Final Year Project Report

Heart Disease Prediction

BS Software Engineering

Submitted by

Haris Bin Irfan REG. # 37888

.

Yasir Ul Haq REG. # 37728

.

Syed Usama Ahmed Hashmi REG. # 38633

Project Advisor

Dr. Muhammad Raza

February, 2020

Faculty of Engineering, Sciences and Technology

IQRA University, Karachi

INTRODUCTION TO GROUP MEMBERS

Haris Bin Irfan Reg. # 37888
(Leader/Project Developer)
Immediate Contact: (+923042723854, haris.bin.irfan2012@outlook.com)

Yasir Ul Haq Reg. # 37728
(Assistant Project Developer)
Immediate Contact: (+923350237955, yasirkhanchandia8@gmail.com)

Syed Usama Ahmed Hashmi Reg. # 38633
(Designer/Project Support Assistant)
Immediate Contact: (+923423579322, uxamasyed60@gmail.com)

ACKNOWLEDGMENT

First of all, we thank Almighty Allah who praise us with the ability to think, work and deliver what we are assigned to do. Secondly we are so much grateful to HOD of Software Engineering "**Dr. Mansoor Ebrahim**" & all the Faculty Members of Software Engineering of Iqra University for the valuable effort made through during this course, in order to achieve us a great experience of outcome a successful software project at the end. For the support, he had given towards the success of this project and for giving us the domain knowledge and providing necessary information and documents regarding the project. We wish to express our gratitude to our project supervisor "**Dr. M . Raza**", and our co-supervisor "**Miss. Sania Marium**", who gave constant supervision despite there busy schedule, for the technical guidance provided to move us in a correct path, truly for the patience they had when working with our team, and without then it would not be a success. Iqra University and its panel of dedicated lecturers would be acknowledged our thankfulness for the knowledge and the wisdom we have gathered while being students in there. We are also grateful to our family and friends, for supporting and encouraging us to complete this project successfully. Finally, we would like to thank all the colleagues of Iqra University who have been with us in all difficult times with suggestions and supportive words which carry us to make this project a reality.

ABSTRACT

Heart disease is one of the world's most common human diseases, and has a very significant impact on human life. In heart disease the heart couldn't push the amount of blood required to other parts of the body.Precise diagnosis of heart disease on time is critical for the prevention and treatment of heart failure. In many cases, the diagnosis of heart disease by traditional medical practice has been seen as unreliable.Non-invasive methods such as machine learning are reliable and efficient to differentiate the healthy people and people with heart disease.We developed a machine-learning-based diagnostic system for predicting heart disease through the use of heart disease data set in the proposed study.We used six prominent machine learning algorithms, two feature selection algorithms, the process of cross-validation, and one performance evaluation metric for classifiers.The proposed system is capable of quickly recognizing and classifying healthy people with heart disease. Similarly, positive receiver curves and area under the curves were determined for each classifier.We addressed all of the performance evaluation criteria used in this project, feature selection algorithms, preprocessing methods, validation process and classifiers.The proposed system's output has been tested on complete features and a reduced collection of features. The reduction of features impacts the efficiency of classifiers in terms of classifier accuracy and execution time.The proposed decision support system focused on machine learning would help the doctors identify heart patients effectively.

PREFACE

This report is submitted as part of the requirement for the degree of Bachelor of Science (BS(SE) in Software Engineering at the Iqra University. The work in this report is a product of our innovative idea and our efforts. This report can be copied and circulated on condition that the source is acknowledged.

TABLE OF CONTENTS

LIST OF ACRONYMS

Cardiovascular Disease (CVD)

Small and mid-size enterprises (SMEs)

Light Gradient Boosting Machine (LightGBM)

eXtreme Gradient Boosting (XGboost)

Neural Networks (NN)

Artificial Neural Networks (ANN)

Decision Tress (DT)

Naïve Bayes (NB)

Hadoop Distributed File System (HDFS)

Support Vector Machine (SVM)

Data Mining (DM)

Machine learning (ML)

Random forest (RF)

LIST OF FIGURES

Chapter 1: INTRODUCTION

Data mining is the process of finding patterns and trends previously unknown in datasets, and using this information to build predictive models. To extract secret patterns and associations from large databases, data mining incorporates statistical analysis, machine learning and computer technologies. The World Health Statistics 2012 study highlights the fact that one in three people globally has elevated blood pressure-a condition that causes roughly half of all stroke and heart disease deaths. Heart disease, also known as cardiovascular disease (CVD), includes a number of heart attack conditions-not just heart attacks . Heart attack was the leading cause of deaths in the various countries like India. In the United States, heart disease affects a single adult every 34 seconds. Other types of heart disease include coronary heart disease, cardiomyopathy, and cardiovascular disease. The word "cardiovascular disease" involves a wide range of conditions that affect the heart and blood vessels, and how blood is pumped through the body and distributed. Diagnosis is a complicated and important task which requires accurate and efficient execution.

Often the diagnosis is made, based on the experience & expertise of the doctor. This adds to unnecessary patient outcomes & disproportionate medical cost of therapies. Hence an automatic system of medical diagnosis would be extremely beneficial. The healthcare industry collects huge amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Discover of hidden patterns and relationships often go unexploited [1].

1.1 PROJECT OBJECTIVE

A mixture of lifestyle changes, medication and, in some cases, surgery, will effectively manage heart disease. The signs of heart disease can be reduced with the right treatment, and heart functioning increased. The projected outcomes can be used to avoid and thus reduce cost of surgical and other expensive treatments. The heart is a very critical part of the human body. It pumps blood into the whole body If the flow of blood in the body becomes insufficient, organs like the brain suffer, and if the heart stops working completely, death occurs within minutes. Some of the heart disease risk factors are:

1. Smoking: Smokers are twice as likely to have a heart attack as non smokers.
2. Cholesterol: A low cholesterol diet and Tran's saturated fat can help lower cholesterol levels and reduce heart disease risk.
3. Blood pressure: High BP leads to heart Attack.
4. Diabetes: If not controlled, diabetes can result in severe heart damage including heart attack and death.
5. Sedentary life style: Simple leisure activities, such as gardening and walking, can reduce our risk of heart disease.
6. Eating Habits: A healthy diet of the soul, low in salt, calories, saturated fat, trans fat, cholesterol and refined sugars can decrease our risk of heart disease.
7. Stress: Poorly controlled stress can result in heart attacks and strokes.

There are many methods related to prediction of disease. Yet heart-related disease in particular has been analyzed and the level of risk is produced. But there are usually no such tools that are used for specific disease prediction.

The main objective is to predict the Boolean class heart disease prediction, which represents whether a patient has heart disease or not:

- False does not represent heart disease.
- True represents heart disease present.

1.2 RELEVANCE OF DOMAINS

The work done on this project relies on a research work that we have done in the Computer Science Field from past education as well as learning from personal development, this section discusses some of the key features that are extensively used and incorporated in our project.

1.2.1 MACHINE LEARNING

Machine learning is an artificial intelligence (AI) application which provides systems with the ability to learn and improve automatically from experience without being explicitly programmed. Machine learning focuses on the creation of computer programs that can access data and know for themselves. Machine Learning concepts is also applied in our project Because in this Project we use supervised machine learning approach to build our prediction model.

1.2.2 HUMAN COMPUTER INTERFACE (HCI)

In terms of website design, interaction between human computers is concerned with how visitors to the website perceive, navigate and use it. This is an important part of website design. It should be composed of two areas. Firstly the' user experience' in terms of how most users see the website. In terms of website design, interaction between human computers is concerned with how visitors to the website perceive, navigate and use it. This is an important part of website design. It should be composed of two areas. Firstly the' user experience' in terms of how most users see the website. Human computer interaction should take all these technical aspects into consideration to ensure that website design works well under all conditions required. It involves blind users as well as others who use aided technology to access the website.

1.2.3 WEB APPLICATION DEVELOPMENT

Web application development will typically be led by a small development team with a short life-cycle. Front-end web application development is accomplished through client-side programming. Client refers to a computer application like a web browser for example. Programming on the client side will typically use HTML, CSS, and JavaScript. HTML programming will instruct a browser to display the content of Web pages on-screen, while CSS will keep the information displayed in the correct format. JavaScript can run JavaScript code on a web page, which will interactivate some content. Server-side scripting allows programming on the client side, and is used to build the scripts that web applications use. Scripts can be written in multiple languages including Ruby, Java, and Python. Server-side scripting can create an end-user custom interface and mask the source code which makes up the interface. A Server such as MySQL or PostgreSQL may be used to store data during the creation of web applications.

1.3 SCOPE OF THE PROJECT

The aim of the project here is that the incorporation of clinical decision support with computer-based patient records will eliminate medical errors, increase patient safety, decrease unintended discrepancies in training, and improve patient results.This proposal is encouraging because methods of data modeling and evaluation,e.g data mining, have the potential to generate an intelligence-rich enviroment that can help improve the quality of medical decisions dramatically.Prediction of heart disease also helps to reduce the cost of treatment and also increase visualization and analysis services.With deep knowledge in this field and accurate data, large companies are investing heavily in this type of activity to help focus on future incidents and risks involved.This work brings together all historical and current data available as a framework for establishing reasonable expectations of the future.

1.4 CHALLENGES

Medical diagnosis is considered to be an important yet complex task that needs to be accurately and efficiently carried out. It would be very useful to automate the same. Clinical decisions are often made based on the expertise and experience of the practitioner rather than the knowledge-rich information that is hidden in the server. This practice leads to unacceptable biases errors and high medical costs that impact patient's quality of service. Data mining has the potential to create a knowledge-rich environment that can significantly enhance the quality of clinical decision-making. Choosing the best algorithm for the purpose of training. Implementing the high Accuracy rate algorithm in system to generate accurate heart disease predictions.

1.5 MOTIVATION AND NEED

Healthcare organizations face a major challenge in providing quality facilities at affordable rates. Quality service includes correct diagnosis of patients and effective treatments. Bad medical decisions can lead to catastrophic and therefore unacceptable consequences. Hospitals also need to reduce the expense of clinical testing. By using effective computer-based data and/or decision support systems, they can achieve these outcomes. Many hospitals today use some kind of hospital information systems to monitor their health care or patient information. Usually, these systems create huge amounts of data in the form of numbers, text, graphs and photographs. These data are, sadly, rarely used to assist medical decision-making. Such data contain a variety of hidden information that is largely unused. This raises a critical question: "How will we have a tendency to turn data into helpful knowledge that helps health care professionals to create smart medical decisions?" This is the main driving force behind this study.

Chapter 2: TECHNOLOGY BACKGROUND

Machine learning is a data analytics method which automates the building of analytical models. It is a branch of artificial intelligence based on the idea that systems with minimal human intervention can learn from data, identify patterns and make decisions.

2.1 MACHINE LEARNING

Machine learning is without question one of the most powerful and influential innovations in the world today. What is more, we are far from seeing its full potential. No doubt, it will keep making news for the foreseeable future. This article is intended as an introduction to the principles of machine learning, discussing all of the fundamental ideas without being too large.

Machine learning is an instrument for turning knowledge into intelligence. There's been an explosion in data in the last 50 years. If we study it and find the patterns hidden within, this collection of data is useless. Machine learning methods are used to identify the important underlying patterns within complex data that we would otherwise have trouble finding. The latent patterns and experience of a subject can be used to predict future events and conduct all sorts of complex decision taking.

2.1.1 MACHINE LEARNING APPROACHES

While performing Machine Learning, there are many methods that can be taken. Typically they are grouped in the areas listed below. Supervised and Unsupervised are the most widely used and well known methods. Semi-supervised and reinforcement learning is newer and more complex, but the findings have been remarkable. Machine Learning is famous for the No Free Lunch Theorem. It states that no single algorithm works well for all tasks. Each problem you try to solve has their own idiosyncrasies. Therefore, there are lots of algorithms and approaches to match every single quirk problem. A lot more Machine Learning and AI types will continue to be implemented which best suits different problems.

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

2.2 HISTORY OF MACHINE LEARNING

In 1943, the first case of neural networks was when neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on neurons, and how they work. They decided to create a model of this using an electrical circuit and the neural network was born accordingly. Alan Turing created the world-famous Touring Test in 1950. The test is quite clear-in order for a computer to succeed it must be able to convince a human being that it is a human being and not a machine. In 1952 the first computer program that could learn as it was running was seen. It was a game created by Arthur Samuel, that played checkers. In 1958 Frank Rosenblatt developed Perceptron, the first artificial neural network. This was primarily aimed at pattern and form identification.

Another extremely early example of a neural network came in 1959 when Bernard Windrow and Mercian Hoff at Stanford University developed two versions of these. The first one was called ADELINE, and binary patterns could be observed. For example it could predict what the next one would be in a stream of bits. The next version was named MADELINE, and it could remove noise on phone lines, so it had a useful application in the real world. Today, it is still in use.

Despite MADELINE's success, not much progress had been made until the late 1970s for many reasons, mainly the Von Neumann architecture's popularity. This is an architecture where instructions and data are stored in the same memory, which is actually easier to understand than a neural network, and so many people have designed programs on that basis.

2.3 1980s AND 1990s

1982 was the year in which interest in neural networks began to rise again, when John Hopfield proposed that a network with bidirectional lines be developed, similar to how neurons actually work. In addition, in 1982, Japan announced that it was concentrating on more advanced neural networks, which attracted American investment in the field, generating further research in the area.

Neural networks use back propagation (explained in detail in the Introduction to Neural Networks), and this important step came in 1986, when three Stanford psychology researchers decided to extend an algorithm created by Widrow and Hoff in 1962. This enabled the use of multiple layers in a neural network, creating what are known as ' slow learners, ' which will learn for a long time to come.

The late 1980s and 1990s brought little into the industry. However, the IBM Deep Blue computer, which was a chess-playing computer, beat the world chess champion in 1997. Since then, there have been many more developments in the field, such as in 1998, when work on digit recognition at AT&T Bell Labs led to good accuracy in the identification of handwritten postcodes from the US Postal Service. Type in anything that you want.

2.4 21st CENTURY

Many businesses have realized since the beginning of the 21st century that machine learning will increase the calculation potential. That's why they're doing more research in it, to stay ahead of the competition. Some large projects include:

- Google Brain (2012)-This was a deep neural network developed by Google's Jeff Dean that focused on image and video pattern detection. It was able to use the power of Google, thereby making this incomparable to much smaller neural networks. It was also used for monitoring objects in videos on YouTube.
- AlexNet (2012) — In 2012, AlexNet won a large margin in the ImageNet competition which led to the use of GPUs and Convolutionary Neural Networks in machine learning. They also created ReLU, which is an activation function which significantly improves CNN performance.
- DeepFace (2014) — This is a Facebook-based Deep Neural Network that they believe can recognize people with the same precision as human beings do.
- DeepMind (2014)-Google acquired this business and is able to play basic video games at the same level as humans. It managed to beat a professional at game Go in 2016, which is considered one of the hardest board games in the world.
- OpenAI (2015) — This is a non-profit organization that Elon Musk and others have founded to create safe artificial intelligence that can benefit humanity.
- Amazon Machine Learning Project (2015) — This is part of Amazon Web Services which reveals how the majority of big businesses want to participate in machine learning activities. We say this powers many of their internal systems, from commonly used services like search results and Alexa to more ambitious ones like Prime Air and Amazon Go.
- U-net (2015)-This is a CNN architecture specializing in the segmentation of biomedical images. It introduced an equal amount of layers of upsampling and downsampling, and also skipped connections. See the Semantic Segmentation page for more information on what that means.

2.5 SUPERVISED LEARNING

Through supervised learning, the goal is to learn mapping between a set of inputs and outputs (the rules).

For example, weather forecast could be the inputs, and visitors to the beach would be the outputs. The objective of supervised learning would be to learn the mapping which describes the relationship between temperature and number of visitors to the beach.

During the learning process, for example, labeled data is provided from past input and output pairs to teach the model how to behave, hence' supervised' learning. New inputs can then be fed into the forecast temperature for the beach example, and the Machine learning algorithm will then produce a future prediction for visitor numbers.

The key generalization aspect of machine learning is being able to adapt to new inputs and make predictions. We want to optimize generalization in training so the supervised model determines the true underlying' general' relationship. If the model is over trained, we cause the examples used to be over-fitted and the model would not be able to adapt to new, previously unseen inputs.

A side effect to be mindful that the guidance we have adds bias to the learning of supervised learning. The model can only mimic exactly what it was shown, so accurate, impartial examples are very important to show. Supervised learning often typically takes a lot of data before it learns. Obtaining adequately accurately labelled data is often the toughest and costliest aspect

2.5.1 CLASSIFICATION

Classification is used to classify the similar data points into different sections. Machine Learning is used to find rules about how to distinguish the various data points.

But how do you create magical laws? Okay, the rules are uncovered in multiple ways. We all focus on using data and answers to discover rules which separate data points linearly.

A key concept in machine learning is linear separability means is ' can there be a line separating the different data points? '. Simply put, classification methods are attempting to find the best way to distinguish data points by a line.

The lines drawn between groups are known as the boundaries of decisions. The whole field chosen to describe a class is known as the surface area of decision. The decision surface defines that a given class will be assigned if a data point falls within its boundaries.

2.5.2 ENSEMBLE CLASSIFICATION

The classifiers in the ensemble all predict the correct classification of each unseen instance and their predictions are then combined using some form of voting system.

The idea of a random forest of classifiers is introduced and issues relating to the selection of a different training set and/or a different set of attributes from a given dataset when constructing each of the classifiers are discussed.

A number of alternative ways of combining the classifications produced by an ensemble of classifiers are considered. The chapter concludes with a brief discussion of a distributed processing approach to dealing with the large amount of computation often required to generate an ensemble.

2.6 PRE-EXISTING SYSTEM

The pre-existing system operates on deep learning as well as data mining collections. By applying the powerful prediction algorithm, the existing system modules produce a comprehensive report. The main aim of the existing system was to compare and check the pre-patient with disease outputs and new patient disease and then identify future possibilities of cardiac disease for a specific patient. By implementing following classification models:

- Naive Bayes
- Logistic Regression
- Decision Tree
- Random Forest
- LightGBM
- XGboost.

But the accuracy is much less by using all the existing systems.

Drawbacks of the existing system

1.) It's very difficult to maintain the system.
2.) There is a high chance of obtaining inaccurate results.
3.) GUI is much less user-friendly.
4.) It takes longer to process the activities.

2.7 LITERATURE REVIEW

There have been numerous studies that focus on heart disease diagnosis. Various data mining techniques have been applied for diagnosis and different probabilities have been achieved for different methods.

In this chapter, numerous DM techniques implemented within recent years have been examined and updated for prediction of heart disease. One of the greatest strengths that can be extended to various health science issues is the wide range of methodologies and techniques in DM [2]. Scientists have used specific DM approaches such as association rule mining, clustering, classification to improve disease diagnosis with good accuracy and low risk of errors. Existing literature indicates that DM plays an effective role in the predictive mode of heart disease over clustering, association rule and regression through classification. The study also highlights some research using only one DM method for the diagnosis of heart disease, while many of the other studies have used ensemble / hybrid DM methods in the search for better model accuracy and reliability.

Lee et al suggested a numerical and classification-based approach for the creation of a multi-parameter linear and nonlinear heart rate variability (HRV) relationship [3]. Experimental work was carried out using linear and nonlinear HRV parameters while applying Naïve Bayesian, association rules and SVM classifiers. SVM had better accuracy than other classifiers. Tan and Teoh proposed a hybrid approach based on a wrapper approach based on classification and GAs. The SVM classified the patterns into desired classes based on the sub-set attribute identified by GA. The data set was used for the UCI Machine Learning Repository and the study showed the GA – SVM hybrid approach's effectiveness. In the multi-class setting, an average accuracy of 84.07 percent improved the efficiency of the GA – SVM hybrid model.

(Polaraju, Durga Prasad, & Tech Scholar, 2017)Proposed Heart Disease Prediction using Multiple Regression Model and it shows that Multiple Linear Regression is suitable for predicting possibility of heart disease. Proposed Heart Disease Prediction using Multiple Regression Model and it shows that

Multiple Linear Regression is suitable for predicting possibility of heart disease. The work is carried out using training data set consisting of 3000 instances with 13 different attributes previously mentioned. The data set is divided into two parts, 80% of the data used for training and 20% used for testing [4].

(Deepika & Seema, 2017) Focuses on techniques capable of predicting chronic disease by using Naïve Bayes, Decision Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN) data from historical health records. In order to measure better performance at an accurate rate, a comparative study is performed on classifiers. SVM gives the highest accuracy rate from this experiment, while Naïve Bayes gives the highest accuracy for diabetes.

(Sai & Reddy, 2017) Proposed prediction of heart disease using data mining ANN algorithm. There was a need to develop a new method that can predict heart disease due to increased costs of treatment for heart disease. Prediction model is used after evaluation to predict the patient's condition based on different parameters such as heart beat rate, blood pressure, cholesterol, etc. The system's reliability is demonstrated in java [5].

(Sultana, Haider, & Uddin, 2017) Proposed cardiovascular disease study. This paper proposed techniques for data mining to predict the disease. It is intended to provide the overview of current techniques to extract information from the database and will be useful for practitioners of health care. Based on the time taken to build the decision tree for the system, performance can be obtained. The primary goal is to predict the disease with fewer attributes [6].

(A & Naik, 2016) It is recommended that a predictive system be developed to diagnose heart disease from the medical data set of the patient. 13 The system was considered to be based on risk factors of input attributes. Data cleaning and data integration were performed after analyzing the data from the dataset. To predict heart disease, he used k-means and naive Bayes. This paper is designed to build the system using the diagnostic historical heart database. For the construction of the system, 13 attributes were considered. Data mining techniques such as clustering, classification methods can be used to retrieve information from the database. The Cleveland Heart Database used 13 attributes with a total of 300 files. This model is intended to predict whether or not the patient has heart disease based on 13 attribute values [7].

(Soni, Ansari, & Sharma, 2011) Proposed use of non-linear classification algorithm to predict heart disease. It is suggested to use bigdata tools like Hadoop Distributed File System (HDFS), Map reduction along with SVM for heart disease prediction with optimized set of attributes. This work carried out an investigation into the use of various techniques of data mining to predict heart disease. It suggests using HDFS to store large data in different nodes and use SVM simultaneously to execute the prediction algorithm in more than one node. SVM is used in parallel fashion, resulting in higher computational time than sequential SVM [8].

(Yan and Zheng Proposed) an actual GA-based diagnostic system for heart disease while applying the sub-setting of critical clinical features. The prediction system was designed to identify five major heart diseases, using 352 heart disease data instances with their corresponding diagnostic weights to support or deny each heart disease diagnosis. This provided the functional prediction framework for heart disease with reasonably high accuracy. Based on the machine learning literature and DM, Austin et al developed alternative classification schemes that include bootstrap aggregation bagging, RFs and boosting and SVMs. Research has shown that current DM and ensemble methods offer high precision and performance advantages [9].

Chapter 3: REQUIREMENT & METHODOLOGY

Methodology is a system that includes steps to convert raw data into recognized data patterns to extract user knowledge. The approach suggested involves measures, referred to as the pre-processing stage where the information are thoroughly analyzed. It will deal with missing values, balance information and normalize attributes depending on the algorithms used. Using classification models and EnsembleVoteClassifier, predictive analysis of the data is done after pre-processing of data. Eventually, prescriptive modeling is performed, where different performance metrics are used to test the predictive model in terms of performance and accuracy.

3.1 PROPOSED SYSTEM

The proposed system has details that are categorized according to features in it if patients have heart disease or not. This proposed system will attempt to use this dataset to create a model that tries to predict whether or not a person has this disease.In this proposed system, we use 4 Classification algorithm's BaggingClassifier(GaussianNB), MLPClassifier, SVC and AdaBoostClassifier(RandomForestClassifier). Calculating the score using the sklearn library. Implements VotingClassifier to get best accuracy results. Using the Comparing Models and Confusion Matrix to finally analyze the results. It should be grouped into separate structured data depending on the features of the patient's heart from the dataset we have. First, we have to import the dataset. Read the dataset, the data should contain different variables like age, gender, sex, cp(chest pain),slope, target etc. The data should be explored so that the information is verified. The records are divided into two datasets: dataset training 80% and dataset testing 20%. To avoid bias, the records are selected randomly for each set. Using VotingClassifier, we need to create a model that predicts the disease of the patient.

3.2 DATA SET FOR EXPERIMENT

The data set for this research has been taken from the UCI data repository. Used data is freely available from the UCI Machine Learning Repository [10]. The Cleveland data were collected from the above-mentioned DM warehouse. This database includes 76 attributes and 14 attributes have been picked after neglecting redundant and obsolete attributes.

The list of 14 attributes and their brief description are shown in above table. In particular, several researchers used the Cleveland datasets and found that they were appropriate for the creation of a mining model due to lower missing values and outliers. Before they were submitted to the proposed algorithm for training and testing, the data were cleaned and preprocessed. Therefore, 303 are the appropriate instances for the development of supervised machine-learning model building. Attribute selection technique was used for further data reduction to make patterns easier and more comprehensible, but negligible effects were found on model performance observations undertaken in this study. All 13 attributes are considered in order to develop a model for the classifier and to achieve a predictive outcome for heart disease. GaussianNB, MLPClassifier, SVM, RandomForestClassifier algorithm are the classification techniques used in this research. The EnsembleVoteClassifier was used to evaluate the algorithms involved in the classification. The model was built using the Google Colab tool. In these experiments, 5-fold cross-validations were used to divide the data set into training and test sets that meets the model training and testing purpose requirement. As a result, the accuracy rate of this study was over 85%.

S no	Input variables	Description	Options
1	Age	Age in years	Continuous value
2	Sex	1 = male, 0= female	Male, female
3	Cp	Chest pain type	Chest pain type. Values from 1 to 4. 1: typical angina. 2: atypical angina. 3: non-anginal pain. 4: asymptomatic.
4	Trestbps (blood pressure)	Resting blood pressure in mmHg	Continuous value in mmHg
5	Chol (cholesterol)	Serum cholesterol in mm/dL	Continuous value in mm/dL
6	Fbs (fasting blood sugar)	Fasting blood sugar in mg/dL	Fasting blood sugar attributes value "1" for greater than 120 mg/dL, else the attribute value is 0 (false). Value 1 = true. Value 0 = false.
7	Restecg (ECG)	Electrocardiographic results (ECG result)	Resting electrocardiographic results value ranging from 0 to 2. 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
8	Thalach (heart rate)	Maximum heart rate	Continuous value categorized into normal and abnormal
9	Exang	Exercise Induced angina	Exercise induced angina. Values from 0 to 1. Value 1 = yes. Value 0 = no.
10	Old peak	ST depression induced by exercise relative to rest	Continuous values
11	Slope	Slope of the peak exercise ST segment	Measure of slope for peak exercise. Values can be 1, 2, or 3. Value 1: up sloping. Value 2: flat. Value 3: down sloping.
12	Ca	Number of major vessels colored by fluoroscopy	Number of major vessels from 0 to 3
13	Thal	Heart rate of patient	Represents heart rate of the patient. It can take values 3, 6, or 7. Value 3 = normal. Value 6 = fixed defect. Value 7 = reversible defect.
14	Class	Class labels (predicted outcome)	Contains a numeric value between 0 and 1. Each value represents heart disease or absence of disease. Value 0: absence of heart disease. Value 1: presence of heart disease.

Figure 1 13 Attributes Of Data Set

3.3 CLASSIFIERS USED FOR EXPERIMENT

3.3.1 GAUSSIAN NB:

A special type of NB algorithm is a Gaussian Naive Bayes algorithm. It is specifically used when there are continuous values for the features. It is also assumed that all features follow the normal distribution of a Gaussian distribution.

Bayes' Theorem

Theorem of Bayes considers the probability of an event occurring given the likelihood of another event already occurring. Mathematically, the theorem of Bayes is stated as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where are events A and B and events P(B)? 0.

- Basically, given that event B is true, we are trying to find the probability of event A. Event B is also known as proof.
- P(A) is the priori of A (previous probability, i.e. probability of occurrence before proof is seen). The proof is an undefined instance's attribute value (here it is case B).
- P(A) is a posteriori probability of B, i.e. probability of occurrence after proof.

Now, with regard to our dataset, Bayes ' theorem can be applied as follows:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where y is a variable of class and X is a vector of dependence (size n) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

3.3.2 MLP CLASSIFIER:

A multilayer perceptron (MLP) is an artificial neural feedforward network producing a collection of outputs from a set of inputs. Multiple layers of input nodes connected as a direct graph between the input and output layers characterize an MLP. MLP uses network learning backpropagation. MLP is a form of deep learning\

3.3.3 SVM:

SVM is a rough realization of structure-based risk minimization and linear / non-linear data categorization. SVMs optimize the dividing hyperplane gap. Each training sample sub-set (support vectors) defines boundary decision functions. SVM includes three steps, the creation of the support vector, the formation of the maximum distance between the found points and the boundary of the perpendicular decision. For many practical problems, the maximum margin linear classifier is an inapplicable approach. For practical applications, where the non-linear separable data set is to be

separated by hyperplane, the non-linear data is to be mapped via kernel functions to another feature space.

The data points are separated by hyperplane in the original input space, and the functions of the kernel map non-linear training samples to high-dimensional space. A search for the best hyperplane is then performed to split the transformed data into two different classes. For classification purposes, the margin of the hyperplane is maximized thus reducing classification errors. For multi-dimensional hyperplane, the algorithm predicts the risk of heart disease and optimally categorizes the data into different labels by generating the margin between data clusters.

3.3.4 RANDOM FOREST CLASSIFIER:

RandomForestClassifier, one of the most accurate algorithms for machine learning, is a decision-based ensemble classifier approach that contains a flowchart like a tree structure. RF is a combination of { h(x, n) } tree-structured classifiers where random trees for data classification are distributed for "x" data input and "n.". The decision tree in a random tree-structured forest casts a vote indicating the knowledge group decision. RF uses the Gini index to assess that tree's final category. This algorithm selects optimal attributes for each tree from the total number of input attributes "M." Using this selected attribute to develop a decision tree template, the best possible split is generated using the Gini index. For each of the branches, this is an iterative process until the terminating nodes are too small to further break. For data set x classes with "n", You can define Gini-index, Gini(x) by:

$$Gini(X) = \sum_{i=1}^{n} (Rj)^2$$

Where "Rj" in the data set "X" is the relative frequency of class j The split with the lowest Gini index is picked at the split value in RF.

3.3.4.1 Random forest algorithm applications:

The random algorithm used in wide varieties applications. Below are some of the applications where random algorithm forest is commonly used.

1. Banking
2. Medicine
3. Stock Market
4. E-commerce

1.BANKING:

Random forest algorithm commonly used in two major applications in the banking sector. Which are for keeping the consumer loyal and keeping the customers of the scam.

The loyal customer means not the customer who pays well, but also the customer whom can take the huge amount as loan and pays the loan interest properly to the bank. As the growth of the bank purely depends on the loyal customers. The bank customers data highly analyzed to find the pattern for the loyal customer based the customer details.

Similarly, customers that aren't lucrative for the bank need to be found, such as taking the loan and paying the interest on the loan correctly or seeking outgoing customers. If the bank can define certain

customer forms before giving the customer the loan. Bank would have a chance not to authorize the loan to these clients. In this case, the random forest algorithm is also used to classify the clients that are not profitable for the bank.

2.MEDICINE

In the field of medicine, random forest algorithm is used to determine the right combination of components for validating the medication. Random forest algorithm also helps in the diagnosis of the disease by reviewing the medical history of the patient.

3.STOCK MARKET

On the stock market, the random forest algorithm used for buying the particular stock to define the stock actions as well as the predicted loss or profit.

4.E-COMMERCE

In e-commerce, the random forest used on similar types of customers only in the limited segment of the recommendation engine to classify the probable hood of customer who likes the recommended products. High-end GPU systems are needed to run random forest algorithms on very large datasets. If you don't have any program with GPU. The machine learning models can always be run in the cloud host desktop You can use online cloud computing platform to run high-end machine learning models from every corner of the world sitting.

When we use the random forest algorithm in any classification problem, the problem of overfitting will never come. The same random forest algorithm can be used for tasks of classification as well as regression.

3.3.5 ADABOOST CLASSIFIER:

AdaBoost, short for Adaptive Boosting, is a meta-algorithm for machine learning developed by Yoav Freund and Robert Schapiro, who received the Gödel Prize for their research in 2003. It can be used to improve performance in conjunction with many other forms of learning algorithms.

3.3.6 ENSEMBLE VOTE CLASSIFIER:

Ensemble method is a well-proven technique used in research to achieve highly accurate data classification by hybridizing multiple classifiers in order to achieve more reliable and accurate prediction results. The enhanced quality of prediction is a well-known in-built ensemble methodology feature. The EnsembleVoteClassifier is a meta-classifier for combining, for majority or plurality voting, similar or conceptually different machine learning classifiers. Through hard voting, we predict the label of the final class as the class label that the classification models predicted most frequently. Hard voting is where a model to make the final prediction by a simple majority vote for accuracy is chosen from an ensemble. Soft Voting can only be achieved if all the classifiers can measure the outcomes probabilities.

3.4 HARDWARE & SOFTWARE REQUIREMENTS

3.4.1 Hardware requirements

Hardware	Minimum requirements
Computer	2 GHz minimum, multi-core processor
Memory (RAM)	At least 4GB, preferably higher
Hard disk space	At least 10 GB

3.4.2 Software requirements

Software	Minimum requirements
Operating system	Windows Server 2012 R2 or above
IIS (Internet Information Services)	Version 8 or higher

3.4.3 Browser Compatibility

While the Heart Disease Prediction System may work with various browsers, we recommend using one of the following for the best experience:

- Microsoft Internet Explorer (IE) v11 or later
- Microsoft Edge
- Firefox v45 or later
- Google Chrome v58 or later.

In other browsers, performance and graphical consistency may suffer.

3.5 REQUIREMENTS

3.5.1 FUNCTIONAL REQUIREMENTS :

- System shall predict presence of heart disease with the given symptom values provided by user.
- System shall save feedback about disease prediction from user.
- System shall maintain heart disease prediction results history of every user.
- User must Signup/Login before using main functionality of the system.
- Admin shall be able to add/Delete/Update user.
- Admin shall be able to Import/Delete/Update dataset.
- Admin shall be able train the model with the imported dataset.

3.5.2 NON FUNCTIONAL REQUIREMENTS :

Non-functional requirements include requirements for performance, security and portability.
Performance requirements:
Since this program is going to be web-based, it needs a large server machine with high-speed(100mbs) internet connectivity. The server machine must have a powerful CPU and high-speed internet connectivity so it can accommodate multiple users concurrently and the storage space is the other performance requirement. Higher storage means more users and greater space per device, higher the storage higher performance. The user-side reliability requirement is that the web application must be designed as a lightweight web application so that it can operate on almost any network even with slower internet connections. Expected number of concurrent devices, application must be able to deal concurrently with 100 users. The system server also handles at least 1000 client at any time.

Security requirements:
Because this program is hosting on a cloud sever, all user information must be stored on the cloud sever. The user's workspace shall be accessible exclusively by means of user credentials that no other user can access the user's private data. User privileges are also limited in order to prevent users from damaging the system by the programs they run or by the commands the run on the terminal. Since all data is transmitted through the network, the system shall implement a tool for encryption and decryption that only the intended user can decode the data and work on the data by using HTTPS (HTTPS is a combination of HTTP with SSL/TLS).

Portability requirements:
The main purpose of web-based IDE architecture is to improve software development framework portability. Application can run on a variety of platforms and variety of connection speeds to boost portability. As described in the section on performance requirements, code must be lightweight so it can run slow internet connectivity on the computer. Simple libraries and tools are to be used during development to keep the program lightweight. Portability also means running without any additional effort on most different platforms. To achieve this, web applications will be built using technology and tools provided by all common web browsers and operating systems like HTML5, JQuery, etc.

3.6 KEY BENEFITS AND BENEFICIARIE

Predictive modeling is useful as it offers accurate insight into any problem and allows users to create outcomes. In order to maintain a competitive advantage, insight into future events and results that challenge key assumptions is crucial.

An advertisement of the product will be placed on some social and high traffic websites through which user would be able to get more traffic on our heart disease prediction website. Banners are placed in various universities as well.

3.7 TOOLS & LANGUAGES

- Python
- Visual Studio Code
- PostgreSQL
- Django
- Google Colab
- Adobe Photoshop
- Adobe XD
- Adobe Illustrator
- HTML / CSS / JavaScript

3.8 PROJECT FEATURES

- Efficiency: Diagnose within few seconds.
- Portable: Application can run on a variety of platforms and variety of connection speeds to boost portability.
- Cost Effective: System requires low maintenance.

Chapter 4: PROJECT PLAN & INITIAL DESIGN

4.1 PROJECT PLAN

4.1.1 PHASES

This project is completed in different Phases in the earlier stages the requirement analysis & design phase is done by implementing the water fall model. We use water fall model because it can be implemented for any size of the project. In our project every stage has to be done separately at the right time.

All the Documentations can be done in every stage of the project flow. The testing and the debugging process can be done in every of the stage. We attempted to make sure that all the features of this application is implemented in every phase. Weekly Evaluation is done on the basis of weekly task which is assigned by our course teacher.

When the 3D Vision application is in a testing phase meant that changes could be applied to the application quickly with the design and development stages taking place simultaneously, using the information collected from the teacher's design & functionality of the application until it satisfies the goals of this project, which can be proven in the assessment phase.

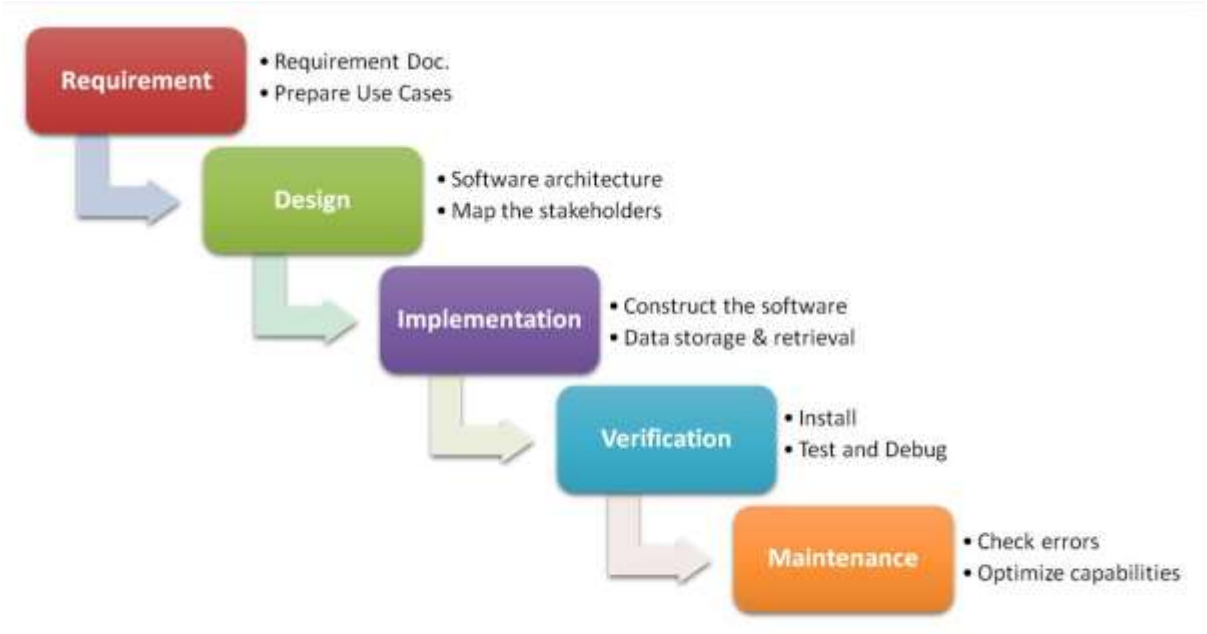


Figure 2 Project Implementation Phases

4.1.2 GANTT CHART

Task Name	Duration	Start	Finish	Predecessors
Project Idea	13 days	Mon 9/2/19	Wed 9/18/19	
System Requirements Gathering	28 days	Thu 9/19/19	Mon 10/28/19	1
Designing System Diagrams	28 days	Tue 10/29/19	Thu 12/5/19	2
Logo Design, Product name and prototypes.	21 days	Tue 12/24/19	Tue 1/21/20	3
Complete Report	12 days	Wed 1/22/20	Thu 2/6/20	4
Implementation	68 days	Fri 2/7/20	Tue 5/12/20	5
Testing	12 days	Wed 5/13/20	Thu 5/28/20	6
Delivery/Training	12 days	Fri 5/29/20	Mon 6/15/20	7

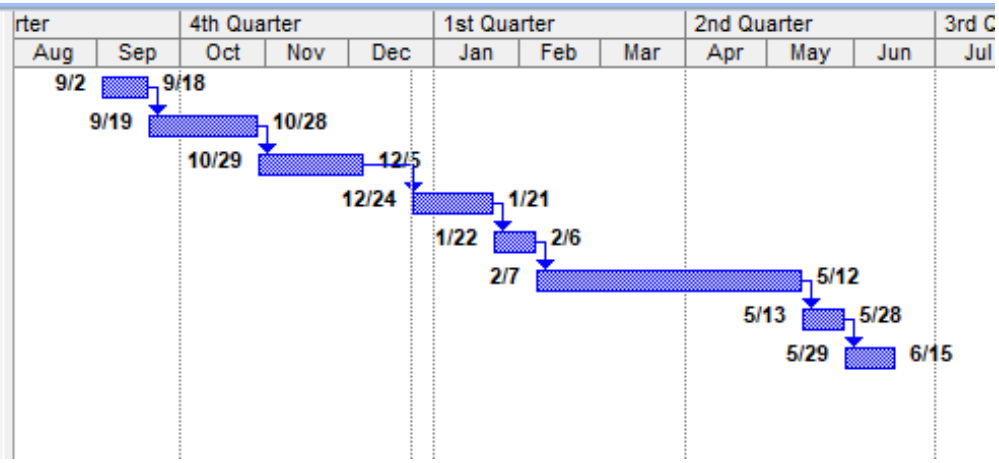


Figure 3 Project Schedule / Milestone Chart

4.2 RISK ANALYSIS

Risk	Likelihood (Low, Med, High)	Impact	Mitigation
Project completion delays	Med	Serious	Paying a lot of attention to project planning and Constantly track and measure the progress.
The budget is not enough / exceeds	Med	High	Place a condition in the contract if more expenses are incurred, it must be covered by the funded party to avoid the risk.
Lack of team members cooperation	Med	High	Tools for project management will be used to support individual activities.
Technology	High	Critical	Best performance servers will be used for hosting of the website.

Figure 4 Risk Analysis

4.3 PROPOSED MODEL

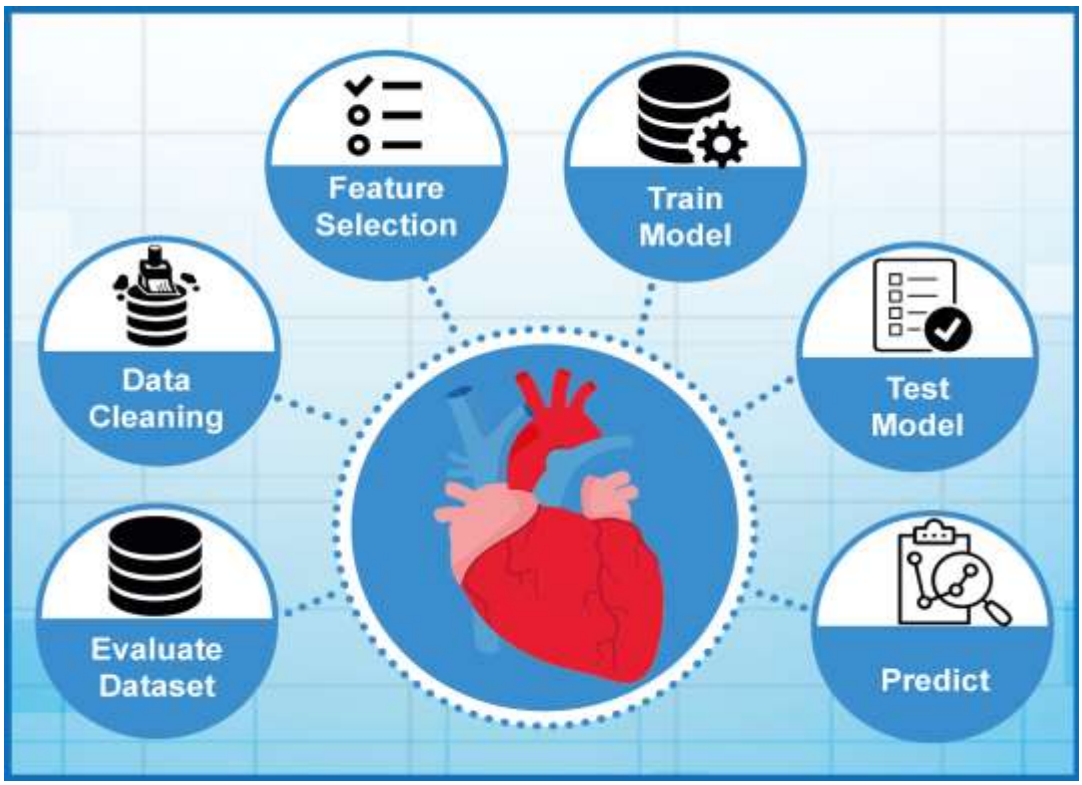


Figure 5 Proposed Model Diagram

4.4 USE CASE

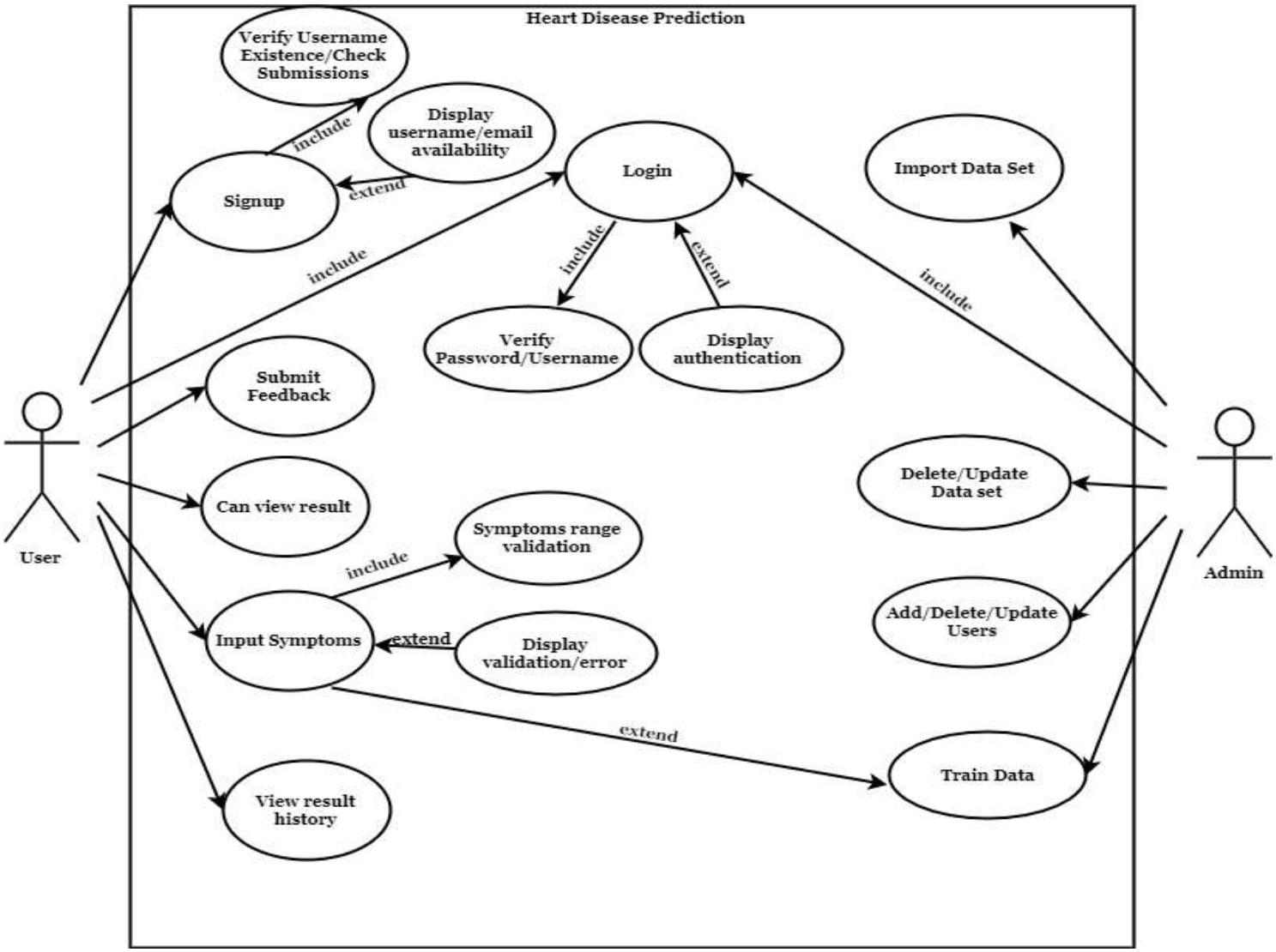


Figure 6 Usecase Diagram

4.5 USE CASE NARRATIVES :

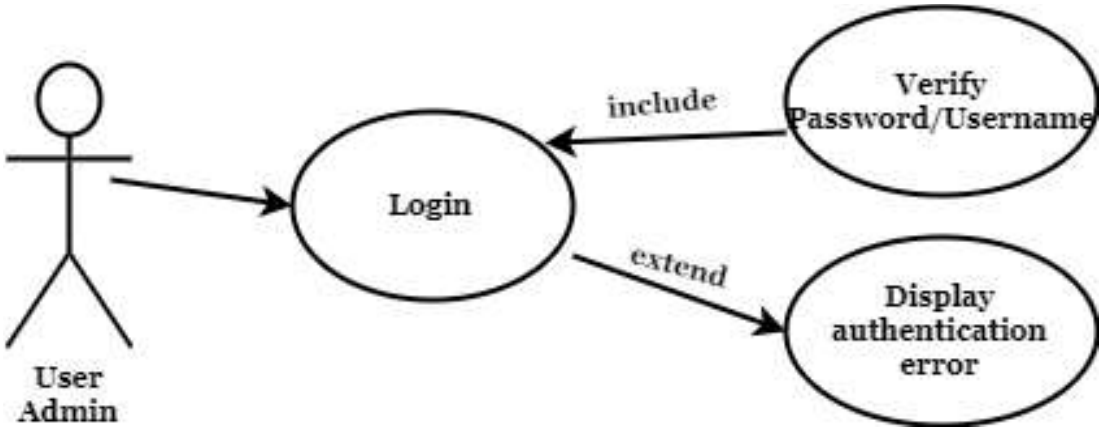


Figure 5.5.1 Login

Use Case Name:	Login	
ID:	HDP-001	
Actors Involved:	User, Administrator	
Brief Description	Actor enters the email address and password to login to the system.	
Pre-Conditions	For User Signup[HDP-002]	
Post-Conditions	Verify Password/Username. [HDP-011] Display authentication error.[HDP-010]	
Priority:	High	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter email address and password. 2. Clicks the submit button.	1. System displays the actor mail page on successful login. 2. System displays error message on invalid login.

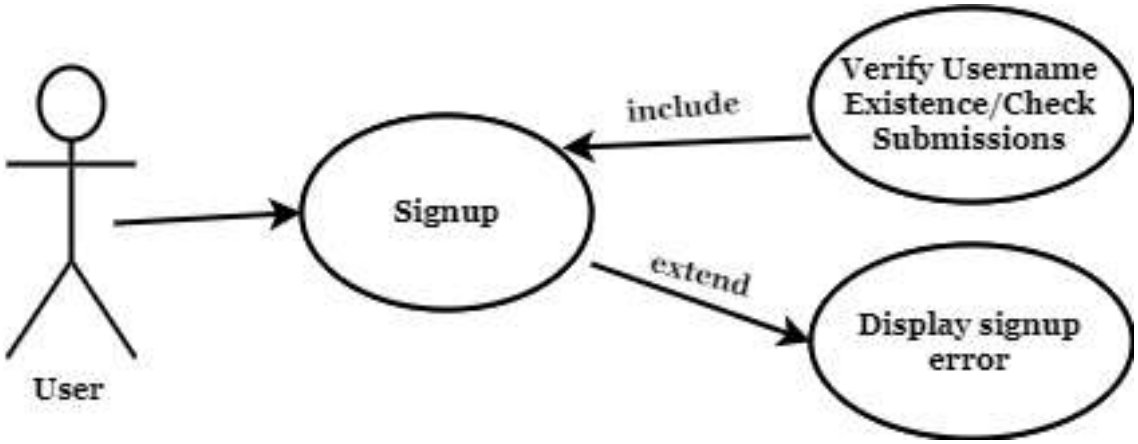


Figure 5.5.2 Signup

Use Case Name:	Signup	
ID:	HDP-002	
Actors Involved:	User	
Brief Description	Actor enters the username email address and password all other info required to signup to the system.	
Pre-Conditions	-	
Post-Conditions	Verify Username Availiblity/Check Submissions.[HDP-013] Display signup error. [HDP-012]	
Priority:	High	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter username,email address and password etc all info. 2. Clicks the submit button.	1. System Verify Username Availiblity. 2. System check submissions and displays error message on invalid submissions.

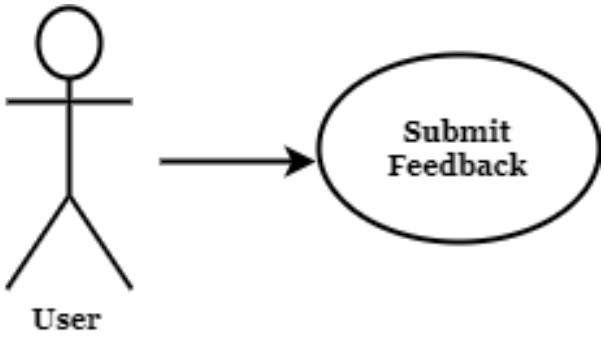


Figure 5.5.3 Submit Feedback

Use Case Name:	Submit Feedback	
ID:	HDP-003	
Actors Involved:	User	
.Brief Description	Actor enters the feedback then submit to the system.	
Pre-Conditions	Login.[HDP-001] Input Sysmptoms.[HDP-006]	
Post-Conditions	Input Sysmptoms.[HDP-006] view result history[HDP-004]	
Priority:	low	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter feedback. 2. Clicks the submit button.	1. System displays feedback. 2. System submits feedback to database.

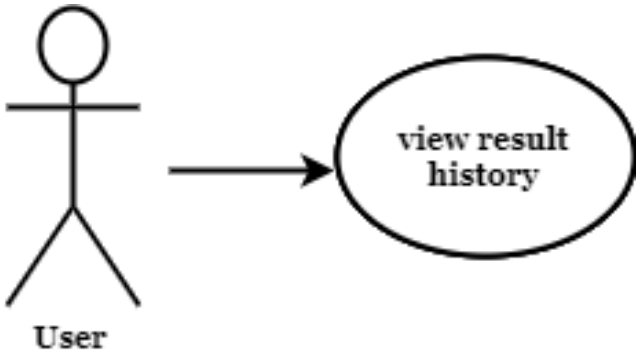


Figure 5.5.3 View result history

Use Case Name:	View result history	
ID:	HDP-004	
Actors Involved:	User	
Brief Description	Actor can view previous result history.	
Pre-Conditions	Login.[HDP-001]	
Post-Conditions	Input Sysmptoms.[HDP-006] Submit Feedback.[HDP-003]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Click View Results button.	1. System displays history of previous results.

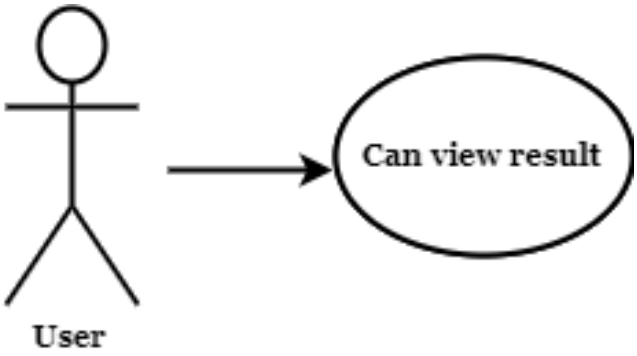


Figure 5.5.4 Input Sysmptoms

Use Case Name:	View result	
ID:	HDP-005	
Actors Involved:	User	
Brief Description	Actor can view current predicted result.	
Pre-Conditions	Login.[HDP-001] Input Sysmptoms.[HDP-006]	
Post-Conditions	Submit Feedback.[HDP-003]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Clicks the submit button.	1. Systems show the current predicted result.

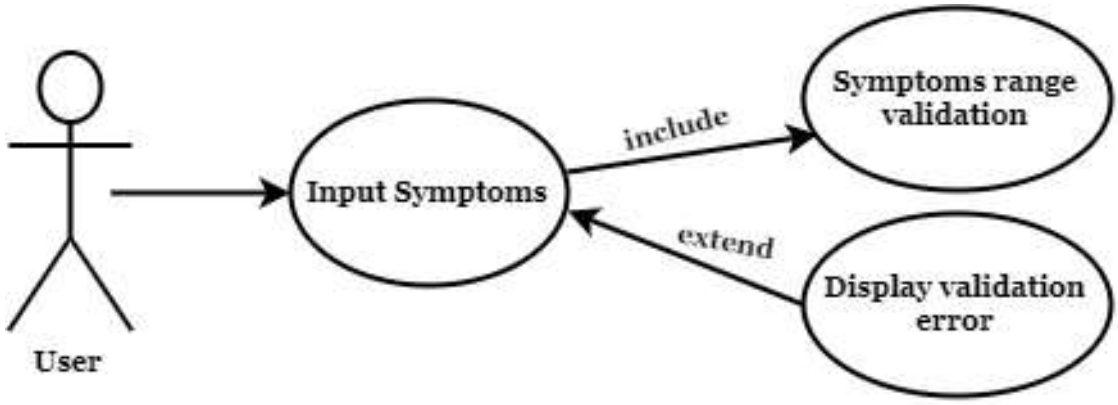


Figure 5.5.5 Input Sysmptoms

Use Case Name:	Input Sysmptoms	
ID:	HDP-006	
Actors Involved:	User	
Brief Description	Actor inputs Sysmptoms.	
Pre-Conditions	Login.[HDP-001]	
Post-Conditions	Symptoms range Validation.[HDP-015] Display validation error.[HDP-011] View result.[HDP-005]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter Sysmptoms. 2. Clicks the submit button. s[HDP-005]	1. System collects inputs Sysmptoms. 2. System submits to database for prediction and displays the result.

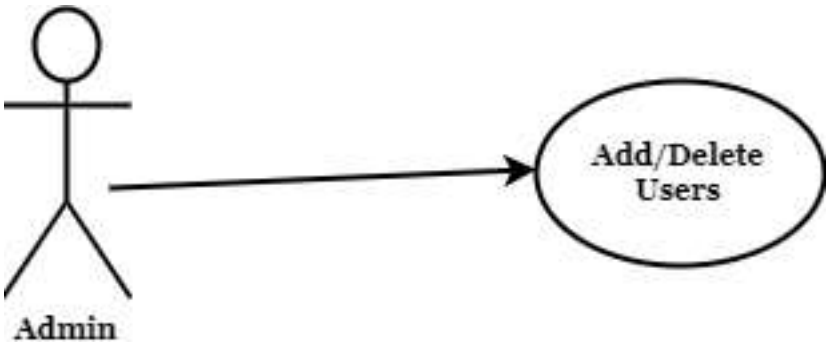


Figure 5.5.5 Input Sysmptoms

Use Case Name:	Add/Delete Users	
ID:	HDP-007	
Actors Involved:	Admin	
Brief Description	Actor can Add/Delete Users.	
Pre-Conditions	Login.[HDP-001]	
Post-Conditions	Login.[HPD-001] for User. Import Dataset.[HDP-009]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter user info select options to delete, update or add new user. 2. Clicks the submit button.	1. System collects input user info. 2. System performs required tasks by admin update/delete/add.

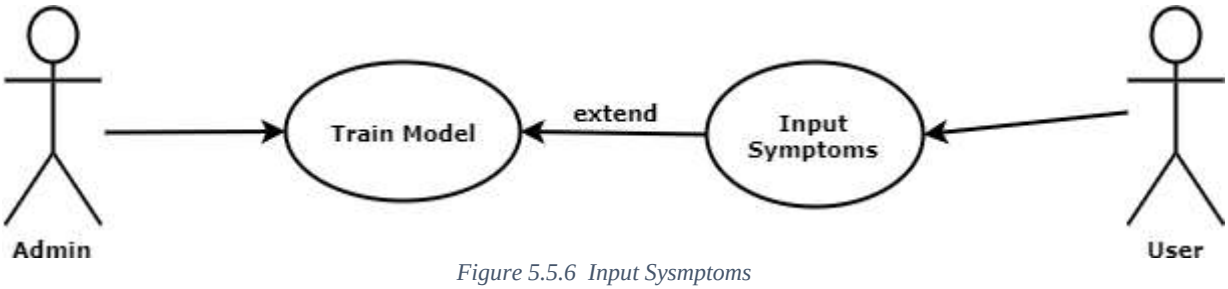


Figure 5.5.6 Input Sysmptoms

Use Case Name:	Train Model	
ID:	HDP-008	
Actors Involved:	Admin	
Brief Description	Actor can train model by using uploaded data set.	
Pre-Conditions	Login.[HDP-001] Import Data set.[HDP-009]	
Post-Conditions	Input Symptoms.[HDP-006]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Clicks the Train model sbutton.	1. System checks data set availability then trains the model with the available data set.

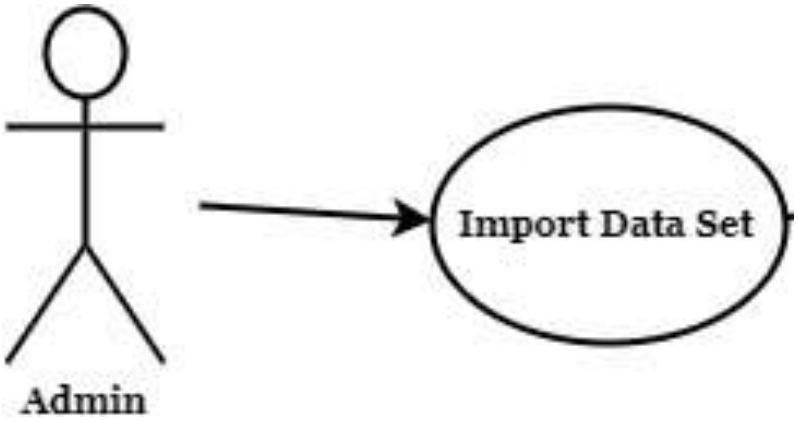


Figure 5.5.7 Import Dataset

Use Case Name:	Import Dataset	
ID:	HDP-009	
Actors Involved:	Admin	
Brief Description	Actor can Add Dataset	
Pre-Conditions	Login.[HDP-001]	
Post-Conditions	Train Model[HDP-008]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Upload Dataset csv file. 2. Clicks the submit button.	1. System collects csv file. 2. System Validate data set format and updates the database.

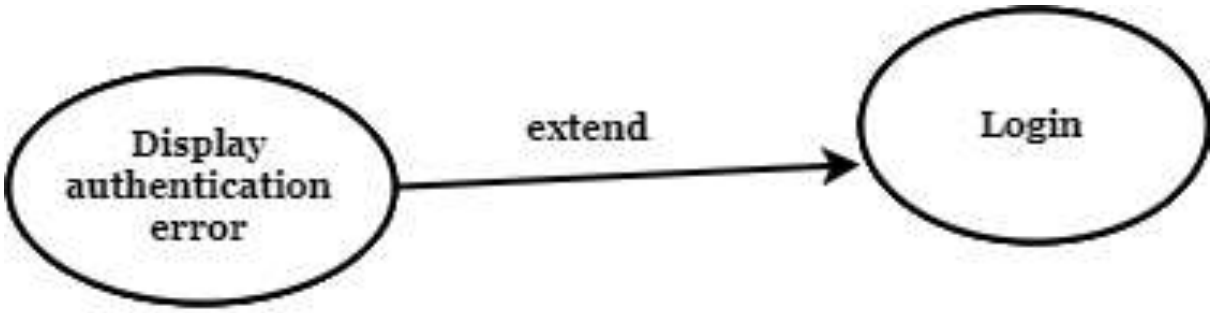


Figure 5.5.8 Display authentication error

Use Case Name:	Display authentication error	
ID:	HDP-010	
Actors Involved:	Admin,User	
Brief Description	Displays Login authentication error after actor incorrect username password inputs.	
Pre-Conditions	Login[HDP-001] Verify Password/Username[HDP-011]	
Post-Conditions	Login[HDP-001]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enters Username password. 2. Clicks Login button.	1. System collects inputs . 2. Authenticate inputs with records in database if wrong display authentication error.

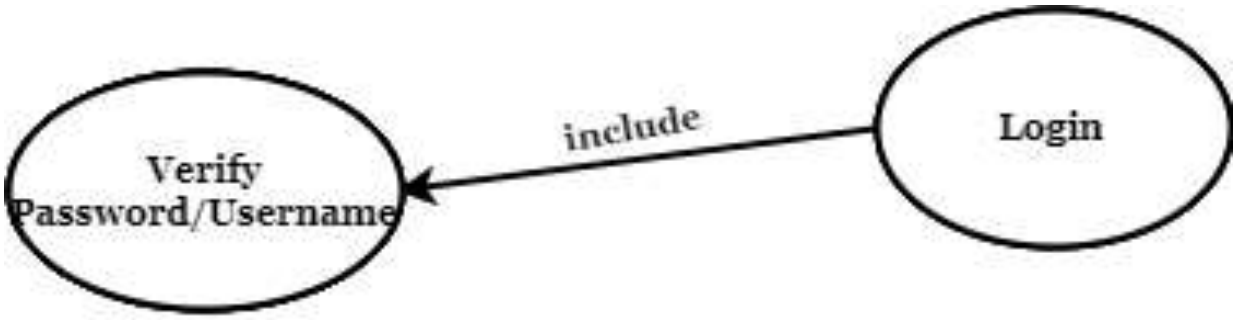


Figure 5.5.9 Verify Password/Username

Use Case Name:	Verify Password/Username	
ID:	HDP-011	
Actors Involved:	Admin,User	
Brief Description	Verify Password/Username that user entered with records in database.	
Pre-Conditions	Login[HDP-001]	
Post-Conditions	Display authentication error[HDP-010]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enters Username password. 2. Clicks Login button.	3. System collects inputs . 4. Authenticate inputs with records in database.

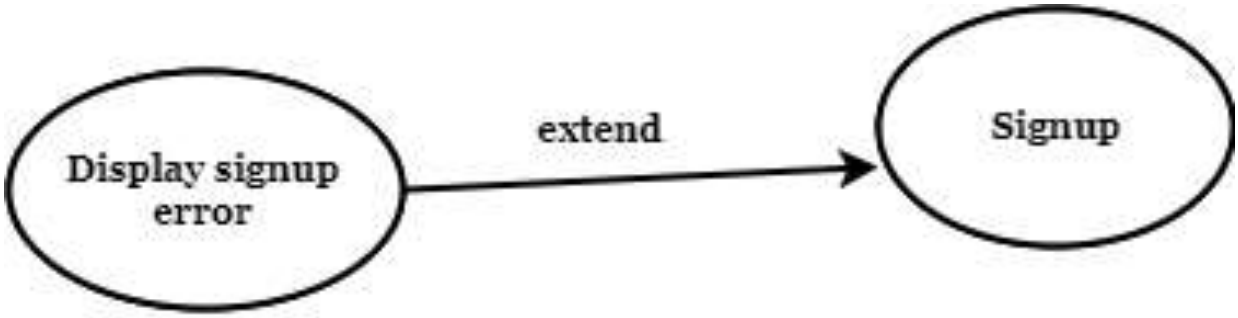


Figure 5.5.10 Display signup error

Use Case Name:	Display signup error	
ID:	HDP-012	
Actors Involved:	User	
Brief Description	Actor gets a error if the signup info has any issue(ex Username aready taken).	
Pre-Conditions	Signup[HDP-002] Verify Username Existence/Check submissions [HDP-013]	
Post-Conditions	Signup[HDP-002]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter All info required in signup form. 2. Clicks the submit button.	1. System collects signup form inputs. 2. Verify submissions if correct submits to database else display signup error.

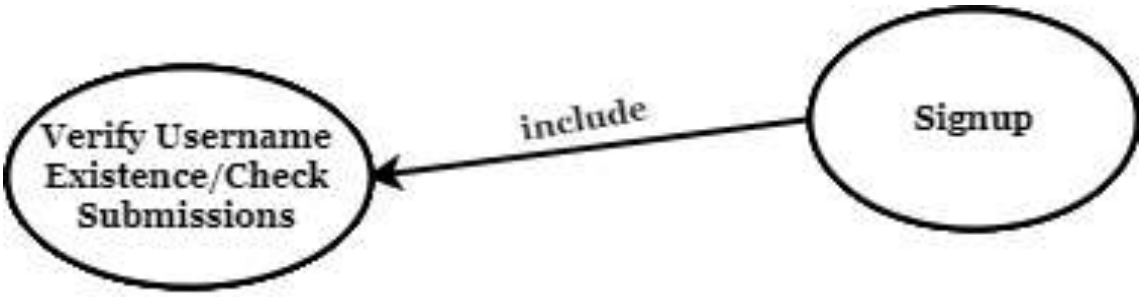


Figure 5.5.11 Verify Username Existence/Check submissions

Use Case Name:	Verify Username Existence/Check submissions	
ID:	HDP-013	
Actors Involved:	User	
Brief Description	System Verify Username Existence/Check submissions.	
Pre-Conditions	Signup[HDP-002]	
Post-Conditions	Login[HDP-001] Display signup error[HDP-012]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter All info required in signup form. 2. Clicks the submit button.	1. System collects signup form inputs. 2. Verify submissions if correct submits to database else display signup error.

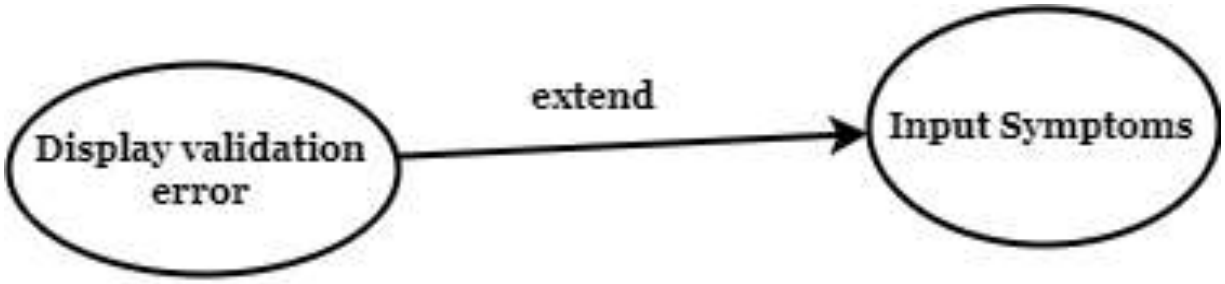


Figure 5.5.12 Display validation error

Use Case Name:	Display validation error	
ID:	HDP-014	
Actors Involved:	User	
Brief Description	Actor gets a validation error if there is any wrongly input symptoms.	
Pre-Conditions	Login[HDP-001] Symptoms range validation[HDP-015]	
Post-Conditions	Input Symptoms[HDP-006]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter Sysmptoms. 2. Clicks the submit button.	1. System collects inputs Sysmptoms. 2. System displays validation error on wrongly input symptoms.

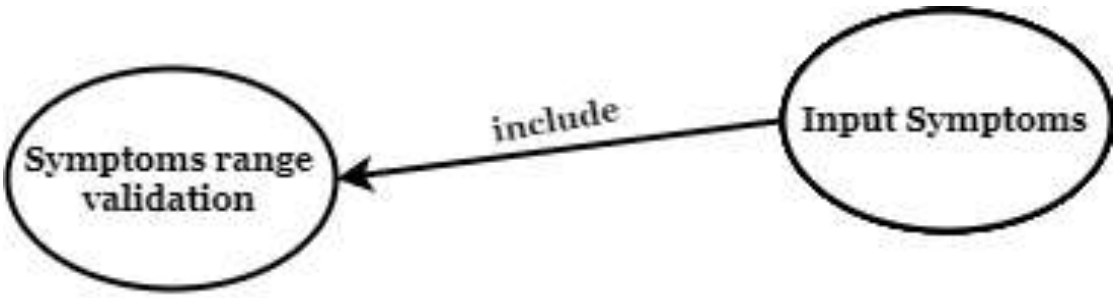


Figure 5.5.13 Symptoms range validation

Use Case Name:	Symptoms range validation	
ID:	HDP-015	
Actors Involved:	User	
Brief Description	System validate Symptoms entered by user	
Pre-Conditions	Login[HDP-001] Input Symptoms[HDP-006] Display validation error[HDP-014]	
Post-Conditions	View result[HDP-005]	
Priority:	high	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Enter Sysmptoms. 2. Clicks the submit button.	1. System collects inputs Sysmptoms. 2. System validates Symptoms.

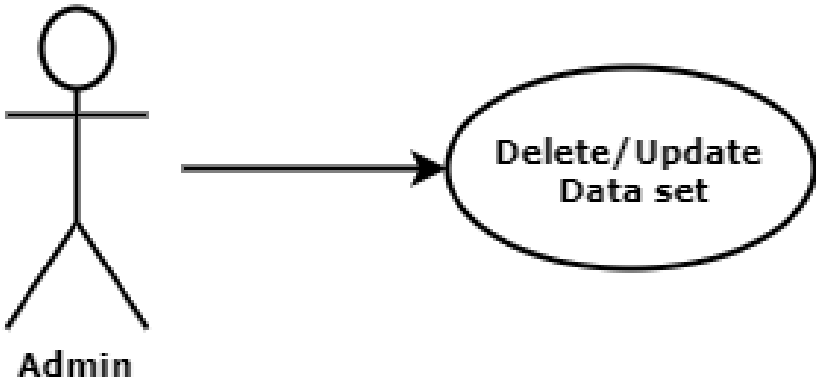


Figure 5.5.14 Delete/Update Data set

Use Case Name:	Delete/Update Data set	
ID:	HDP-016	
Actors Involved:	Admin	
Brief Description	Actor	
Pre-Conditions	Login[HDP-001] Import dataset[HDP-016]	
Post-Conditions	Train model[HDP-008]	
Priority:	medium	
Normal Flow of Events:	**Actor Action**	**System Response**
	1. Admin Delete/update dataset 2. Clicks the submit button.	1. System collects changes. 2. System submits to database.

4.6 ENTITY RELATIONSHIP DIAGRAM :

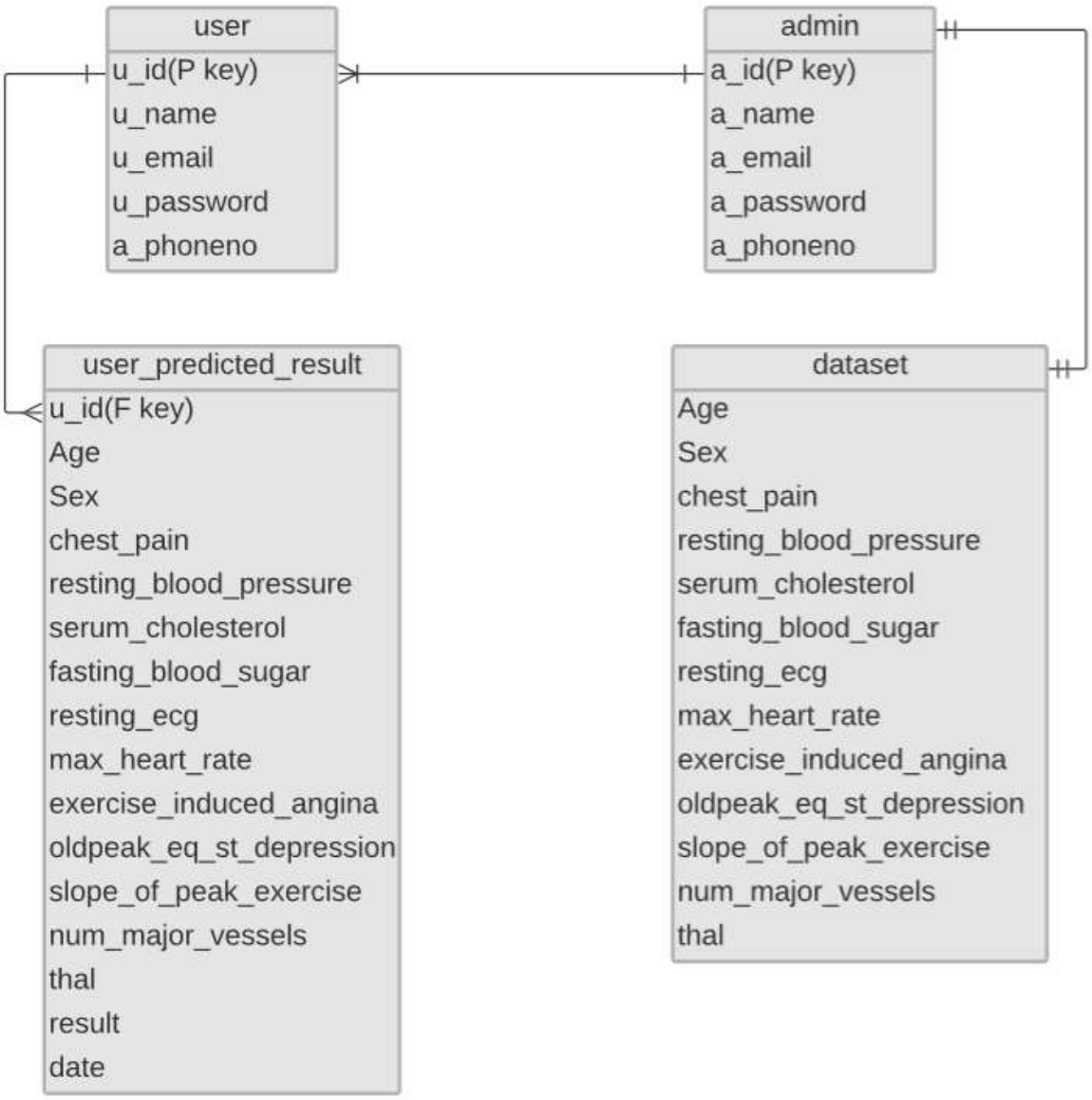


Figure 7 ERD Diagram

4.7 PROTOTYPE:

4.7.1 Login :

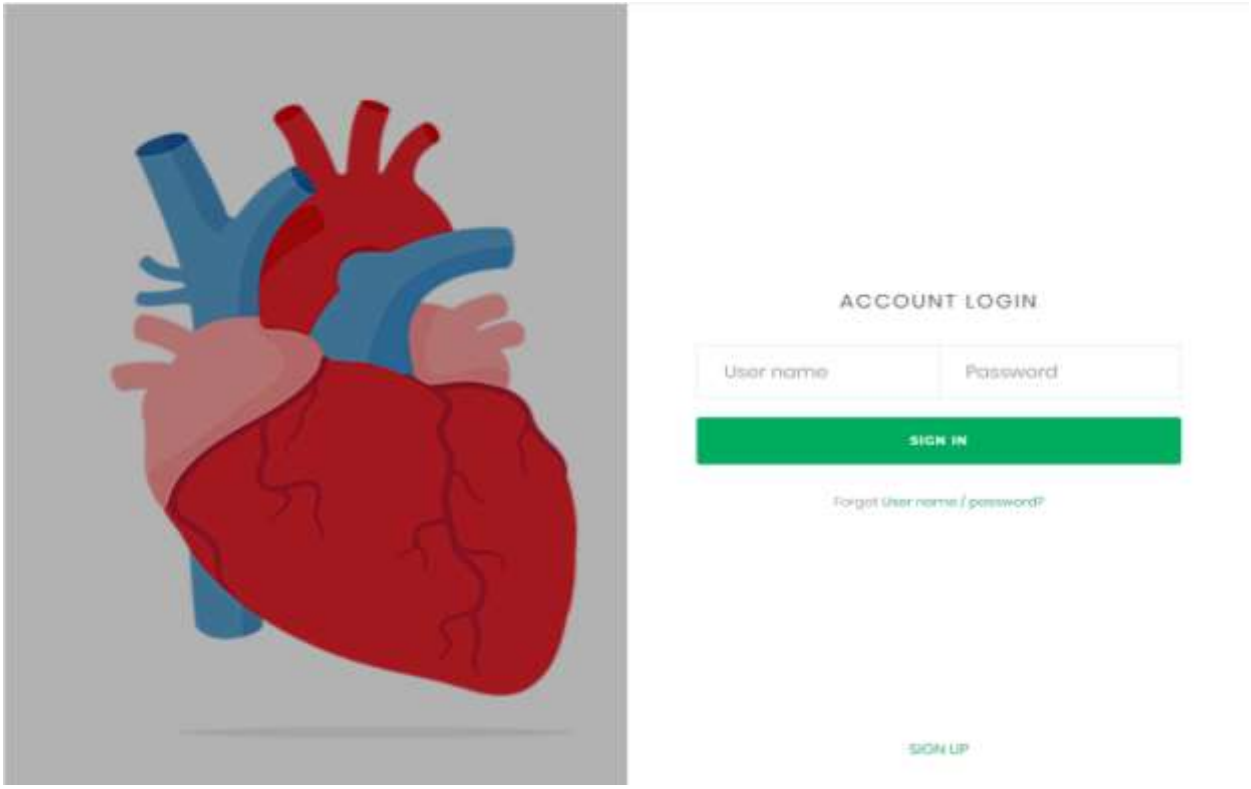


Figure 8 Prototype Login

4.7.2 Signup:

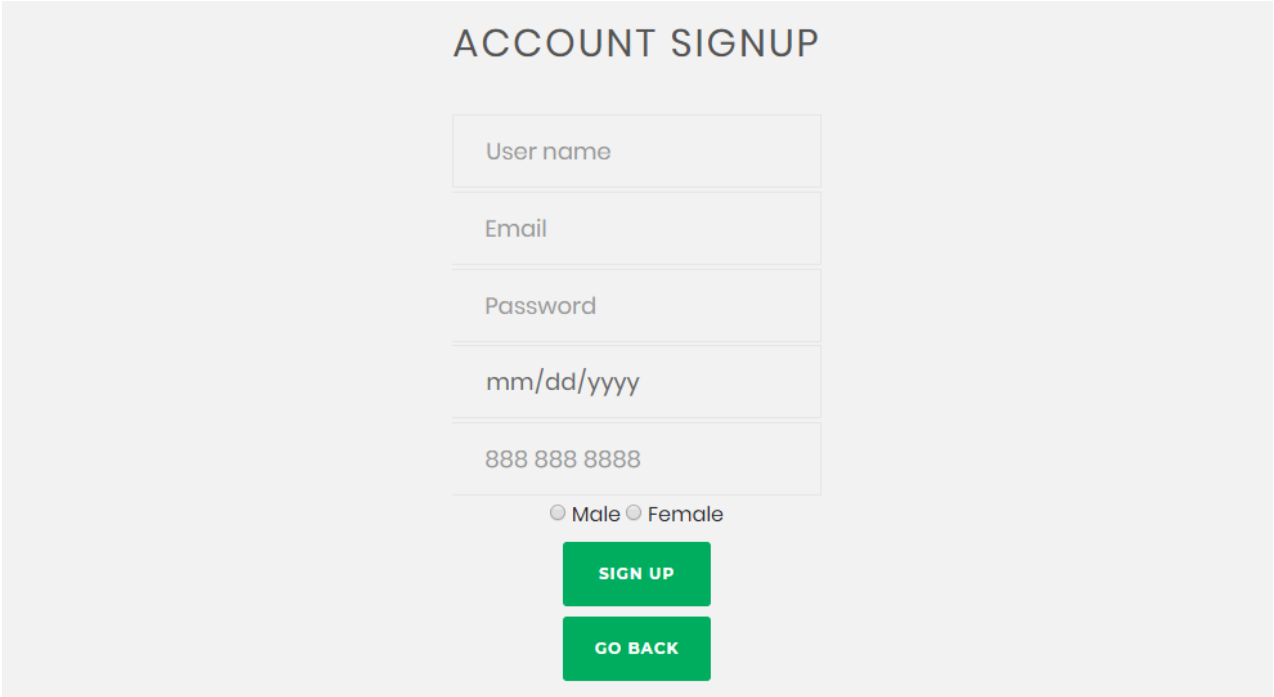


Figure 9 Prototype Signup

4.7.3 Password Recovery:

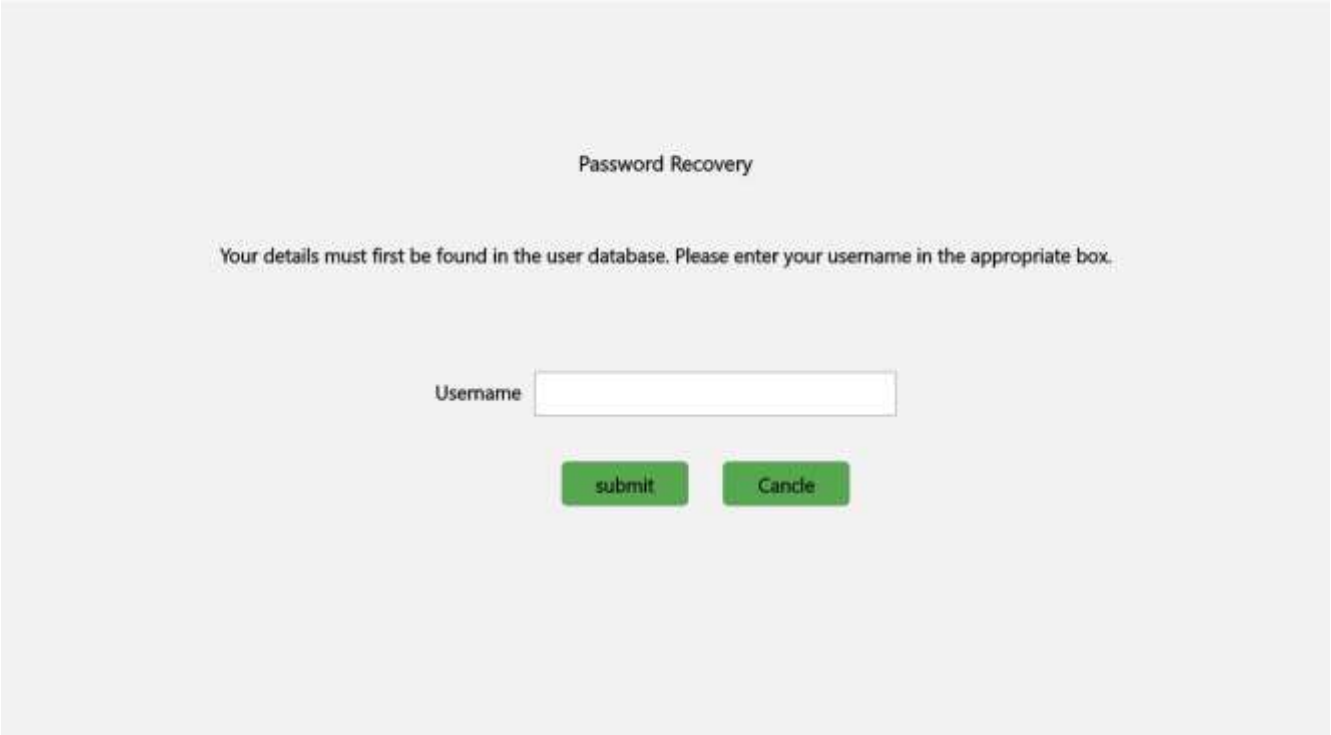


Figure 10 Prototype ForgetPassword

4.7.4 Admin Dashboard:

Figure 11 Prototype Adim Dashboard

4.7.6 User Dashboard:

Figure 12 Prototype User Dashboard

Chapter 5: PROJECT DESIGN & DEVELOPMENT

In this system we use multiple classification algorithm to implement successful heart attack prediction system. We may transfer the input to the device as in the CSV file or manual entry. On that input the algorithms apply after the input has been taken. The operation is performed after accessing data set and effective heart attack level is generated. By consulting expert doctors and medical experts, the proposed system would add several more criteria relevant to heart attack with their weight, age, and priority levels. The heart attack prediction system was developed to help classify various heart attack risk rates such as normal, low or high and also include the dosage information relevant to the expected outcome.

5.1 DESIGN :

An effective website design should accomplish its intended purpose by conveying its particular message while at the same time engaging the visitor. Various factors like clarity, colors, typography, graphics, usability and accessibility all contribute to good website design.

Images: To meet the user needs the website will be basic and will be very simple to use. Also, We will be adding elements of interactivity in the form of images slide in images etc. This will make the website easier for the users to use as they will be able to find information easier. We will also include images next to the written content that relates to the content.

Support: Clients often have questions or concerns about the websites. The website will have a support page. In the support page, the users can talk to us through the chat or can call us right away. If customers give us a call we will be more than happy to pick up the phone and deal with any enquiries they have whether it be sale enquiries, pricing problems or simple to ask for an advice on a costume. However, the website will also have an email so customers can email us with any questions. By doing this the website will increase customers satisfaction by providing better service.

Visibility on search engines: There are actually dozens of search engines but the most frequently used one is Google, Yahoo and Bing. I will make sure the website easily comes on the search engines. It is important for the users that they can find us on google quickly when they type heart disease. The website will be registering on Google my business which can help to appear the website in relevant search results. Also, the website will have content that matches the keywords on the website and have bolded keywords that match the products descriptions/theme on the page. By doing this customer won't have any problem to find us on.

Appropriateness of graphics: I will be using appropriate graphics like picture related to the costumes. I would not use any graphics that is not serving any purpose at all because some users may become annoyed at the unneeded use of graphics. I will not be using any inappropriate graphics as it won't attract the customers.

Mobile Friendly: More people are surfing the web using their phones or other apps. Construction of your website with a responsive layout, where your website can adjust to different screens, is essential.

5.1.1 WEB INTERFACE :

Project Front End Template

5.1.1.1 Login Page :

This Section contains the “Login” page. All users are required to access our website. The main operation on this page is login and its components are login credentials (Email/Username, Password.

This section also provides the facility for the user to redirect on register page, in case the user doesn’t have an account. It also allow user redirect to forget password page in case of loss of password.

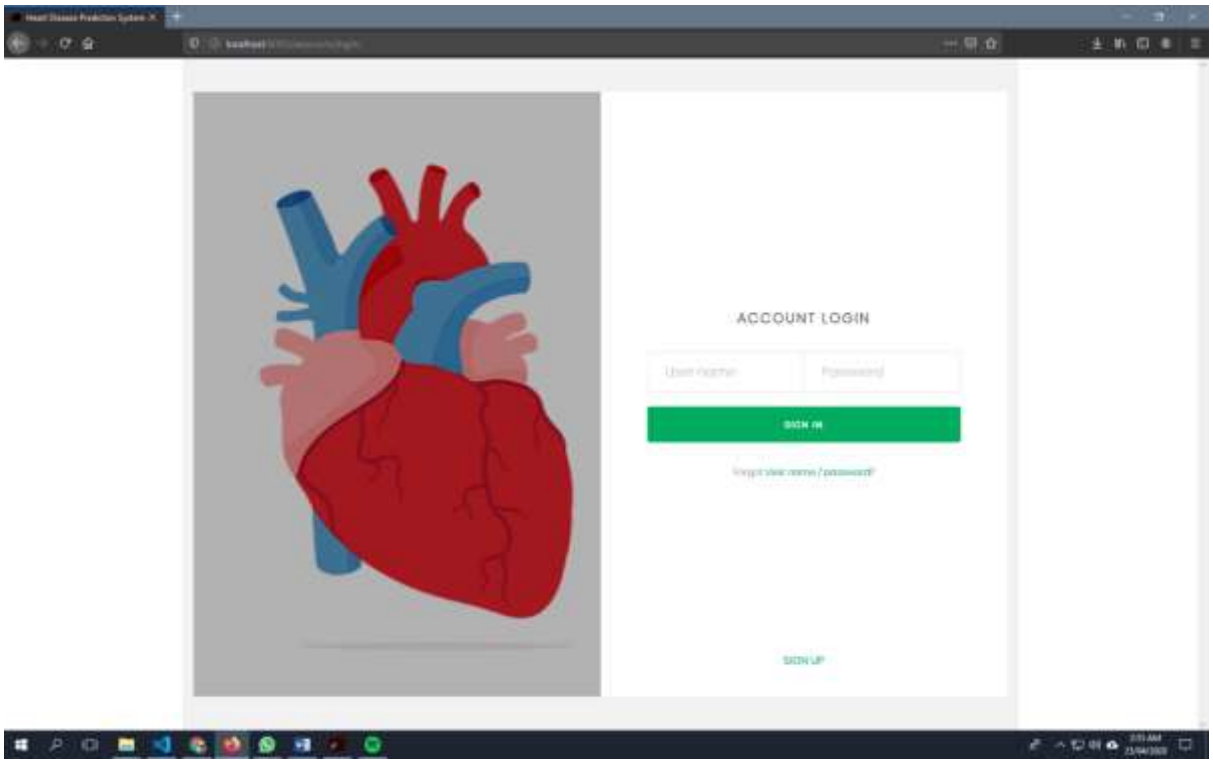
Purpose

This section holds the authenticity of the user and contains the following elements:

- Login Form
- Redirect to registration form
- Redirect to forget password page

Navigation & User Interaction

This user is required to insert the right email and password. After successful login, the user will have further options to see and explore.



5.1.1.2 Signup Page :

This Section provides the “Signup” facility to user to create a new account. The components required to perform this operation are Signup credentials (Names, Email, Date of Birth, etc).

The user must fill up the fields which include (Complete Name, Email, Password and Confirm Password, Gender, Date of birth, Country etc.)

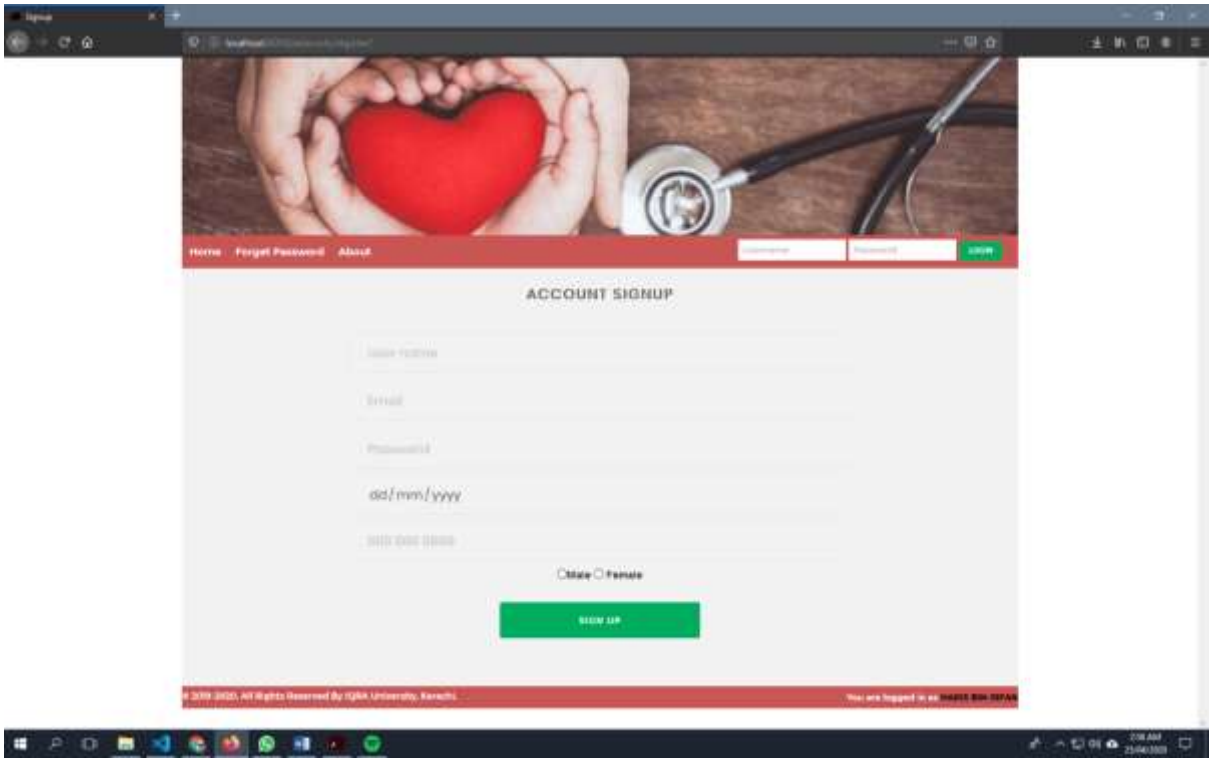
Purpose

This section allows the user to submit personal credential and get registered:

- Signup Form.
- Validation for not to put special characters in name and city fields.
- Submit button to submit a request for creating an account.

Navigation & User Interaction

The user is required to insert the correct name, email, password, and other credentials, After successful registration, the user will have logged in and can explore further.



5.1.1.3 **Profile Page**:

This Section provides the “Profile” facility to user to edit his/her personal data. As the user click on Edit profile the field will get enabled with all the information about the user stored in our database.

The user can edit fields which include (Complete Name, Email, Password and Confirm Password, Gender, Date of birth, Country etc.)

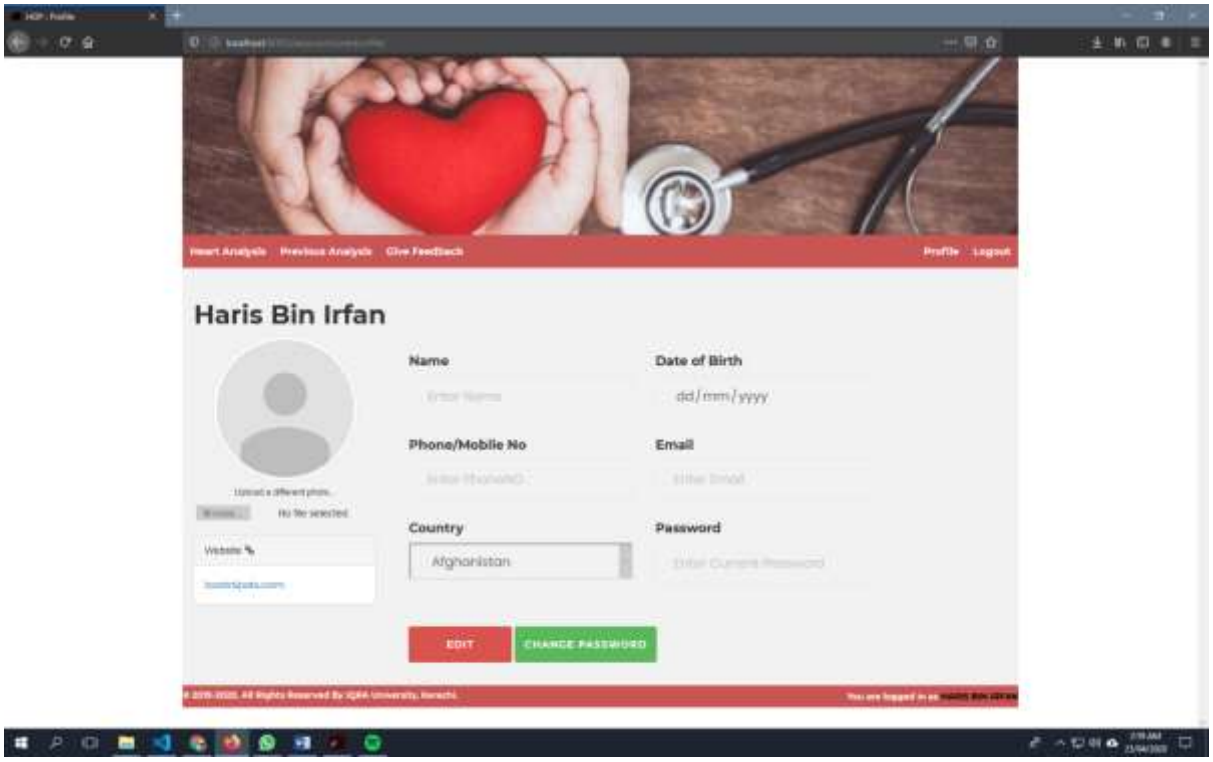
Purpose

This section allows the user to edit personal credential and update there accounts:

- Edit Form.
- Save button to submit a request for updating an account.
- Change Password Form.

Navigation & User Interaction

The user can navigate to this section first by clicking profile icon in the menu bar then by clicking the Edit button on profile page the field will get enabled with all the information about the user stored in our database. The user can edit this information so the system can keep updated information about the user.



5.1.1.4 Heart Analysis :

This Section provides the “Heart Analysis” facility to user to submit there symptoms to the system for heart disease prediction. As the user click on submit the field will sends all the symptoms information database and gets prediction result.

The user to submit there symptoms to the system using fields which include (Chest pain sex, age, Resting ECG and Thal, Slope, Cholesterol etc.)

Purpose

This section allows the user to submit there symptoms to the system:

- Heart Analysis Form.
- Submit button symptoms to the system for heart disease prediction.

Navigation & User Interaction

The user can navigate to this section first by clicking Heart Analysis icon in the menu bar the field submit user symptoms about there health to the system using fields which include (Chest pain sex, age, Resting ECG and Thal, Slope, Cholesterol etc) to the system for heart disease prediction.

5.1.1.5 **Password Recovery Page :**

This section contains the " Password Recovery " page. The main operation of this page is to restore the password for the user and its component is Email Address. The user only needs to write an email in the field and click the given button.

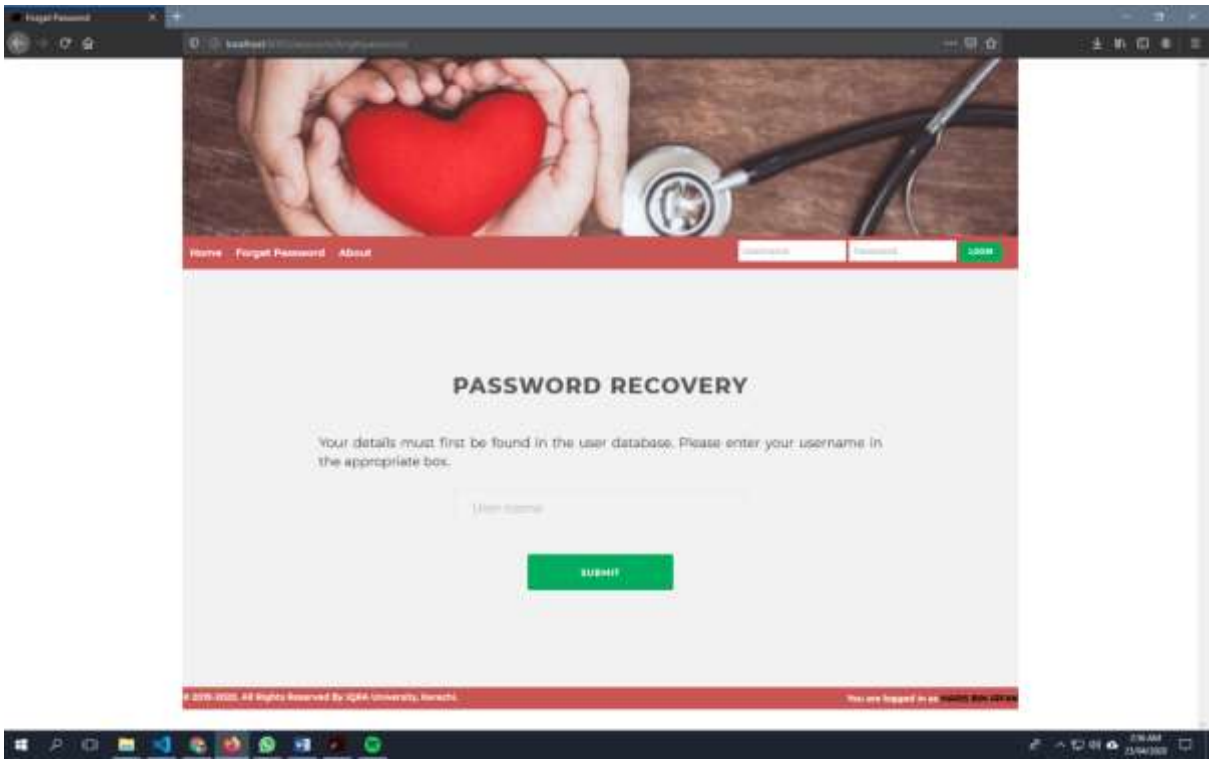
This section uses the server 's email service to send the current/changed password to user email address.

Purpose

This page allows the user to recover its password through email confirmation.

Navigation & User Interaction

The user is required to insert the registered email address and click the "Reset My Password" button. A forgot password email notification "Please reset your password by clicking here" is send to the user from the heartdisease.prediction@gmail.com which allows the user to recover current/changed password.



5.1.1.6 Change Password Page :

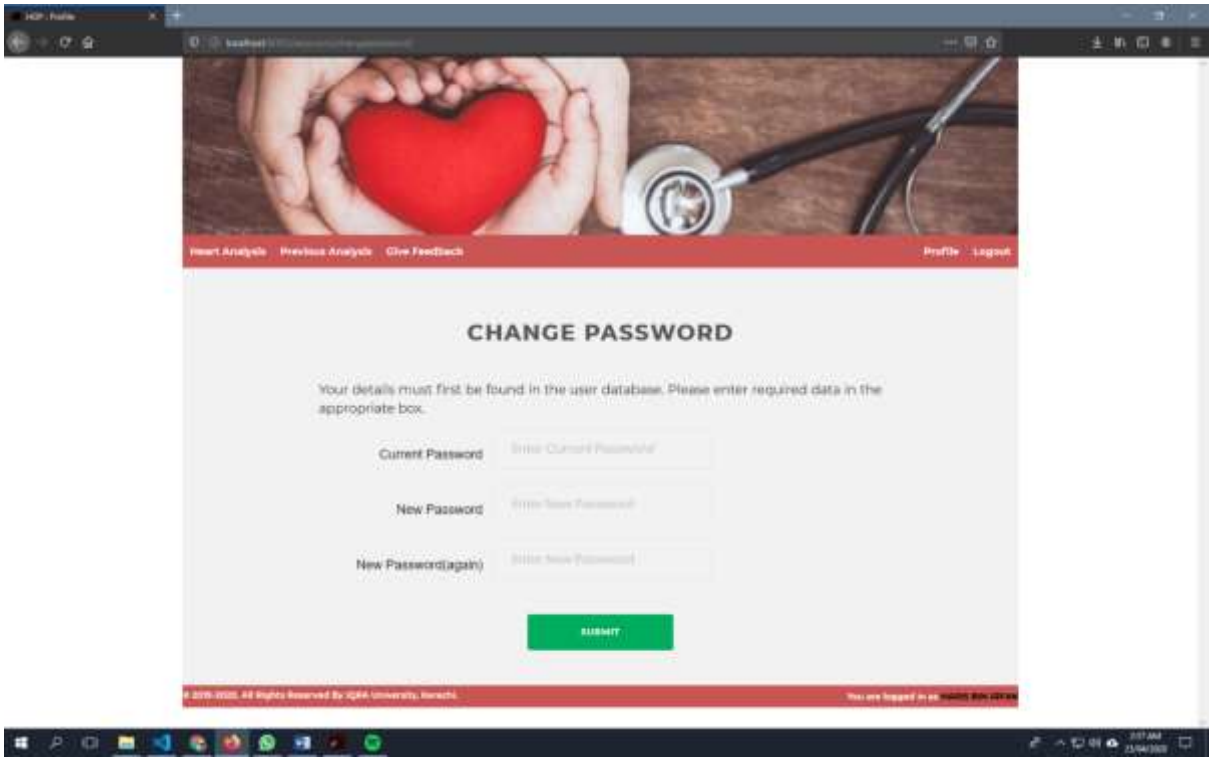
The user is allowed to reset his/her password from Rest your password page. This page is sent to the user on email when the user clicks on forgot the password button on profile page. This is due to the security purpose a user is not allowed to change the information of another user. The information of the user is taken from this page and against the User ID and Email store in the database. The user should enter current password, new password, confirm password. After entering all the information, the user must click on "Reset" button in order to set the new password.

Purpose

This page allows the user to set a new password if he/she does not remember it.

Navigation & User Interaction

This will navigate the user to a " profile" page of HDP. The current user is now Logged in to HDP a can get a recommendation.



5.1.1.7 **Analysis History Page :**

This Section provides the “Analysis History” facility to user to there previous analysis records. As the user click on Analysis History page a table with all the field and with all the information about the user stored in our database.

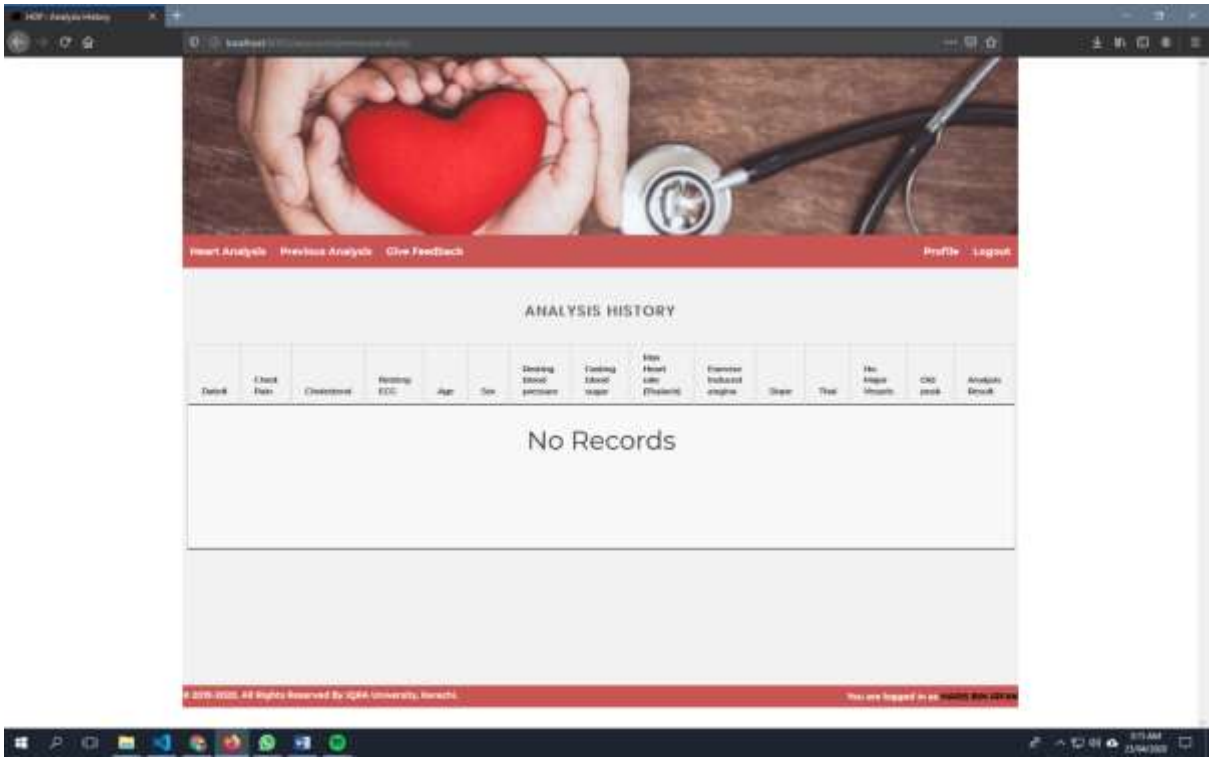
Purpose

This section allows the user to there previous analysis records:

- Analysis History page.

Navigation & User Interaction

The user can navigate to this section first by clicking Analysis History icon in the menu bar a table with all the field and with all the information about the user stored in our database.



5.1.1.8 **About Page** :

This section give a brief explanation of HDP to its users.

Purpose

The purpose of this section is to give a brief explanation of HDP to its users.

Navigation & User Interaction

The user can navigate to "About" page by clicking menu button "About" on the menu bar.'

5.1.2 BACKEND DESIGN :

Database Heart Disease Prediction Query's :-

```
CREATE DATABASE /*!32312 IF NOT EXISTS*/`heart disease prediction` /*!40100 DEFAULT CHARACTER SET utf8mb4 */;

USE `heart disease prediction`;

/*Table structure for table `dataset` */

DROP TABLE IF EXISTS `dataset`;

CREATE TABLE `dataset` (
  `date` varchar(50) DEFAULT NULL,
  `age` varchar(50) DEFAULT NULL,
  `sex` varchar(50) DEFAULT NULL,
  `chest_pain` varchar(50) DEFAULT NULL,
  `resting_blood_pressure` varchar(50) DEFAULT NULL,
  `serum_cholesterol` varchar(50) DEFAULT NULL,
  `fasting_blook_sugar` varchar(50) DEFAULT NULL,
  `resting_ecg` varchar(50) DEFAULT NULL,
  `max_heart_rate` varchar(50) DEFAULT NULL,
  `exercise_induced_angina` varchar(50) DEFAULT NULL,
  `oldpeak_eq_st_depression` varchar(50) DEFAULT NULL,
  `slope_of_peak_exercise` varchar(50) DEFAULT NULL,
  `num_major_vessels` varchar(50) DEFAULT NULL,
  `thal` varchar(50) DEFAULT NULL,
  `result` varchar(50) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;

/*Data for the table `dataset` */

/*Table structure for table `user` */

DROP TABLE IF EXISTS `user`;

CREATE TABLE `user` (
  `id` int(11) NOT NULL,
```

```
  `name` varchar(50) DEFAULT NULL,
  `username` varchar(50) DEFAULT NULL,
  `email` varchar(50) DEFAULT NULL,
  `password` varchar(50) DEFAULT NULL,
  `phone_no` varchar(50) DEFAULT NULL,
  `address` varchar(100) DEFAULT NULL,
  `type` varchar(50) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;

/*Data for the table `user` */

insert  into `user`(`id`,`name`,`username`,`email`,`password`,`phone_no`,`address`,`type`) values (0,NULL,NULL,NULL,NULL,NULL,NULL,NULL);

/*Table structure for table `user_predicted_result` */

DROP TABLE IF EXISTS `user_predicted_result`;

CREATE TABLE `user_predicted_result` (
  `u_id` int(11) DEFAULT NULL,
  `date` varchar(50) DEFAULT NULL,
  `age` varchar(50) DEFAULT NULL,
  `sex` varchar(50) DEFAULT NULL,
  `chest_pain` varchar(50) DEFAULT NULL,
  `resting_blood_pressure` varchar(50) DEFAULT NULL,
  `serum_cholesterol` varchar(50) DEFAULT NULL,
  `fasting_blook_sugar` varchar(50) DEFAULT NULL,
  `resting_ecg` varchar(50) DEFAULT NULL,
  `max_heart_rate` varchar(50) DEFAULT NULL,
  `exercise_induced_angina` varchar(50) DEFAULT NULL,
  `oldpeak_eq_st_depression` varchar(50) DEFAULT NULL,
  `slope_of_peak_exercise` varchar(50) DEFAULT NULL,
  `num_major_vessels` varchar(50) DEFAULT NULL,
```

```
  `thal` varchar(50) DEFAULT NULL,

  `result` varchar(50) DEFAULT NULL,

  KEY `u_id` (`u_id`),

  CONSTRAINT `user_predicted_result_ibfk_1` FOREIGN KEY (`u_id`) REFERENCES `user`
(`id`)

) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

5.1 DEVELOPMENT :

The work of software development isn't confined to coders or development teams. Professionals such as scientists, device fabricators and hardware makers also create software code even though they are not primarily software developers. Nor is it confined to traditional information technology industries such as software or semiconductor businesses. In fact, according to the Brookings Institute, those businesses "account for less than half of the companies performing software development."

5.1.1 DJANGO FRAMEWORK :

With Django, you can take Web applications from concept to launch in a matter of hours. Django takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source. Django includes dozens of extras you can use to handle common Web development tasks. Django takes care of user authentication, content administration, site maps, RSS feeds, and many more tasks right out of the box.

- **Ridiculously fast** : Django was designed to help developers take applications from concept to completion as quickly as possible.
- **Reassuringly secure** : Django takes security seriously and helps developers avoid many common security mistakes, such as SQL injection, cross-site scripting, cross-site request forgery and clickjacking. Its user authentication system provides a secure way to manage user accounts and passwords.
- **Exceedingly scalable** : Some of the busiest sites on the planet use Django's ability to quickly and flexibly scale to meet the heaviest traffic demands.
- **Incredibly versatile** : Companies, organizations and governments have used Django to build all sorts of things — from content management systems to social networks to scientific computing platforms.

5.1.2 LIBRARIES USED :

- **Psycopg: -** Psycopg is the most popular PostgreSQL database adapter for the Python programming language. Its main features are the complete implementation of the Python DB API 2.0 specification and the thread safety (several threads can share the same connection). It was designed for heavily multi-threaded applications that create and destroy lots of cursors and make a large number of concurrent "INSERT"s or "UPDATE"s.
- **Pillow: -** Python Imaging Library (Fork). Pillow is the friendly PIL fork by Alex Clark and Contributors. PIL is the Python Imaging Library by Fredrik Lundh and Contributors.
- **matplotlib.pyplot**:- is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- **Numpy**: - is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- **Pandas: -** In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.
- **Scikit-learn :-** (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- **Joblib :-** library is intended to be a replacement for Pickle, for objects containing large data. We'll repeat the save and restore procedure as with Pickle. As seen from the example, the Joblib library offers a bit simpler workflow compared to Pickle. While Pickle requires a file object to be passed as an argument, Joblib works with both file objects and string filenames. In case your model contains large arrays of data, each array will be stored in a separate file, but the save and restore procedure will remain the same. Joblib also allows different compression methods, such as 'zlib', 'gzip', 'bz2', and different levels of compression.

Chapter 6: TESTING AND EVALUATION

- Testing is a program execution process which aims to find an error.
- Research poses a surprising electronic engineering phenomenon.
- The software review helps to reassure device developers and consumers that the software is reasonably suitable for operational use. Testing is a process which aims to create trust in the software.
- Training is a collection of tasks which can be regularly organized and carried out in advance.
- Application testing is also referred to as Test & Validate.

6.1 TYPE OF TESTING :

The various types of testing include

- White Box Testing
- Black Box Testing
- Alpha Testing
- Beta Testing
- Win Runner And Load Runner
- Unit Testing
- System Testing

6.1.1 WHITE BOX TESTING :

- Also known as glass-box inspection. This is a form of test case design, using the formal process control system to extract test cases.
- The Software Developer will derive test cases using white box test methods Using white box testing methods, the software engineer can derive test cases that
  1. Ensure that all individual parts within a module are exercised at least once.
  2. Practice all rational decisions on their real and false hands.

6.1.2 BLACK BOX TESTING :

- Also called as a behavior check. It focuses upon the software's practical requirements.
- A complementary method is likely to show a different class of errors than errors in the White Box.
- A black box check helps the software engineering to determine a set of input conditions that will completely exercise all usable system specifications.

6.1.2 ALPHA TESTING :

Alpha testing is the prototype stage of the program, where the program will run first. Not all of the intended features will be usable, but it will have core functions and will be able to accept inputs and produce outputs. An alpha check is normally performed on a different device in the developer's offices.

6.1.2 BETA TESTING :

The beta test is a "live" software program in an environment that the developer can not monitor. The beta check is performed by the end user of the app at one or more client sites.

6.1.2 WIN RUNNER TESTING :

We use Win Runner as a load monitoring tool running on the GUI layer, since it helps us to record and replay user activities from a wide range of software applications as if those activities were manually performed by a real person.

6.1.2 LOAD RUNNER TESTING :

You can get an accurate picture of end-to-end efficiency of the system with Load Runner. Verify the new or modified applications meet the performance criteria set out.

6.2 TESTING USED IN THIS PROJECT :

6.1.2 SYSTEM TESTING :

Testing the debugging programs is one of the most important aspects of the computer programming triggers, the system will never generate the output for which it was designed without the programs that function. Testing is best performed when product development is called upon to help find any mistakes and bugs. The sample data were used to check. It is not quantity but the consistency of the data that used the research matters. Testing is intended to ensure that the device before live process commands was correctly an effective one.

6.1.2 VALIDATION TESTING :

The program is fully installed as a kit, interfacing errors have been discovered and corrected and a final set of software tests have been carried out. That is, validation tests start, validation testing can be specified in several ways, but a clear meaning is that validation works when the program performs in a manner that can reasonably be assumed to be the client. Here all the validations given to the design is completely checked. All validations given to the design are reviewed in full here.

Chapter 7: CONCLUSION AND FUTURE WORK

The proposed model is a web application system based on GUI, user-friendly, scalable, responsive and expandable. The proposed working model will also help to reduce the cost of care by offering prompt initial diagnostics. The model will also serve the function of the medical student training program, and will be a soft diagnostic method available to doctors and cardiologists. General practitioners can use this method for initial cardio-patient diagnosis. There are several potential enhancements which could be explored to enhance this prediction system's scalability and accuracy. Since we've built a simplified method, we can use this method to analyze various data sets in future. The dimensionality of the heart database is typically high in DM warehouse so defining and selecting significant attributes for better diagnosis of heart disease are very daunting tasks for future study.

REFERENCES

[1] C. S. D. a. S. S. Apte, "“Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”," *International Journal of Computer Applications,* vol. Vol. 47, no. pp. 0975 – 888, p. No. 10, 2012.

[2] L. HG, N. KY and R. KH, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV. In: Washio T, Zhou ZH, Huang JZ, et al, editors. Emerging Technologies in Knowledge Discovery and Data Mining. PAKDD 2007, Lecture Notes in Computer Scien," 2007. [Online].

[3] C. Blake and . C. Mertz, "UCI Machine LearningDatabases," 2004. [Online]. Available: http://mlearn.ics.uci.edu/databases/heartdisease/.

[4] K. Polaraju, D. Durga Prasad and . M. Tech Scholar, "Prediction of Heart Disease using Multiple Linear Regression Model.," 2017. [Online]. Available: International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from www.ijedr.org.

[5] P. P. &. R. C. Sai, "HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.," 2017. [Online]. Available: International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from www.ijcsmc.com.

[6] M. H. A. &. U. Sultana, "Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).," 2017. [Online]. Available: https://doi.org/10.1109/CEEICT.2016.7873142.

[7] A. S. &. N. C. A, "Different Data Mining Approaches for Predicting Heart Disease, 277– 281.," 2016. [Online]. Available: https://doi.org/10.15680/IJIRSET.2016.0505545.

[8] J. A. U. &. S. D. Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease," 2011. [Online]. Available: Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5). https://doi.org/10.1109/CEEICT.2016.7873142.

[9] Z. J. Yan H, "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm.," 2008. [Online]. Available: J Appl Soft Comput..

[10] U. M. L. Repository., Accessed November 01, 2018. [Online]. Available: Available from: https://archive.ics.uci.edu/ml/index.php. .