confidently say that this experience has not only enriched me with technical knowledge but also has unparsed the maturity of thought and vision. The attributes required being a successful professional.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project.

Last but not least, we acknowledge our friends for their contribution in the completion of this project.

Signature : _____

Name of Candidate: Jyoti Gupta

University Roll No. : 171500149

Signature : _____

Name of Candidate: Varsha Varshney

University Roll No. : 171500375

ABSTRACT

Heart related diseases or Cardiovascular Diseases are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (HDPS) is developed using SVM, Decision Tree, Logistic Regression and Naives_Bayes/Data Mining Techniques for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The HDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g. Relationships between medical factors related to heart disease and patterns, to be established. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

Table of Content

Declaration.....	I
Acknowledgement.....	II
Abstract.....	III
1. Introduction	
1.1 Objective.....	1
1.2 Existing System.....	2
1.3 Proposed System.....	2
2. Software Requirement Specification	
2.1 Technical Feasibility.....	3
2.2 Scope.....	3
3. Data Mining Algorithm.	
3.1 Decision Tree Algorithm.....	4
3.2 Support Vector Machine.....	4
3.3 Naive Bayes.....	4
3.4 Logistic Regression.....	5
4. Software Design	
4.1 Use Case.....	6
4.2 Data Flow Diagram.....	7
5. Testing.....	9
6. Implementation and Interface.....	11 - 18
7. Conclusion.....	19
8. References.....	20

CHAPTER 1

Introduction

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease .

1.1 Objective

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. The overall objective of this project will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered 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 few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

1.2 Existing System

The before all existing system works on sets of both Deep learning and data mining. The existing system modules generates comprehensive report by implementing the strong prediction algorithm. The main aims of the existing system to compare and check the before patient whose having disease outputs and new patient disease and determine future possibilities of the heart disease to a particular patient. By Implementing the above mentioned model we will get the goal of developing a system with increased rate of accuracy of estimating the new patient getting heart attack percentage. The model which is proposed for Heart Attack Prediction System is invented for using Deep learning algorithms and approach. But by using all the existing systems the accuracy is very less.

Proposed System

This proposed system have a data which classified if patients have heart disease or not according to features in it. This proposed system can try to use this data to create a model which tries predict(reading data and data Exploration) if a patient has this disease or not. In this proposed system, use SVM, Decision Tree, Logistic Regression and algorithm. By using sklearn library to calculate score. Implements Naïve Bayes algorithm to getting accuracy result. From the data we are having, it should be classified into different structured data based on the features of the patient heart .By using logistic Regression, naïve_bayes the accuracy rate increases.

Chapter 2

Software Requirement Specifications

1.3 Technical Feasibility

Heart Disease Prediction System is generally based on the analysis of cardiovascular Diseases; so the first and foremost need is of medical information from the users, we can get the information from the medical information from given in the user interface module, it works as an input to the medical information form, Next step is to classify the disease information using Data Mining Algorithms, after taking a certain amount of information from the medical information form based on the Data Mining, Diseases Prediction module can identify different scenarios, after this step if an diseases has been recognized, further it can predict the different algorithm accuracy about the Heart Disease.

1.4 Scope

The Heart Disease Prediction has always been beneficial to the world, whether they are Plant Disease Prediction or Climate Change Disease Prediction. Our Heart Disease Prediction System predict the disease for the occurring information thus preventing loss of life and money. 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 .

Chapter 3

Data Mining Algorithms

1. Decision Tree

A Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences including chance event outcomes and utility. It is one of the ways to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis to help and identify a strategy that will most likely reach the goal. It is also a popular tool in machine learning. A Decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one. Finally by following these rules, appropriate conclusions can be reached.

2. Support Vector Machine(SVM)

It is a supervised learning method which classifies data into two classes over a hyper plane. Support vector machine performs a similar task like C4.5 except that it doesn't use Decision trees at all. Support vector machine attempts to maximize the margin (distance between the hyper plane and the two closest data points from each respective class) to decrease any chance of misclassification. Some popular implementations of support vector machine are scikit-learn, MATLAB and of LIBSVM.

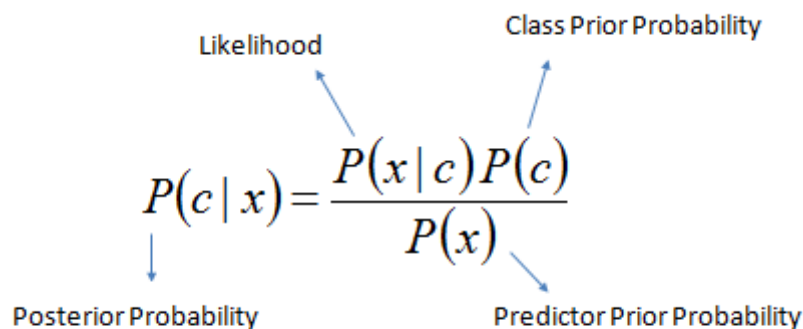
3. Logistic Regression

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

4. Naive Bayes(NB):

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$


$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

- $P(c/x)$ is the posterior probability of *class (target)* given *predictor (attribute)*.
- $P(c)$ is the prior probability of *class*.
- $P(x/c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

Chapter 4

Software Design

The project will be structured into three phases: An investigation into the available frameworks; The creation of a software prototype to meet the minimum requirements specified above; A discussion of possible enhancements to the prototype, and implementation of as many features as possible.

The first phase of the project will focus on finding the most appropriate framework for implementing the model company's prototype solution. A selection of frameworks will be chosen from the web, and then disposable prototype systems will be created in the two frameworks deemed most suitable for the second phase of the project. Based on the experiences with these two disposable prototypes, a decision will be made for which framework will be utilized for creating the prototype solution for the next phase.

In the second phase, Data Analysis processes will be modeled, and a requirement specification created for a web-based system. The design for the Analysis will be produced, and implemented in the framework selected in phase 1. The system will then be tested by the Backhand Big Data Analysis.

Finally in the third phase, some system enhancements will be detailed and appropriate solutions to problems arising from user testing will also be described. Design issues in implementing the new features will be considered, and the prototype from phase 2 will be enhanced as much as time permits.

4.1 UML Diagram

A diagram is the graphical presentation of a set of elements, most often rendered as a connect graph of vertices and arcs. You draw to visualize a system from different perspective, so a diagram is a projection into a system. For all but most trivial systems, a diagram represents an elided view of the element that make up a system. The same element may appear in all diagrams, only a few diagrams or in no diagrams at all. In theory, a diagram may contain any combination of things and relationships. In practice,

however, a small number of common combination arise, which are consistent with the five most useful views that comprise the architecture of a software-intensive system.

1. Use case Diagram
2. Data flow Diagram
3. Sequence Diagram

4.2 Use case Diagram

A use case diagram is the Unified Modelling Language (UML) is a type of behavioral diagram defined by and created from a use-case analysis. Its purpose is to present a graphical overview of the functionality provide by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

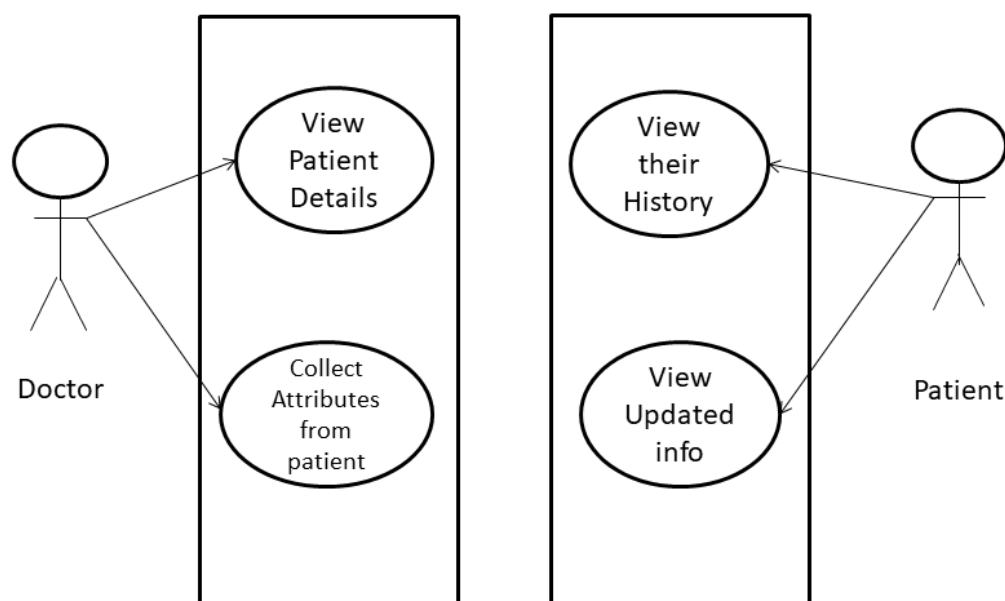


Fig.4.1 Use Case Diagram

4.2 Data Flow Diagram

The **database schema** of a database system is its structure described in a formal language supported by the database management system (DBMS). The term "schema" refers to the organization of data as a blueprint of how the database is constructed (divided into database tables in the case of relational databases). The formal definition of a database schema is a set of formulas (sentences) called integrity constraints imposed on a database. These integrity constraints ensure compatibility between parts of the schema.

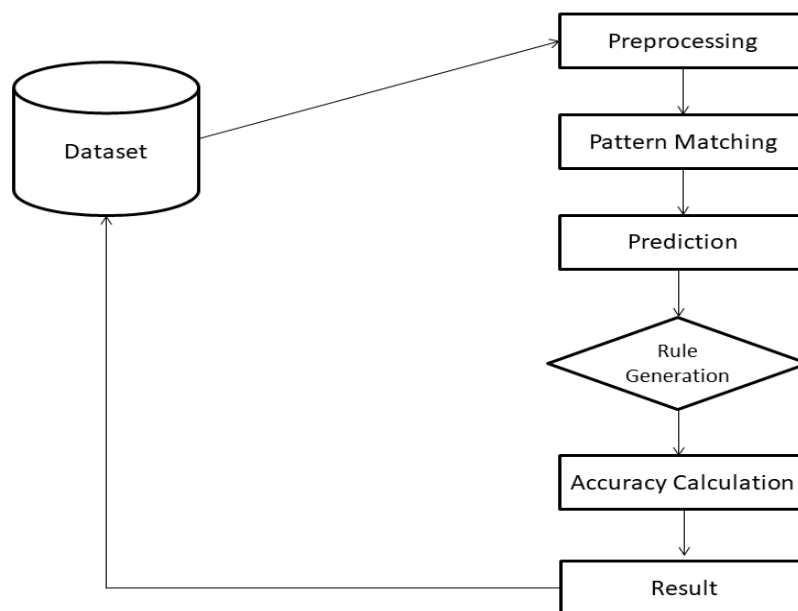


Fig. 4.2 Data Flow Diagram

Chapter 5

Testing

5.1 Introduction

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. In fact, testing is the one step in the software engineering process that could be viewed as destructive rather than constructive.

A strategy for software testing integrates software test case design methods into a well-planned series of steps that result in the successful construction of software. Testing is the set of activities that can be planned in advance and conducted systematically. The underlying motivation of program testing is to affirm software quality with methods that can be economically and effectively apply to both strategies to both large and small scale system.

The following are the Testing Objectives:

- Testing is a process of executing a program with the intent of finding an error.
- A good test has a high probability of finding an as yet undiscovered error.
- A successful test is one that uncovers an as yet undiscovered error.

5.2 Design of Test cases & scenarios

The objective is to design tests that systematically uncover different classes of errors and do so with a minimum amount of time and effort. Testing cannot show the absence of defects. It can only show that software defects are present.

5.2.1 Integration Testing

Modules integrated by moving down the program design hierarchy. Can use depth first or breadth first top down integration verifies major control and decision points early in design process. Top- level structure tested most. Depth first implementation

allows a complete function to be implemented, tested and demonstrated and does depth first implementation of critical function early. Top down integration forced (to some extent) by some development tools in program with graphical user interfaces.

Begin construction and testing with atomic modules (lowest level modules). Bottom up integration testing as its name implies begins construction and testing with atomic modules. Because modules are integrated from the bottom up, processing required for modules subordinates to a given level is always available and need for stubs is eliminated.

5.2.2 Validation Testing

Validation testing aims to demonstrate that the software functions in a manner that can be reasonably expected by the customer. This tests conformance the software to the Software Requirements Specification.

5.2.2.1 Validation Test Criteria

A set of black box test is to demonstrate conformance with requirements. To check that all functional requirements satisfied, all performance requirements achieved, documentation is correct and 'human engineered', and other requirements are met

e.g. Compatibility ,Error recovery and Maintainability

When validation tests fail it may be too late to correct the prior to scheduled delivery. Need to negotiate method of resolving deficiencies with the customer.

Chapter 6

Implementation And Interface

6.1 Requirement for the Project

The requirement of Prediction website is to allow getting information to solve problems in the corporation; going out and seeking opinions on optimal, actual, feelings, causes, and solutions and the user to put their own disease dataset. It provides developer to check where the accuracy of disease in the dataset on the bar-graph and also they can check their recommended data .

6.1.1 Server Side

We are using Web Server(Django Server) to provide comprehensive user interface to reviewer the will be solely responsible for deciding if a reported event is false positive. Server Side also contains the base module that is responsible for reporting of disease to the reviewer. I will be applying Machine Learning approaches(and eventually comparing them) for classifying whether a person is suffering from heart disease or not, using one of the most used dataset — Cleveland Heart Disease dataset from the UCI Repository.

6.1.2 Client Side

User would need a computer with an updated web browser to interact with our UI and at the client side we also need a reviewer to verify the authenticity of a reported event.

6.2 System Module

In this the module that act as an interface between Client and Server is Describe in detail.

6.2..1 User Interfaces

We are providing the user (reviewer) with all different kind of functionalities such as notifying him/her in case he/she is not afford enough or needs a result related to his disease. In case a critical disease is predict in absence of the reviewer our UI is capable enough to automatically show the result through the Medical information that they would filling and

report directly to him/her profile section.

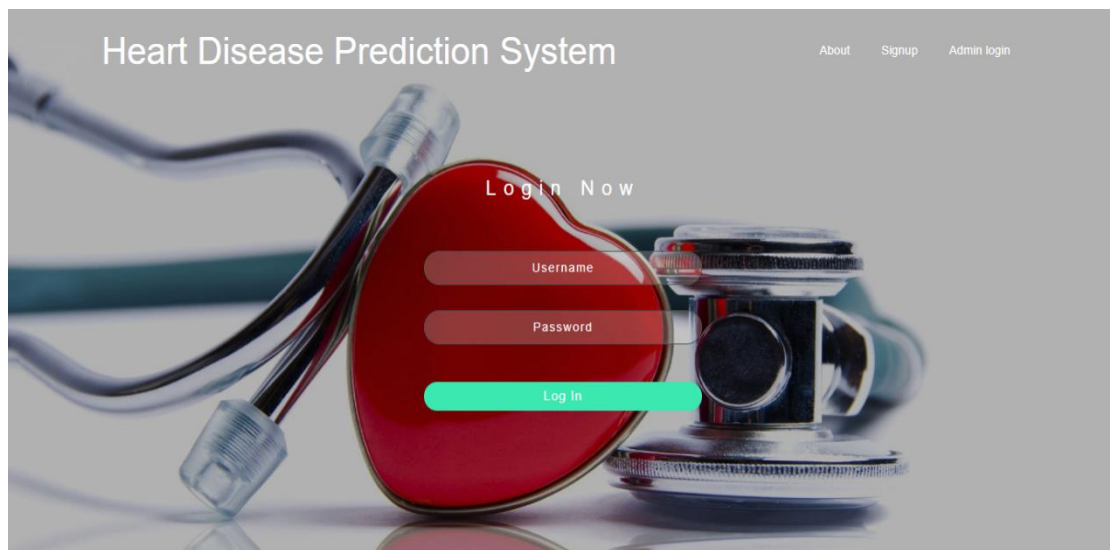


Fig.6.1 Login

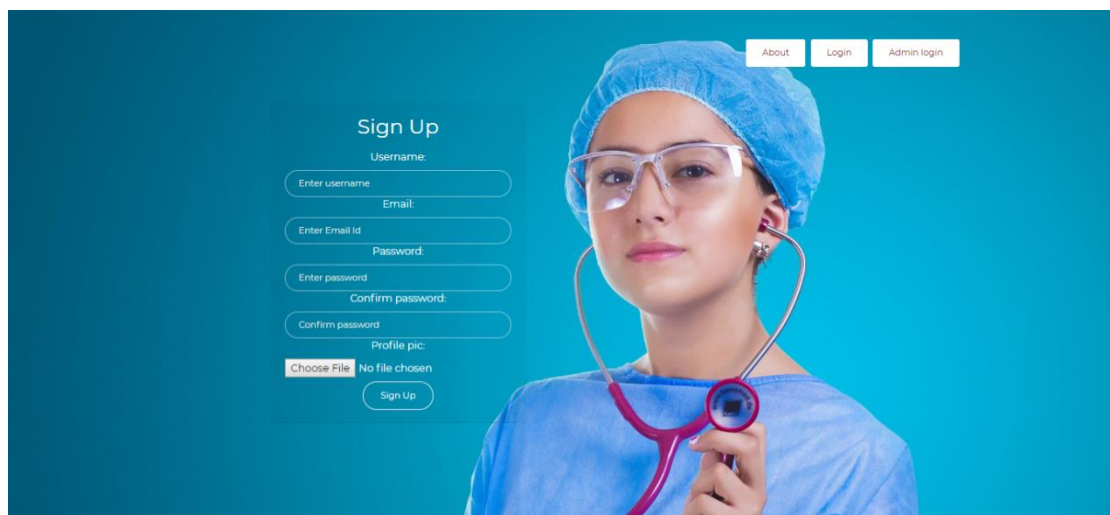


Fig.6.2 Sign Up

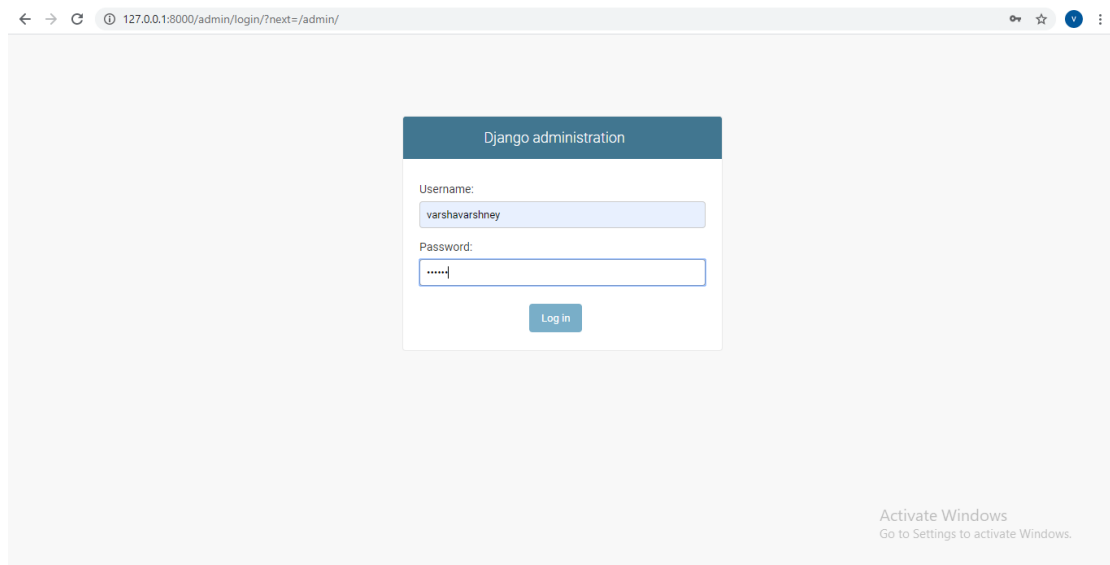


Fig.6.3 Admin Login Page

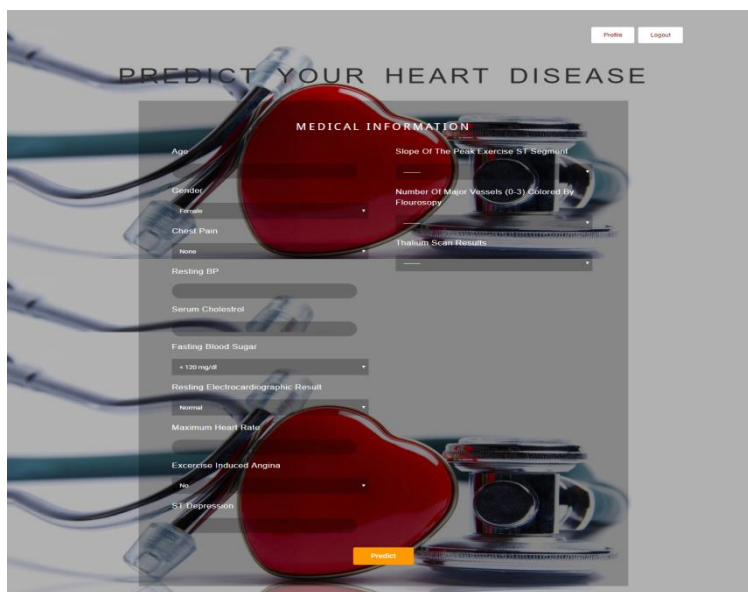


Fig.6.4 Predict Page

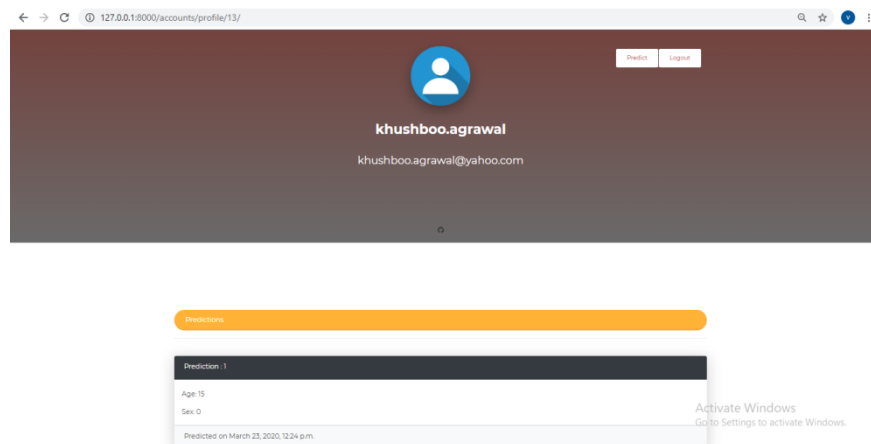


Fig.6.5 User Profile

HealthData - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num							
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0							
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1							
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1							
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0							
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0							
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0							
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	1							
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0							
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	1							
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1							
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0							
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0							
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	1							
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0							
16	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0							
17	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0							
18	48	1	2	110	229	0	0	168	0	1	3	0	7	1							
19	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0							
20	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0							
21	49	1	2	130	266	0	0	171	0	0.6	1	0	3	0							
22	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0							
23	58	0	1	150	283	1	2	162	0	1	1	0	3	0							
24	58	1	2	120	284	0	2	160	0	1.8	2	0	3	1							
25	58	1	3	132	224	0	2	173	0	3.2	1	2	7	1							

Fig.6.6 Cleveland_Dataset

Age: 15

Gender: Female

Chest Pain: Typical Angina

Resting BP: 162

Serum Cholesterol: 250

Fasting Blood Sugar: < 120 mg/dl

Resting Electrocardiographic Result: Having ST-T wave abnormality

Slope Of The Peak Exercise ST Segment: Flat

Number Of Major Vessels (0-3) Colored By Fluoroscopy: Two

Thalium Scan Results: Normal

Fig.6.7 Prediction page filled

Django administration

WELCOME, VARSHAVARSHNEY VIEW SITE / CHANGE PASSWORD / LOG OUT

Home > Predict_Risk > Predictionss

Select predictions to change

ADD PREDICTIONS +

Action: Go 0 of 7 selected

<input type="checkbox"/>	PROFILE	AGE	SEX	CP	RESTING BP	SERUM CHOLESTEROL	FASTING BLOOD SUGAR	RESTING ECG	MAX HEART RATE	EXERCISE INDUCED ANGINA	ST DEPRESSION	ST SLOPE	NUMBER
<input type="checkbox"/>	khushboo.agrawal	15	Female	Typical Angina	162	250	< 120 mg/dl	Having ST-T wave abnormality	85	No	12.00	Flat	Two
<input type="checkbox"/>	nishi45	21	Female	Atypical Angina	110	50	< 120 mg/dl	Normal	80	Yes	45.00	Down Sloping	None
<input type="checkbox"/>	nishi45	21	Female	Non-Angina	120	100	< 120 mg/dl	Having ST-T wave abnormality	80	Yes	90.00	Upsloping	Two
<input type="checkbox"/>	yash18	18	Male	None	45	59	< 120 mg/dl	Normal	58	Yes	60.00	Flat	None
<input type="checkbox"/>	yash18	18	Male	Atypical Angina	120	65	< 120 mg/dl	Normal	98	Yes	80.00	Upsloping	None
<input type="checkbox"/>	yash18	18	Male	Atypical Angina	120	65	> 120 mg/dl	Normal	98	Yes	80.00	Upsloping	Two
<input type="checkbox"/>	yogendra	65	Male	None	110	180	< 120 mg/dl	Normal	80	No	90.00	Upsloping	None

7 predictionss

Activate Windows
Go to Settings to activate Windows.

Fig.6.8 User Database

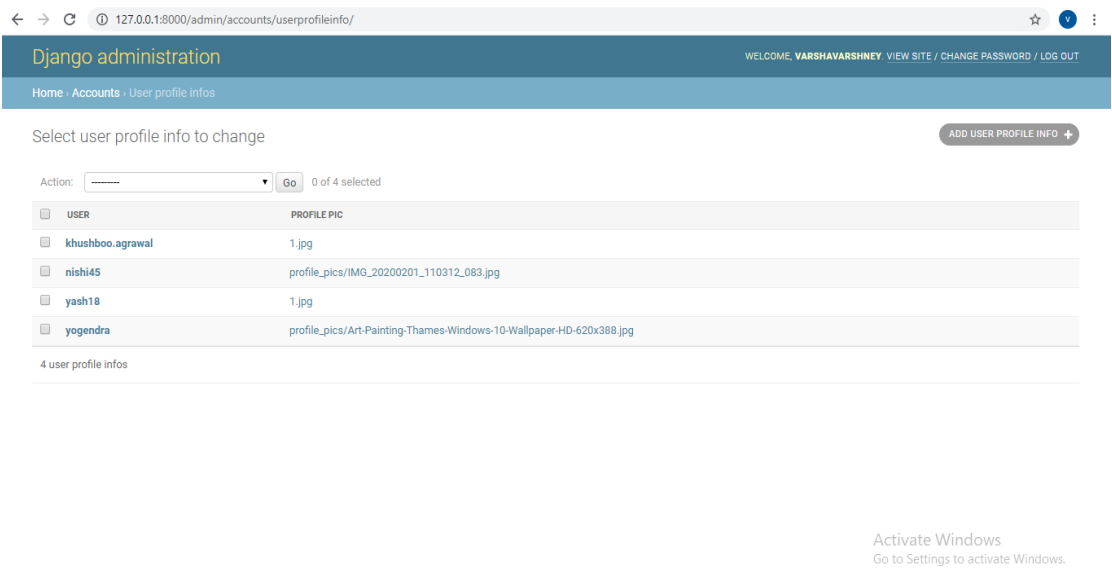


Fig.6.9 User Profile Database

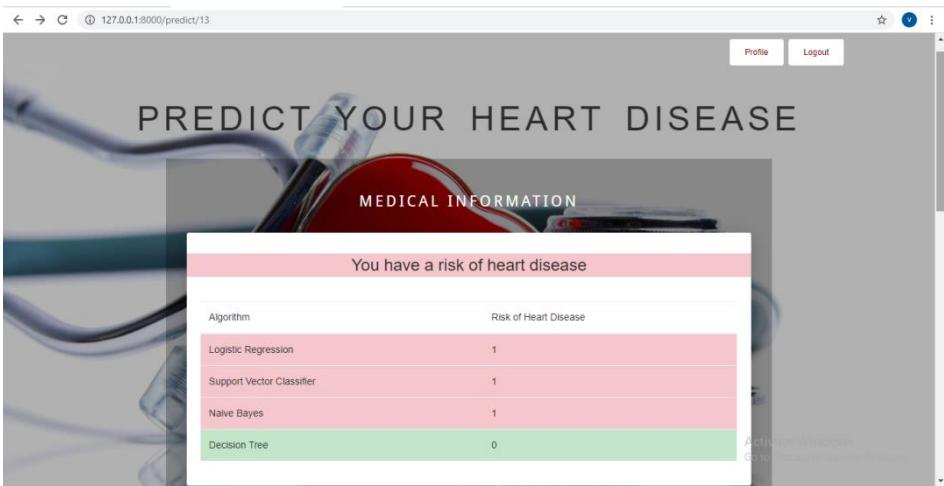


Fig.6.10 Prediction Result

Fig.6.11 Prediction Database



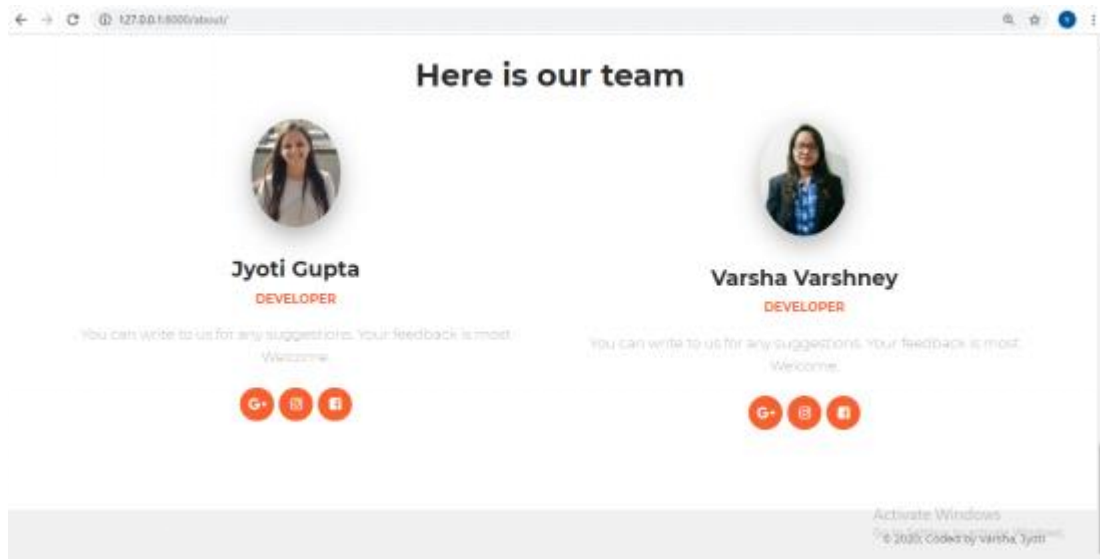


Fig.6.13 Developer Team

Chapter 7

Conclusion

Heart Disease is one of the major concerns for society today. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.

Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. Models based on Naïve Bayes classifier were computationally very fast and have also performed well. SVM performed extremely well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and over fitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.

References

- <https://www.udemy.com/course/machine-learning-basics-classification-models-in-python/learn/lecture/14506724#overview>
- <https://www.udemy.com/course/try-django-2-2-python-web-development/learn/lecture/14258630#overview>
- <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/>
- <https://nevonprojects.com/heart-disease-prediction-project/>