CARDIOVASCULAR DISEASES PREDICTION AND PULSE CHECKER

Main Project Report

Submitted by

AISHA FARHEEN RASHEED A

Reg No : FIT20MCA-2004

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Computer Applications
Of
A P J Abdul Kalam Technological University

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®
ANGAMALY-683577, ERNAKULAM(DIST)
JULY 2022

DECLARATION

I, **AISHA FARHEEN RASHEED A** hereby declare that the report of this project work, submitted to the Department of Computer Applications, Federal Institute of Science and Technology (**FISAT**), Angamaly in partial fulfillment of the award of the degree of Master of Computer Application is an authentic record of my original work.

The report has not been submitted for the award of any degree of this university or any other university.

Date : 21-07-2022

Place: Angamaly **Name : Aisha Farheen Rasheed A**

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®
ANGAMALY, ERNAKULAM-683577

DEPARTMENT OF COMPUTER APPLICATIONS

CERTIFICATE

This is to certify that the project report titled **'Cardiovascular Diseases Prediction And Pulse Checker'** submitted by **Aisha Farheen Rasheed A [Reg No: FIT20MCA-2004]** towards partial fulfillment of the requirements for the award of the degree of Master of Computer Applications is a record of bonafide work carried out by her during the year 2022.

Project Guide **Head of the Department**

Submitted for the viva-voice held on . . . . . at . . . . . .

Examiner:

ACKNOWLEDGEMENT

Gratitude is a feeling which is more eloquent than words, more silent than silence. To complete this project work I needed the direction, assistance and co-operation of various individuals, which is received in abundance with the grace of God.

I hereby express my deep sense of gratitude to **Dr. Manoj George** Principal, FISAT and **Dr. C Sheela** Vice principal, FISAT for allowing me to utilize all the facilities of the college.

My sincere thanks to **Dr. Deepa Mary Mathews** HOD, Department of Computer Applications FISAT ,who had been a source of inspiration. I express heartiest thanks to **Ms. Manju Joy** my internal guide for his encouragement and valuable suggestion. I express my heartiest gratitude to my scrum master **Dr. Sujesh P Lal** and the faculty members in my Department for encouragement and constructive suggestions and comment during the project work.

Finally I wish to express my thanks to my parents, friends and well-wishers who extended their help in one way or other in preparation of my project. Besides all, I thank GOD for everything.

ABSTRACT

Heart-related diseases or Cardiovascular Diseases (CVDs) 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 fora reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart-related diseases. Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. A huge amount of patient-related data is maintained on monthly basis. The stored data can be useful for the source of predicting the occurrence of future diseases. To reduce the large scale of deaths from heart diseases, a quick and efficient detection technique is to be discovered. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop software with the help of machine learning algorithms which can help doctors to decide both prediction and diagnosing of heart disease. The main objective of this research project is to predict the heart disease of a patient using machine learning algorithms.

Contents

Chapter 1

INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them

under suitable headings finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Machine Learning, given by the department of computer science and engineering. Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortal ityrate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience,time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, logistic regression and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications,which can be a great milestone in the field of medicine.

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience,time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data. Our motto in developing this app is to provide a safe environment to women through smart phone as today most of the people are carrying smart phones to wherever they go. Of course, the Delhi Nirbhaya case has made the Government to make the laws tougher, but even though the sexual crime rate in

India have not decreased. So, it is better to take our own safety measures rather than becoming a victim of those crimes.

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models, and further implement Logistic Regression model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop an appropriate computer-based system and decision support that can aid to early detection of heart disease, in this project we have developed a model which classifies if patient will have heart disease in ten years or not based on various features (i.e. potential risk factors that can cause heart disease) using logistic regression. Hence, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Chapter 2

PROOF OF CONCEPT

The development of project are to develop machine learning model to predict future possibility of heart disease by implementing Random Forest. To determine significant risk factors based on medical dataset which may lead to heart disease.To analyze feature selection methods and understand their working principle.

The application is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the three algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better.

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers. The paper propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factor SMO (89 of accuracy) and Bayes Net (87 of accuracy) achieve optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms are still not satisfactory. So that if the performance of accuracy is improved more to give batter decision to diagnosis disease.

In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Na¨ıve Bayes and Random Forest gave the maximum accuracy of 91.2 percent .a system containing two models based on linear Support Vector Machine (SVM). The first one is called L1 regularized and the second one is called L2 regularized. First model is used for removing unnecessary features by making coefficient of those features zero. The second model is used for prediction. Predication of disease is done in this part. To optimize both models they proposed a hybrid grid search algorithm.

This algorithm optimizes two models based on metrics: accuracy, sensitivity, septicity, the Matthews correlation coefficient, ROC chart and area under the curve. They used Cleveland data set. Data splits into 70 percent training and 30percent testing used holdout validation. There are two experiments carried out and each experiment is carried out for various values of C1, C2 and k where C1 is hyper parameter of L1 regularized model, C2 is hyper parameter of L2 regularized model and k is the size of selected subset of features. First experiment is L1-linear SVM model stacked with L2-linear SVM model which is giving maximum testing accuracy of 91.11percent and training accuracy of 84.05percent . The second experiment is L1-linear SVM model cascaded with L2-linear SVM model with RBF kernel. This is giving maximum testing accuracy of 92.22 percent and

training accuracy of 85.02.They have obtained an improvement in accuracy over conventional SVM models by 3.3percent .

The data set used is Cleveland data set. The first step is data pre-processing step. In this the tuples are removed from the data set which has missed the values. Attributes age and sex from data set are also not used as the authors think that it's personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is the combination of Random Forest (RF) and Linear method (LM). In the HRFLM algorithm, the authors have used four algorithms. First algorithm deals with partitioning the input dataset. It is based on a decision tree which is executed for each sample of the dataset.

After identifying the feature space, the dataset is split into the leaf nodes. Output of first algorithm is Partition of data set. After that in second algorithm they apply rules to the data set and output here is the classification of data with those rules. In third algorithm features are extracted using Less Error Classifier. This algorithm deals with finding the minimum and maximum error rate from the classifier. Output of this algorithm is the features with classified attributes. In forth algorithm they apply Classifier which is hybrid method based on the error rate on the Extracted Features. Finally they have compared the results obtained after applying HRFLM with other classification algorithms such a decision tree and support vector machine. In result as RF and LM are giving better results than other, both the algorithms are put together and new unique algorithm HRFLM is created. The authors suggest further improvement in accuracy by using combination of various machine learning algorithms.

To develop a mobile application called women safety which ensures the safety of women. The app includes the following features

- User can generate report of heart diseases prediction along with heart rate.
- User can view details of doctor and chat with them.
- User can purchase heart related products.
- User can see their orders.

Chapter 3

REQUIREMENT ANALYSIS AND SPECIFICATION

3.1 System Study

The implementation phase of the Cardiovascular Diseases Prediction and Pulse Checkerapplication is to ensure safety of the patient health. Thus it translates design specification into source code. The user(Admin as per Cardiovascular Diseases Prediction) tests the developed system and changes are made according to their needs. In this system, it has been successfully implemented. Before implementation, several tests have been conducted to ensure that no errors are encountered during the operation. This project's implementation is done as there are 2 users i.e., Admin,Doctor and Customer. This includes all those activities that take place to convert from old system to new system. The system can be implemented only after testingis done and is found to be working to specifications. The implementation stage isa systems project in its own right.

3.1.1 Existing system

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools techniques are regularly being experimented with to suit the present-day health needs.

Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings finally analyzing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says “Prevention is better than cure”, early prediction its control can be helpful to prevent decrease the death rates due to heart disease.

3.1.2 Proposed system

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

- we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model.

- Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction.

- , data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset.

- Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

3.1.3 Feasibility study

Feasibility is defined as the practical extent to which a project can be performed successfully. To evaluate feasibility, a feasibility study is performed, which determines whether the solution considered to accomplish the requirements is practical and workable in the software. The objective of the feasibility study is to establish the reasons for developing the software that is acceptable to users, adaptable to change and compatible to established standards, to analyze whether the software will meet organizational requirement.

There are three aspects in the feasibility study portion of preliminary investigation

i Technical feasibility

ii Economic feasibility

iii Operational feasibility

Technical feasibility

Cardiovascular Diseases Prediction and Pulse Checker application must be first evaluated from its technical view point. The system is said to be technically feasible,since there will not be much difficulty in getting required resources for the development and maintaining the system as well. All the resources needed for the development of the software as well as the maintenance of the same is available in the organization, here we are utilizing the resources which are available already.

Economic feasibility

Economic analysis is the most frequently used method for evaluating effectiveness of the proposed system most commonly known as cost-benefit analysis. This procedure determines the benefits and savings that are expected from the proposed system and compared with the cost of the existing system. As this system works as a computer based system, reduces a lot manual effort and thus manpower cost. It also introduces faith and goodwill and can be measured as an intangible benefit. As we are generated from the computer based system it reduces cost and time and naturally error prone as compare to manual typewriter.

Operational feasibility

Operational feasibility is the ability to utilize, support and perform the necessary tasks of a system or program. It includes everyone who creates, operates or uses the system. In operational feasibility the entire application is checked whether the system will be used if it is developed and implemented. Also it is checked whether there will be resistance from user that may undermine the possible application benefits. There is no barrier for implementing the system. The system also helps to access the information immediately as need arises. Thus the system is found to be operationally feasible.

3.2 User Characteristics

There are two users characteristics in this system:

- Administrator
- Doctor
- Customer

3.2.1 Administrator

- Administrator is the super user and main controller of this application.
- He/She controls all the activities in Cardiovascular Diseases Prediction and Pulse Checker application. He/She has full control over what happen in thesystem.
- Admin can login and can maintain the application working.

3.2.2 Customer

- Customer can register and login
- Can check their heart diseases prediction.
- Can purchase the medical products available in the site.
- Can check heart rate anywhere and anytime.
- Can chat with the doctors available.
- Can see all the details of orders.

3.2.3 Doctor

- doctor can register and login
- Can check chat with various customer.

3.3 System Specification

The project will create by using Anaconda. Django will be used as backend. Sqlite database will be used to store data.

3.3.1 Hardware specification

Processor	i5 or i7 (i7 is better)
RAM	12GB (Minimum)
Hard Disk	500GB or above
Other	Mouse and Keyboard
Hardware Part	Pulse sensor,MQTT Implementation
Embedded Devices	Microcontroller-NodeMCU, Pulse sensor

3.3.2 Software Specification

Tool	Python IDLE, Arduino IDE
Language	Python, C
Operating System	Windows 7 or later
Front End	Python,CSS,HTML
Back end	Django

3.3.3 About the software tools and platforms

Python IDLE

IDLE (Integrated Development and Learning Environment) is an integrated development environment (IDE) for Python. The Python installer for Windows contains the IDLE module by default.IDLE is not available by default in Python distributions for Linux. It needs to be installed using the respective package managers.IDLE can be used to execute a single statement just like Python Shell and also to create, modify, and execute Python scripts. IDLE provides a fully-featured text editor to create Python script that includes features like syntax highlighting, autocompletion, and smart indent. It also has a debugger with stepping and breakpoints features.

Python installation comes with an Integrated Development and Learning Environment, which you'll see shortened to IDLE or even IDE. These are a class of applications that help you write code more efficiently. While there are many IDEs for you to choose from, Python IDLE is very bare-bones, which makes it the perfect tool for a beginning programmer.Python installation comes with an Integrated Development and Learning Environment, which you'll see shortened to IDLE or even IDE. These are a class of applications that help you write code more efficiently. While there are many IDEs for you to choose from, Python IDLE is very bare-bones, which makes it the perfect tool for a beginning programmer.Python IDLE comes included in Python installations on Windows and Mac. If you're a Linux user, then you should be able to find and download Python IDLE using your package manager. Once you've installed it, you can then use Python IDLE as an interactive interpreter or as a file editor.

Arduino IDE

The Arduino Integrated Development Environment - or Arduino Software (IDE) - contains a text editor for writing code, a message area, a text console, a toolbar with buttons for common functions and a series of menus. It connects to the Arduino hardware to upload programs and communicate with them.Programs written using Arduino Software (IDE) are called sketches. These sketches are written in the text editor and are saved with the file extension .ino. The editor has features for cutting/pasting and for searching/replacing text.

The message area gives feedback while saving and exporting and also displays errors. The console displays text output by the Arduino Software (IDE), including complete error messages and other information. The bottom righthand corner of the window displays the configured board and serial port. The toolbar buttons allow you to verify and upload programs, create, open, and save sketches, and open the serial monitor.

The Arduino Software (IDE) uses the concept of a sketchbook: a standard place to store your programs (or sketches). The sketches in your sketchbook can be opened from the File ¿ Sketchbook menu or from the Open button on the toolbar. The first time you run the Arduino software, it will automatically create a directory for your sketchbook. You can view or change the location of the sketchbook location from with the Preferences dialog.

Libraries provide extra functionality for use in sketches, e.g. working with hardware or manipulating data. To use a library in a sketch, select it from the Sketch ¿ Import Library menu. This will insert one or more include statements at the top of the sketch and compile the library with your sketch. Because libraries are uploaded to the board with your sketch, they increase the amount of space it takes up. If a sketch no longer needs a library, simply delete its include statements from the top of your code.

SQLite

SQLite is an in-process library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine. It is a database, which is zero-configured, which means like other databases you do not need to configure it in your system. SQLite engine is not a standalone process like other databases, you can link it statically or dynamically as per your requirement with your application. SQLite accesses its storage files directly.

Why SQLite?

- SQLite does not require a separate server process or system to operate (serverless).
- SQLite comes with zero-configuration, which means no setup or administration needed.
- A complete SQLite database is stored in a single cross-platform disk file.
- SQLite is very small and light weight, less than 400KiB fully configured or less than 250KiB with optional features omitted.
- SQLite is self-contained, which means no external dependencies.
- SQLite transactions are fully ACID-compliant, allowing safe access from multiple processes or threads.
- SQLite supports most of the query language features found in SQL92 (SQL2) standard.
- SQLite is written in ANSI-C and provides simple and easy-to-use API.
- SQLite is available on UNIX (Linux, Mac OS-X, Android, iOS) and Windows (Win32, WinCE, WinRT).

DJANGO

Django is a Python framework that makes it easier to create web sites using Python.Django takes care of the difficult stuff so that you can concentrate on building your web applications. Django emphasizes reusability of components, also refereed to as DRY (Don't Repeat Yourself), and comes with ready-to-use features like login system, database connection and CRUD operations (Create Read Update Delete).

Django follows the MVT design pattern (Model View Template). Model - The data you want to present, usually data from a database. View - A request handler that returns the relevant template and content - based on the request from the user. Template - A text file (like an HTML file) containing the layout of the web page, with logic on how to display the data.

The model provides data from the database. In Django, the data is delivered as an Object Relational Mapping (ORM), which is a technique designed to make it easier to work with databases. The most common way to extract data from a database is SQL. One problem with SQL is that you have to have a pretty good understanding of the database structure to be able to work with it Django, with ORM, makes it easier to communicate with the database, without having to write complex SQL statements. The models are usually located in a file called models.py View

A view is a function or method that takes http requests as arguments, imports the relevant model(s), and finds out what data to send to the template, and returns the final result. The views are usually located in a file called views.py.

Chapter 4

SYSTEM MODELING

4.1 Modules and Description

4.1.1 Registration

User can add the contact details(name , phno ,gender password and email address).Doctor can add the contact details(name , phno ,gender password and email address).

4.1.2 Heart disease prediction

a) Training using Random forest classifier

- Load heart disease dataset
- Remove null values from rows and columns
- Splitting data's in to training and testing set
- Train Random forest classifier using preprocessed data
- Saving Trained model
- Calculate accuracy

b)Testing

- Load trained RFC model
- Input data's
- Prediction using loaded model
- Train Random forest classifier using pre processed data Show result

4.1.3 Heart rate

- show the count of heart rate in number bpm.

4.1.4 Products

- add and edit medical products such as name ,price, qty etc.

4.1.5 Chat

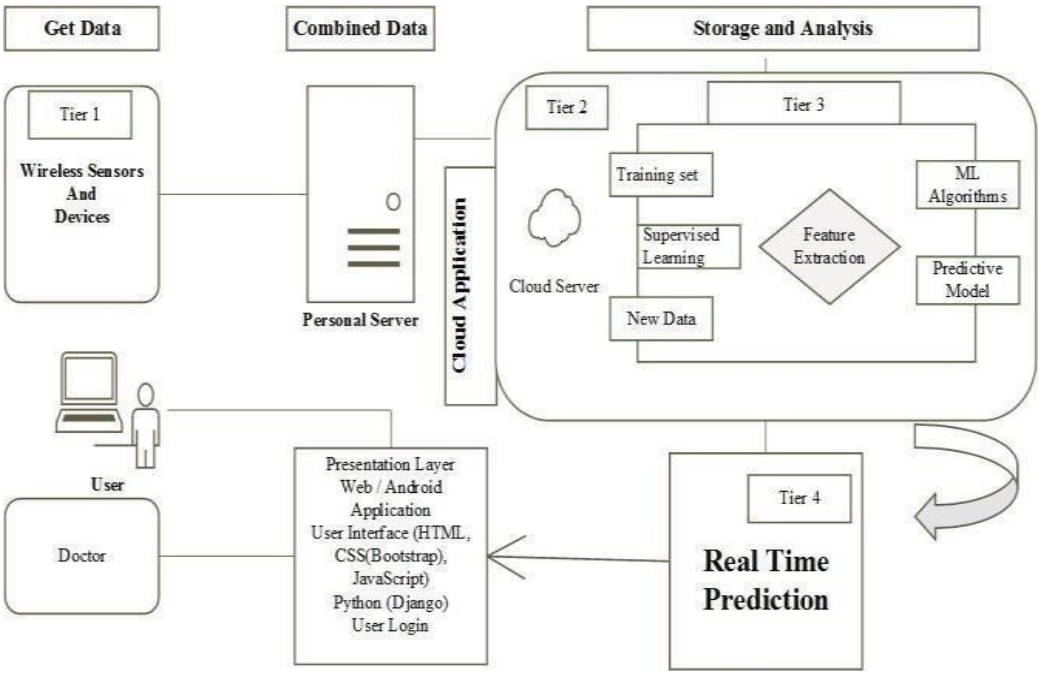
- shows the details between customer and doctor.

4.1.6 Payment

- shows the order and payment of customers wit product details.

4.2 SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system. The working of this system is described as follows: Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.



In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

- Supervised Learning

  Supervised learning is the type of machine learning in which machines are trained using well ”labelled” training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

- Unsupervised learning

  Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format. Unsupervised learning is helpful for finding useful insights from the data. Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to the real AI. Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important. In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

- Reinforcement learning

  Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

4.3 RANDOM FOREST ALGORITHM

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexity of the worst case of learning with Random Forests is O(M(dnlogn)) , where M is the number of growing trees, n is the number of instances, and d is the data dimension.It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Chapter 5

TESTING

5.1 Introduction

Testing is the process of examining the software to compare the actual behaviorwith that of the excepted behavior. The major goal of software testing is to demonstrate that faults are not present. In order to achieve this goal, the tester executes the program with the intent of finding errors. Though testing cannot show absence of errors but by not showing their presence it is considered that these are not present.

System testing is the first Stage of implementation, which is aimed at ensuring that the system works accurately and efficiently before live operations commences. Testing is vital to the success of the system. System testing makes a logical assumption that if all the parts of the system are correct and the goal willbe successfully achieved. A series of testing are performed for the proposed system before the proposed system is ready for user acceptance testing.

-Levels of Testing:

1. Unit Testing
2. Integration Testing
3. Validations
4. Output Testing

A. Unit Testing

In this each module is tested individually before integrating it to the final system. Unit test focuses verification in the smallest unit of software design in each module. This is also known as module testing as here each module is tested to check whether it is producing the desired output and to see if any error occurs.

B. Integration Testing

Integration testing (sometimes called integration and testing abbreviated I&T) is the phase in software testing in which individual software modules are combined and tested as a group. It occurs after unit testing and before validation testing.

Integration testing takes as its input modules that have been unit tested, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output the integrated system ready for system testing. The purpose of integration testing is to verify functional, performance, and reliability requirements placed on major design items.

C. Output Testing

No system could be useful if it does not produce the required output in the specific format. Output testing is performed to ensure the correctness of the output and its format. The output generated or displayed by the system is tested asking the users about the format required by them.

D. Validation Testing

In software project management, software testing, and software engineering, validation is the process of checking that a software system meets specifications and that it fulfills its intended purpose. The errors which are uncovered during integration testing are connected during this phase.

1. Unit Testing

Form	Procedure	Expected Result	Actual Result	Status
Welcome Form	Choose whether to Login or Register			
Login Form	Enter valid username and password	Should validate user and provide link to user accounts	Got entry to accounts	Pass
Category Form	Enter all mandatory fields	Should validate all entered fields and flash a message indicating successful registration	Message indicating successful registration is shown	Pass
Subcategory Form	Enter all mandatory fields	Should validate all entered fields and flash a message indicating successful registration	Message indicating successful registration is shown	Pass

2. Integration Testing

Form	**Expected Result**	**Actual Result**	**Status**
Login and user account forms	Get entry to appropriate user page	Appropriate user page is displayed	Pass
Category Form	Must add category details successfully	Insertion is successful	Pass
Subcategory Form	Must add subcategory details successfully	Insertion is successful	Pass
Item Form	Must add Item details successfully	Insertion is successful	Pass
Order Form	Must update the specified entry in the database	Specified entry updated	Pass
Chat	Must calculate and display the results	Results displayed	Pass

3. Validation Testing

Form	**Expected Result**	**Actual Result**	**Status**
Create user	Check all mandatory fields and validate all entered data fields	If any error found display message and the same screen is displayed else record saved and confirmed	Pass
Edit User	Edit the row corresponding to the value entered	If the value entered is invalid error message is thrown otherwise message indicating successful deletion is flashed	Pass

Form	**Expected Result**	**Actual Result**	**Status**
Relay	It control both bulb and fan loads remotely	Relay works normally as excepted	Pass
Pulse sensor	It gives the lives usage of current for loads	Application works perfectly	Pass

Chapter 6

CONCLUSIONS

Heart diseases are a major killer in India and throughout the world application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases ison a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the seven different machine learning algorithms used to measure the performance are Random Forest applied on the dataset.

The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this n features have to be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed.

All the seven machine learning methods accuracies are comparedbased on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, andf1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 81%.

6.1 Future Scope

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find KNN is best one.

For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rateof the death cases can be minimized by the awareness about the diseases.

Chapter 7

APPENDIXES

7.1 CODING

×

```
mport the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selectionimporttrain_test_split, Stratified KFold, cross_val_score
from sklearn.pipeline import make_pipeline, Pipeline fromsklearn.model_selection
import GridSearchCV
from sklearn.svm import SVC
from sklearn.naive_bayesimportMultinomialNB
from sklearn.linear_modelimportLogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.externals import joblib
from sklearn.metrics import make_scorer, f1_score, recall_score, precision_score
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.metrics import log_loss
```

```
import warnings
warnings.simplefilter(action = 'ignore', category= FutureWarning)
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import AdaBoostClassifier
import numpy as np
from flask import Flask,request,jsonify, render_template
import pickle
app=Flask(__name__,template_folder='template')
app.static_folder = 'static'
model1=pickle.load(open('model1.pkl','rb'))
model2=pickle.load(open('model2.pkl','rb'))
@app.route('/home')
def homepage():
return render_template('index.html')
@app.route('/precautions')
def precautions():
return render_template('precautions.html')
@app.route('/advancedpage')
def advancedpage():
return render_template('index.html')
@app.route('/quick',methods=['POST'])
def quick():
def bmi(height,weight):
bmi=int(weight)/((int(height)/100)**2)
return bmi
int_features1 = [float(x) for x in request.form.values()]
age=int_features1[1]
cigs=int_features1[3]
height=int_features1[8]
weight=int_features1[9]
hrv=int_features1[10]
int_features1.pop(8)
```

int $_f$ *eatures*1.*pop*(9)
bmi=round(bmi(height,weight),2)

int $_f$ *eatures*1.*insert*(8, *bmi*)

if int(int $_f$ *eatures*1[0]) == 1.0 :
sex="Male"

else:

sex="Female"

if int(int $_f$ *eatures*1[2]) == 1.0 :
smoking="Yes"

else:

smoking="No"

if int(int $_f$ *eatures*1[4]) == 1.0 :
stroke="Yes"

else:

stroke="No"

if int(int $_f$ *eatures*1[5]) == 1.0 :
hyp="Yes"

else:

hyp="No"

if int(int $_f$ *eatures*1[7]) == 1.0 :
dia="Yes"

else:

dia="No"

if int(int $_f$ *eatures*1[6]) == 1.0 :
bpmeds="Yes"

Random Forest(RF)

classifierRF=RandomForestClassifier()

classifierRF.fit(x$_t$*rain*, y_t*rain*)
get predictions from best model above

y_p*reds* = *model*$_c$*lassi fierDT.predict*(x_t*est*)

print('Decision Tree accuracy score: ',accuracy$_s$*core*(y_t*est*, y_p*reds*))
print('')

```
import pylab as plt
labels=[0,1]
cmx=confusion_matrix(y_test, y_preds, labels)
print(cmx)
fig = plt.figure()
ax = fig.add_subplot(111)
cax = ax.matshow(cmx)
plt.title('Confusion matrix of the classifier')
fig.colorbar(cax)
ax.set_xticklabels(['']  +  labels)
ax.set_yticklabels(['']  +  labels)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
classifierRF.score(x_test,  y_test)
print('RandomForestLogLoss score'.format(score=log_loss(y_test,
classifierRF.predict_proba(x_test))))
clfs.append(classifierRF)
save best model to current working directory
joblib.dump(classifierRF, "heart_disease.pkl")
load from file and predict using the best configs found in the CV step
model_classifierRF  =  joblib.load("heart_disease.pkl")
get predictions from best model above
y_preds  =  model_classifierRF.predict(x_test)
print('Random Forest accuracy score: ',accuracy_score(y_test, y_preds))
print('')
import pylab as plt
labels=[0,1]
cmx=confusion_matrix(y_test, y_preds, labels)
print(cmx)
fig = plt.figure()
ax = fig.add_subplot(111)
```

```
cax = ax.matshow(cmx)
plt.title('Confusion matrix of the classifier')
fig.colorbar(cax)
ax.set_xticklabels(['']  +  labels)
ax.set_yticklabels(['']  +  labels)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
print('')
print(classification_report(y_test, y_preds))
def precautions():
return render_template('precautions.html')
@app.route('/advancedpage')
def advancedpage():
return render_template('index.html')
@app.route('/quick',methods=['POST'])
def quick():
def bmi(height,weight):
bmi=int(weight)/((int(height)/100)**2)
return bmi
int_features1 = [float(x) forxinrequest.form.values()]
age=int_features1[1]
cigs=int_features1[3]
height=int_features1[8]
weight=int_features1[9]
hrv=int_features1[10]
int_features1.pop(8)
i
```

7.2 EXPERIMENTS

A. Data Preparation

Since the dataset consists of 4240 observations with 388 missing data and 644 observations to berisked for heart disease, two different experiments were performed for data preparation. First, wechecked by dropping the missing data, leaving with only 3751 data and only 572 observations risked for heart disease. This leads to reduced number of the observations providing irrelevant training to our model. So, we progressed with imputation of data with the mean value of the observations and scaling the musing Simple Imputer and Standard Scaler modulesof Sklearn.

B. Exploratory Analysis

Correlation Matrix visualization Before Feature Selection shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through plots and bar graphs.

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heart
0	1.153113	-1.234283	-0.988276	-0.758082	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.940825	-1.196267	-1.083027	0.287258	0.34
1	-0.867217	-0.417904	-0.988276	-0.758062	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.300085	-0.515199	-0.159365	0.719808	1.59
2	1.153113	-0.184349	1.011863	0.928410	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.187218	-0.220388	-0.243328	-0.113213	-0.0
3	-0.867217	1.332233	1.011863	1.787148	-1.758000e-01	-0.077014	1.489778	-0.162437	-0.263965	0.800948	1.018227	0.682815	-0.90
4	-0.867217	-0.417904	1.011863	1.177931	-1.758000e-01	-0.077014	-0.671241	-0.162437	1.089700	-0.106878	0.092865	-0.663664	0.76
...	...	...	...	...	...	...	...	...	...	...	...	...	
4235	-0.867217	-0.184349	1.011863	0.928410	2.059493e-17	-0.077014	-0.671241	-0.162437	0.284061	-0.061487	-0.915087	-0.933810	0.87
4236	-0.867217	-0.650984	1.011863	0.504542	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.002365	-0.265747	0.344465	-1.031304	0.84
4237	-0.867217	0.282295	-0.988276	-0.758062	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.728764	0.051991	0.008565	-1.084628	0.34
4238	1.153113	-1.117623	-0.988276	-0.758082	-1.758000e-01	-0.077014	1.489778	-0.162437	-1.168446	0.382428	1.269138	-0.048834	-0.73
4239	-0.867217	-1.234283	1.011863	1.787148	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.918263	0.009296	0.280498	-1.201616	0.75

4240 rows × 14 columns

Figure 7.1: Datasets

C. <u>Confusion Matrix</u>

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or "classifier") on a set of testdata for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the numberof errors made.

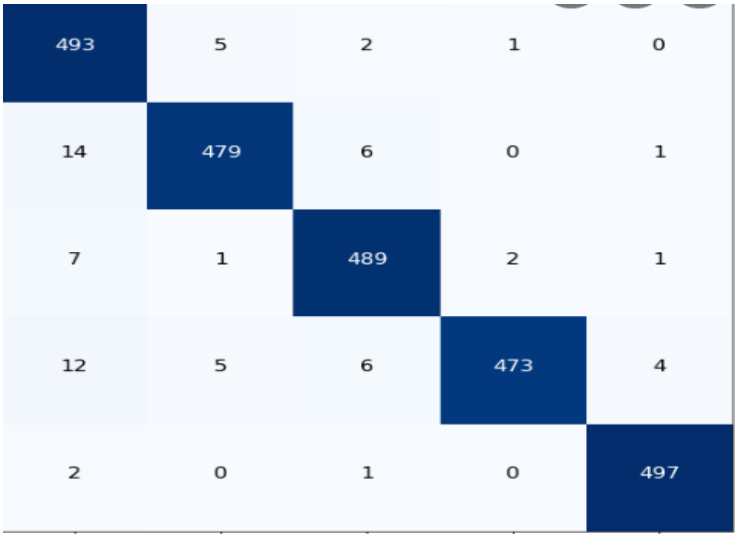


Figure 7.2: Confusion Matrix

7.3 SCREEN SHOTS

- Registration Screen
- Home Screen
- Customer home
- prediction and pulse rate
- product
- cart
- payment
- chat
- doctor home

- doctor chat
- admin home
- add doctor and view doctors
- view customers
- add products and view products
- view payment

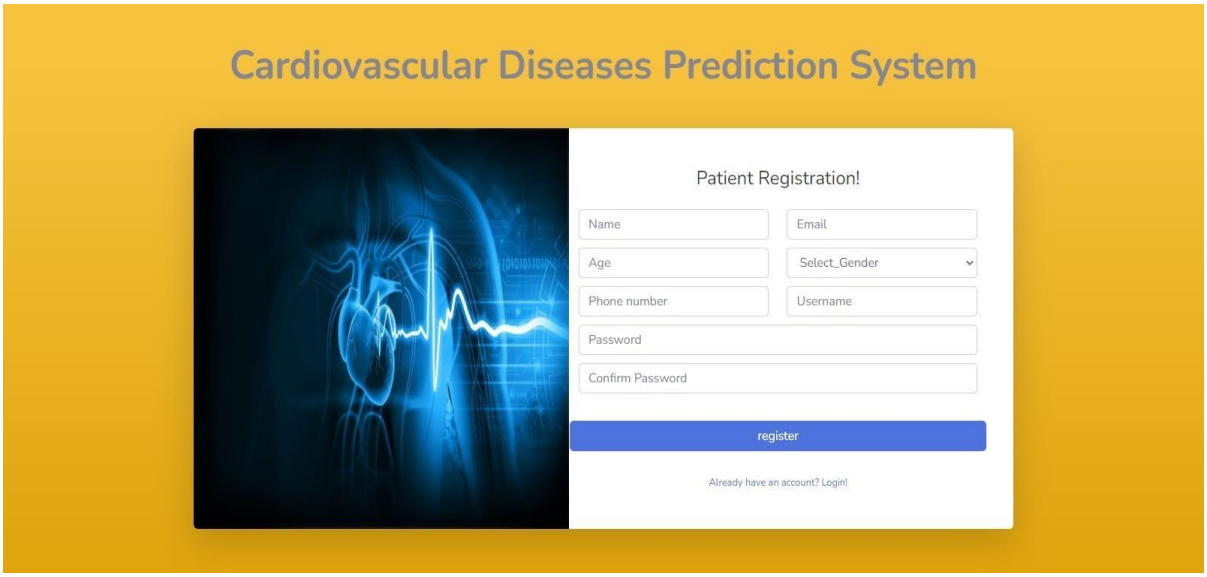


Figure 7.3: User Registration

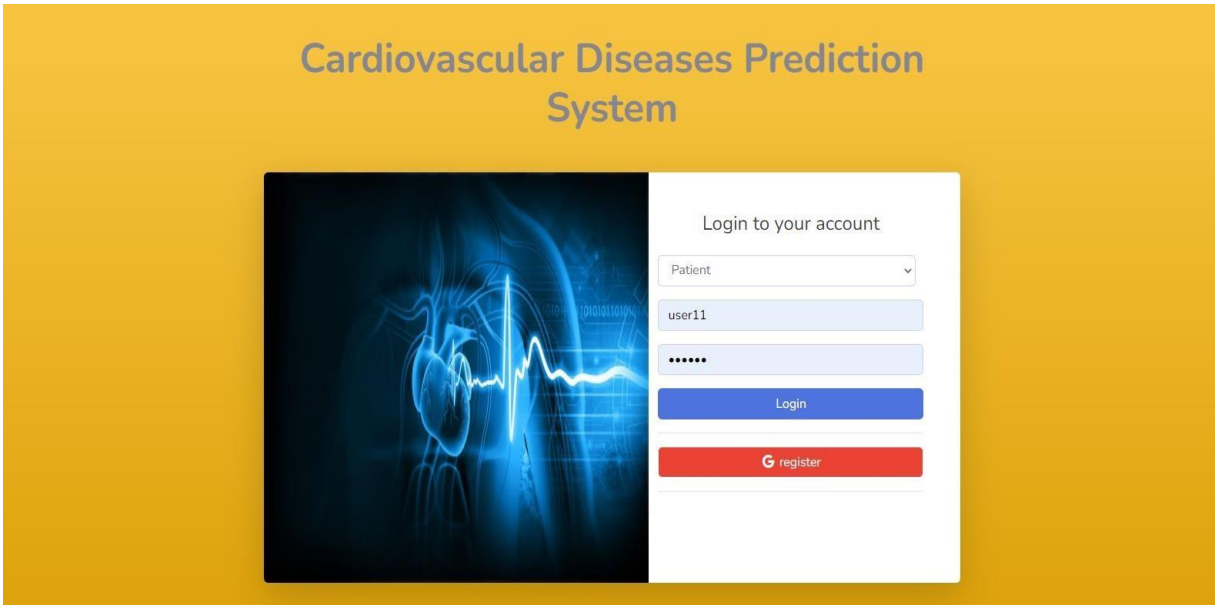


Figure 7.4: Login Screen

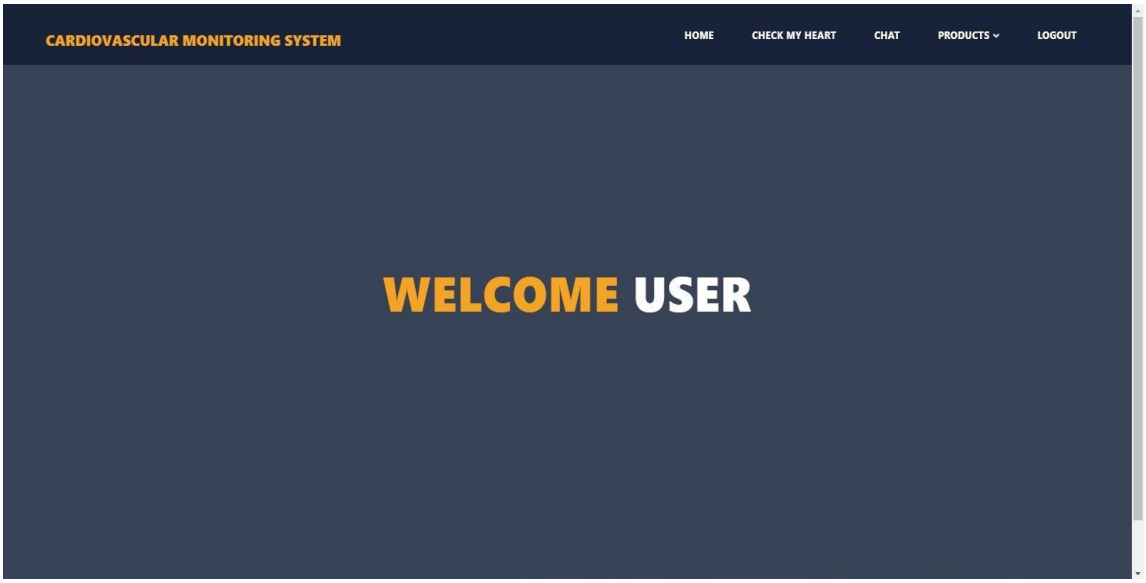


Figure 7.5: Patient Home Screen

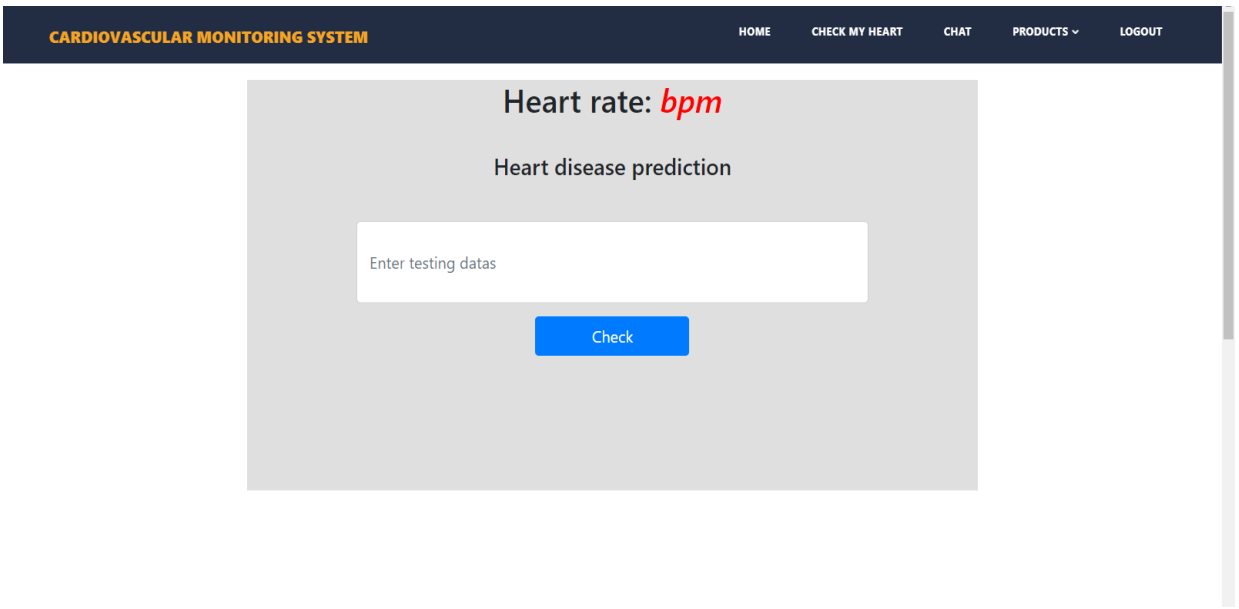


Figure 7.6: Prediction and heart rate

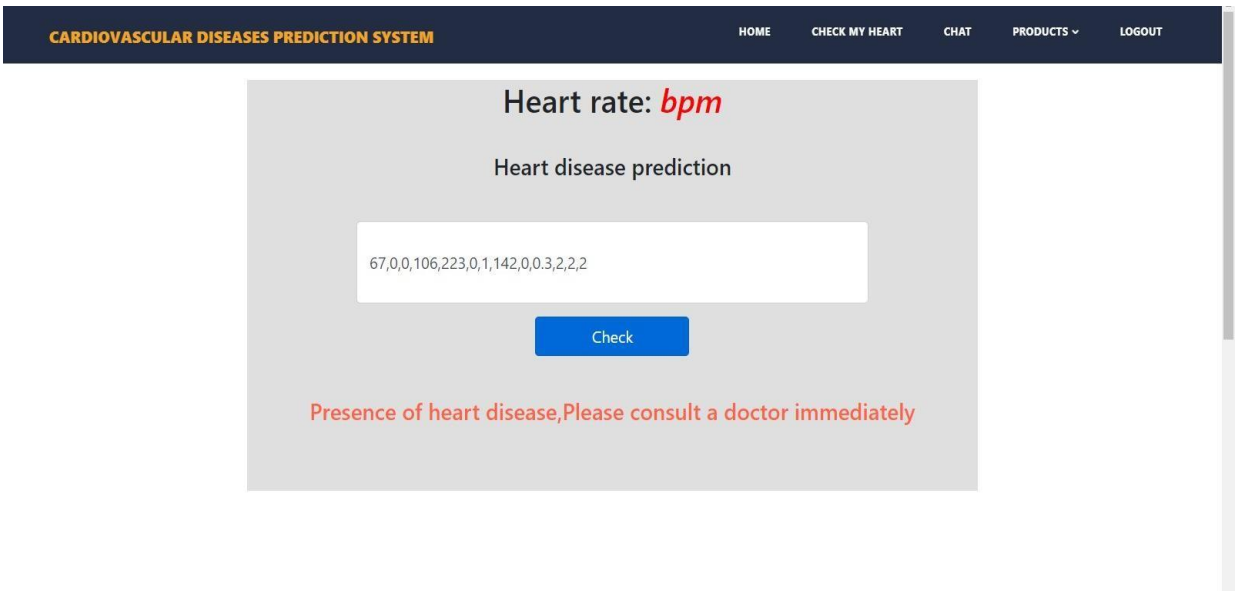


Figure 7.7: Prediction and heart rate with diseases

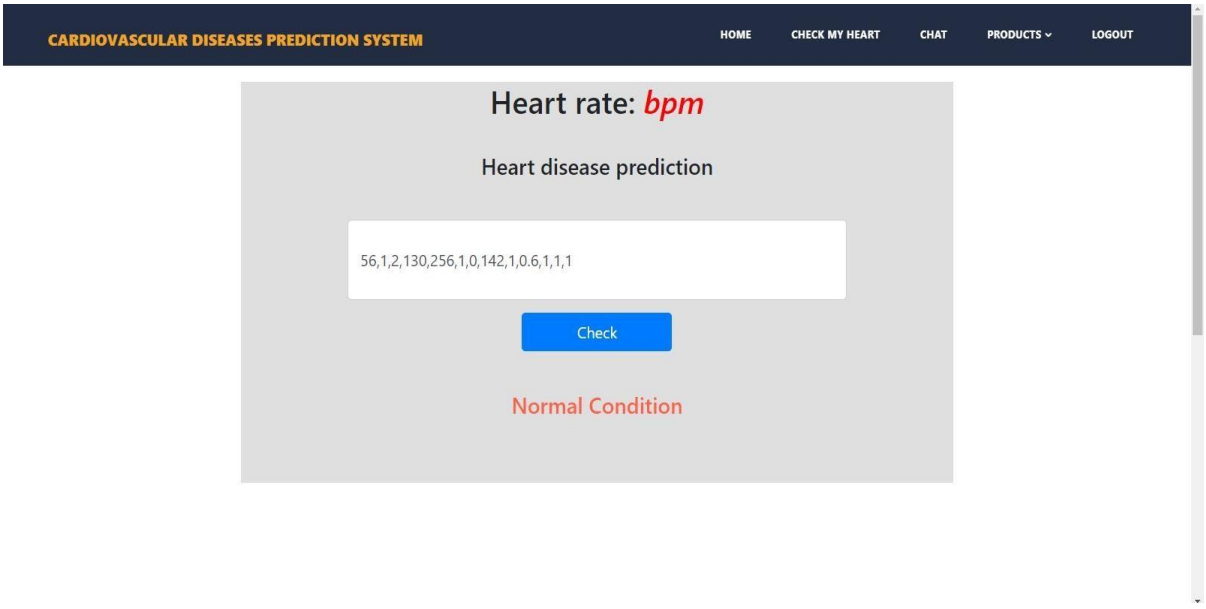


Figure 7.8: Prediction and heart rate with normal condition

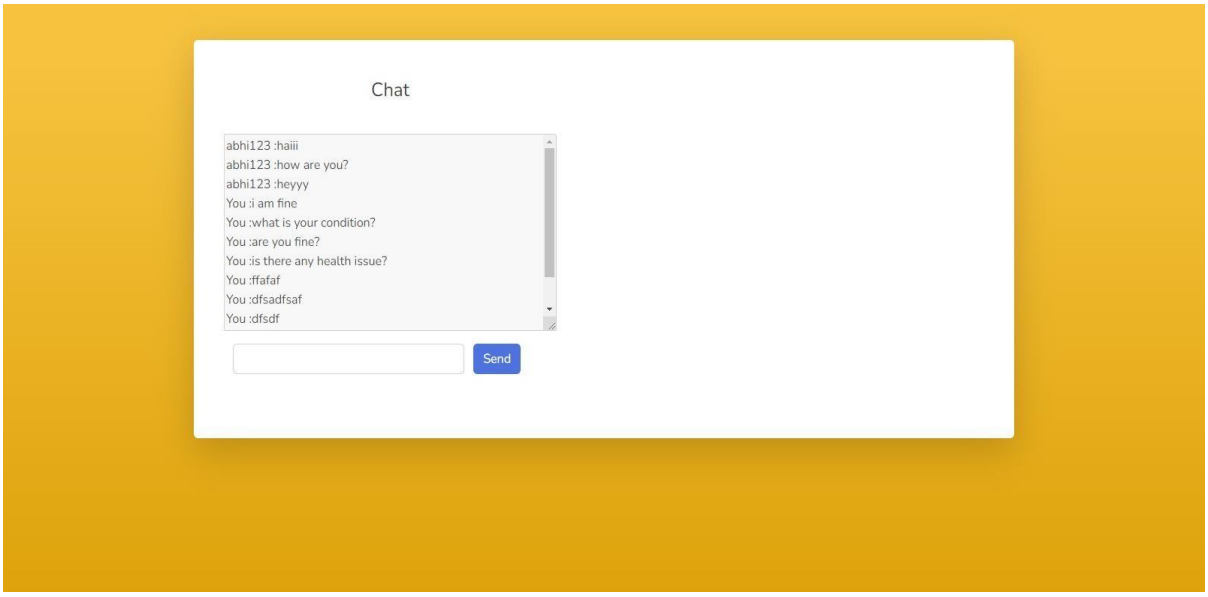


Figure 7.9: chat with doctor

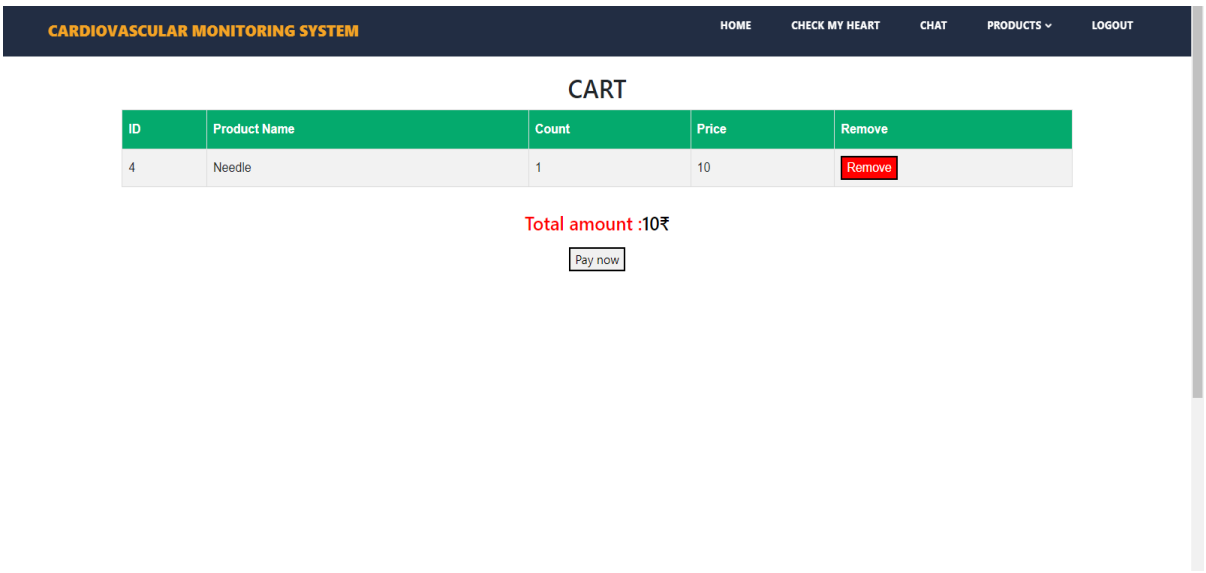


Figure 7.10: cart

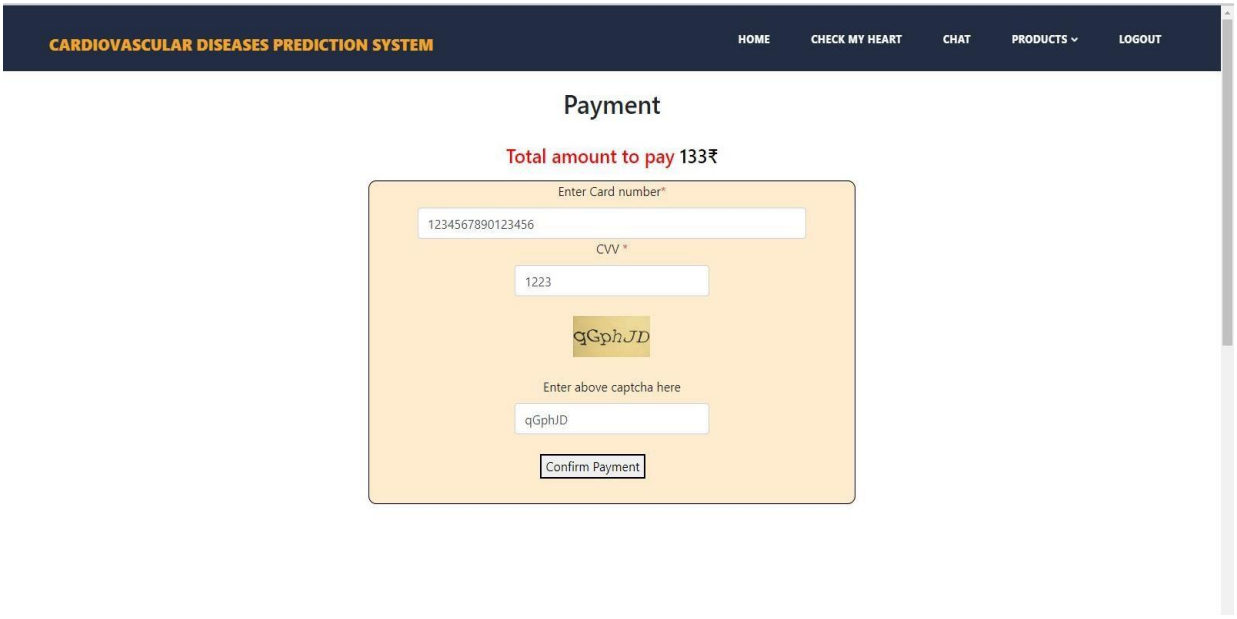


Figure 7.11: payment

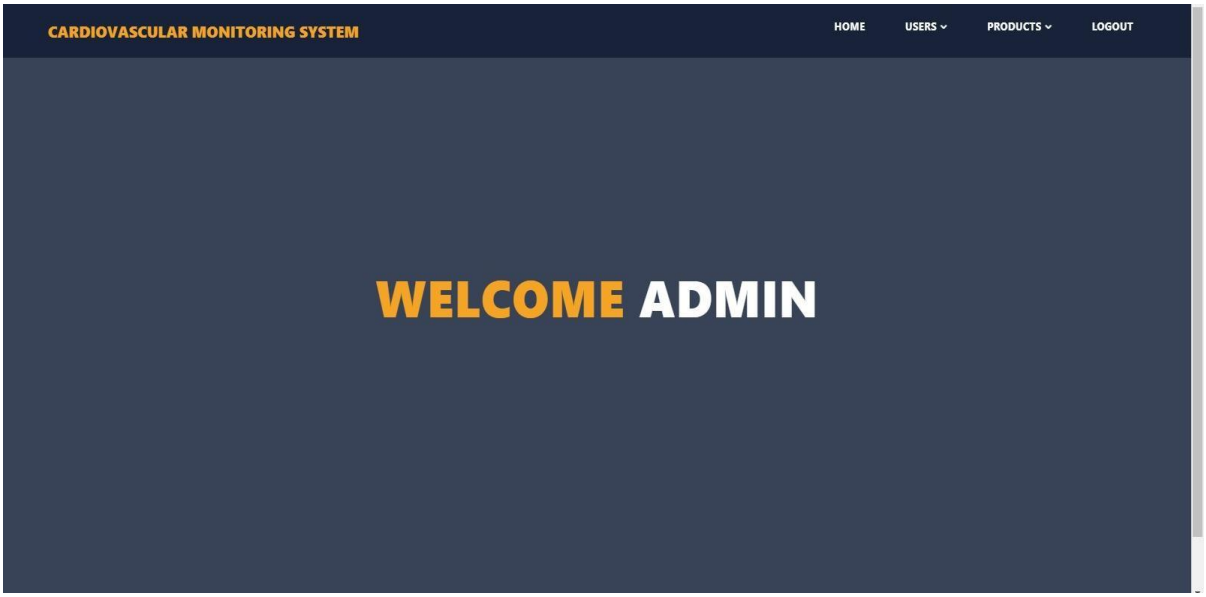


Figure 7.12: admin home

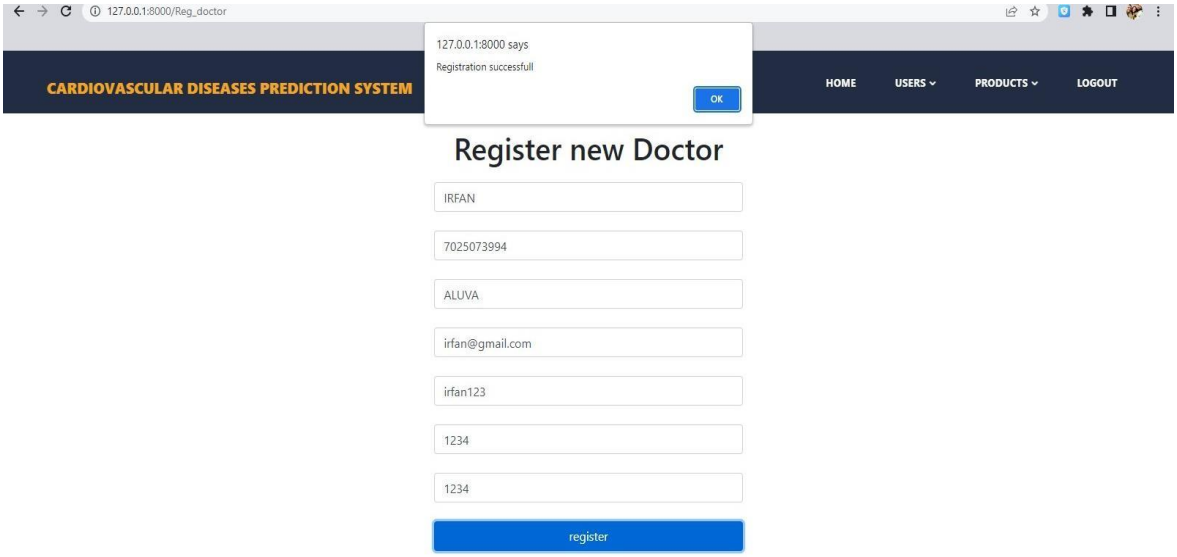


Figure 7.13: add doctor

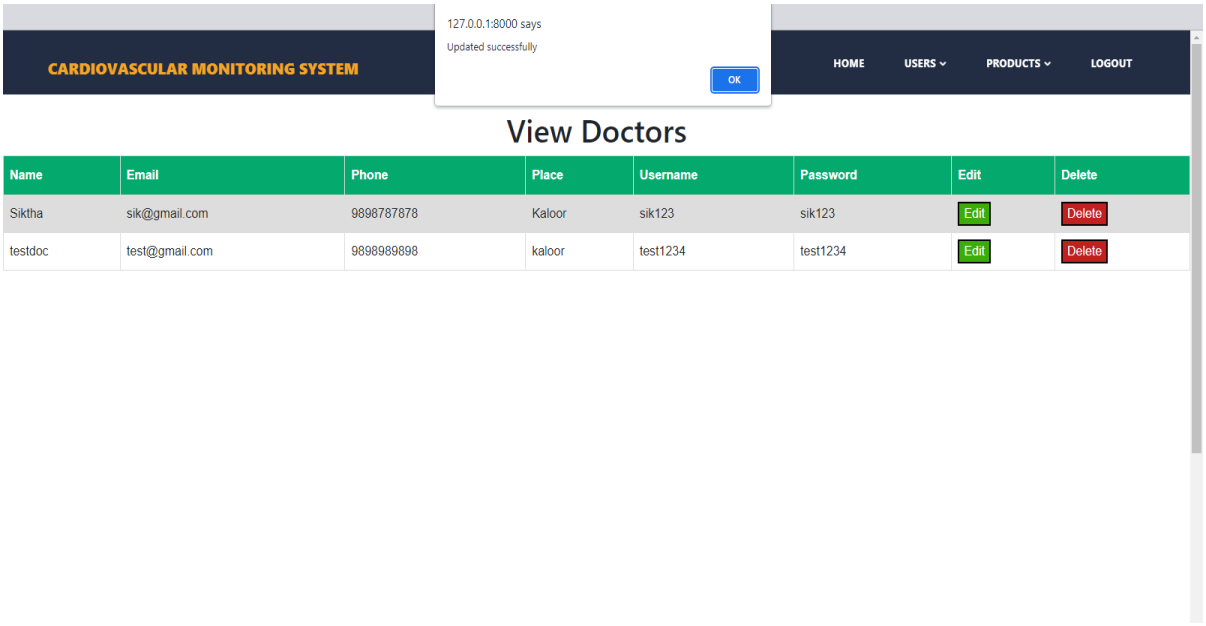


Figure 7.14: View doctor

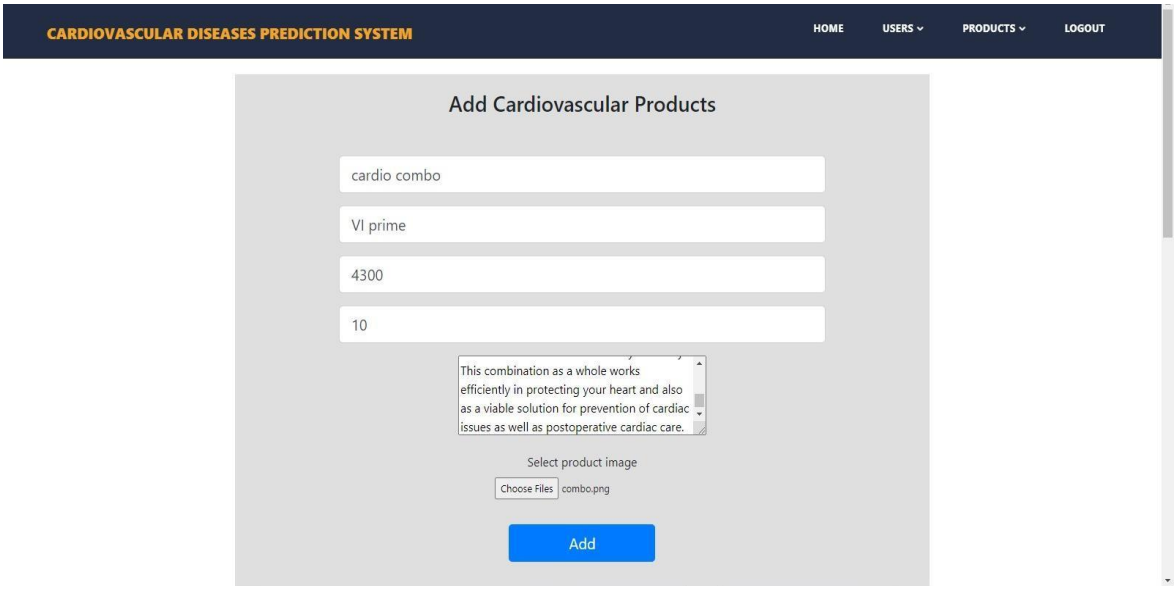


Figure 7.15: Add products

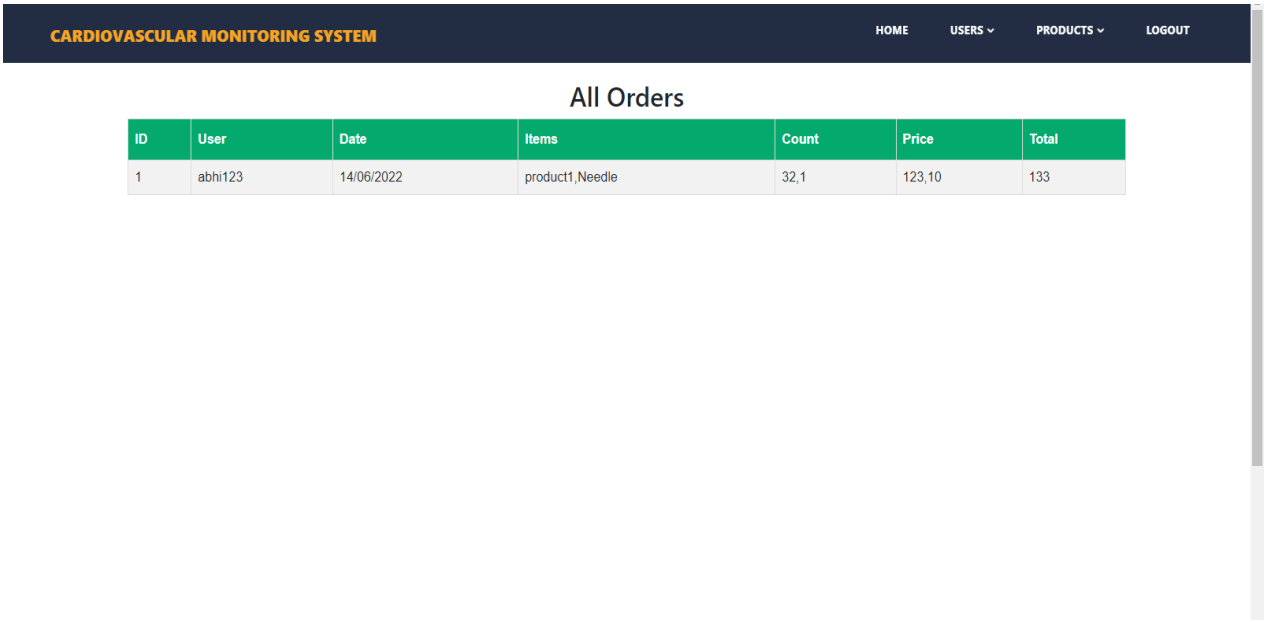


Figure 7.16: Orders

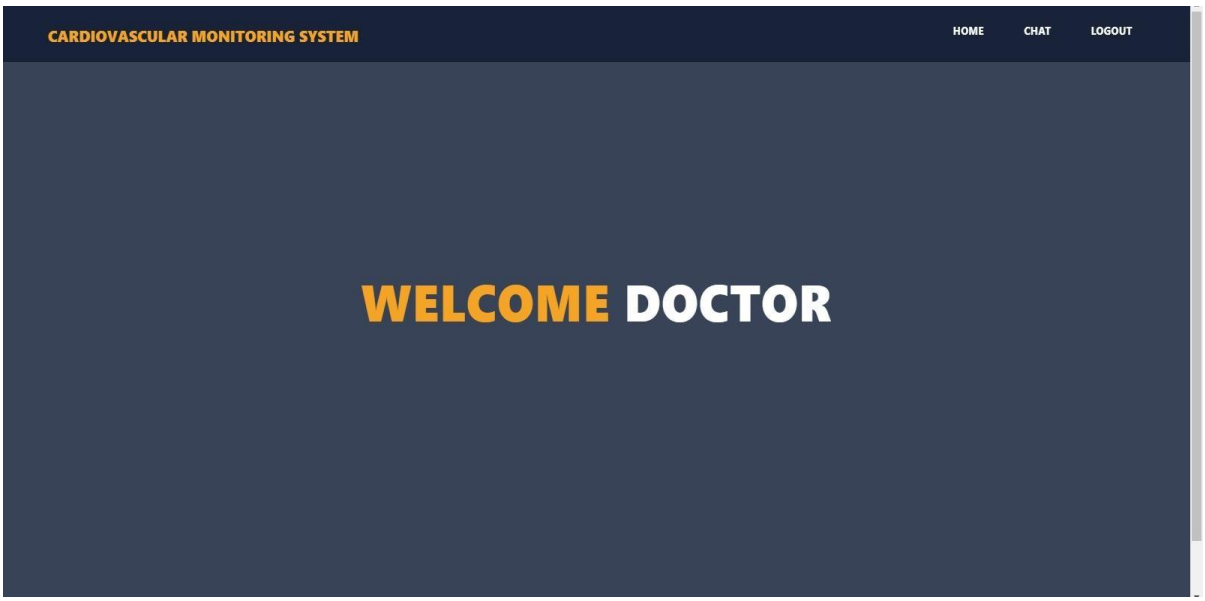


Figure 7.17: doctor home

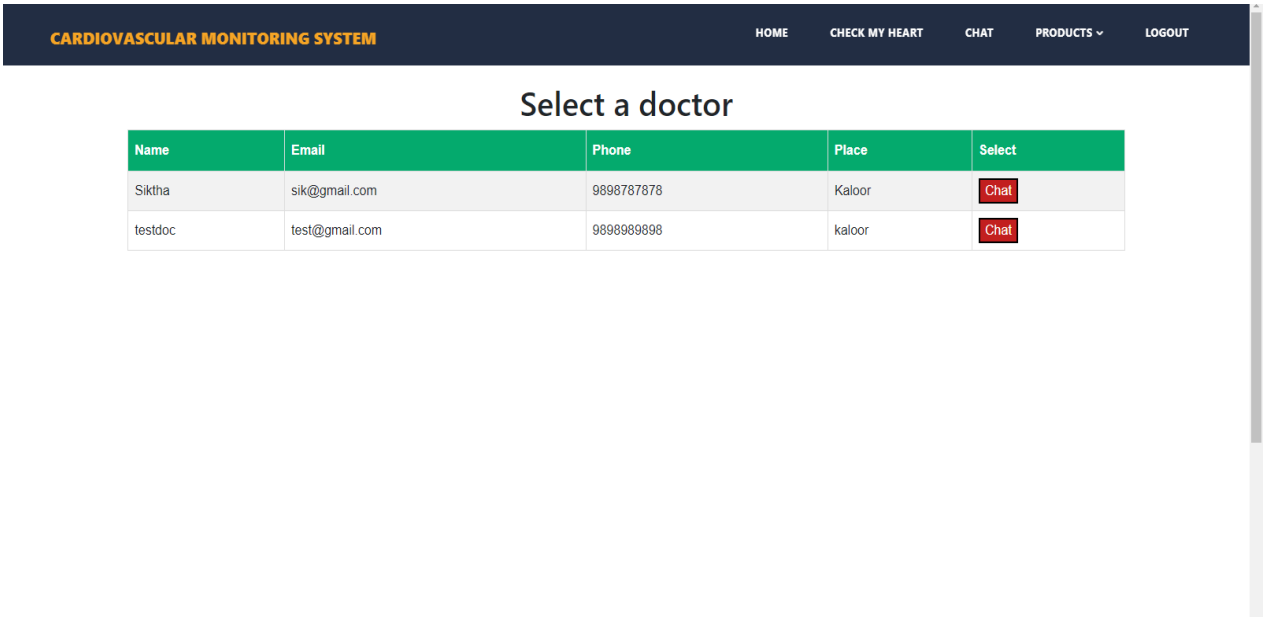


Figure 7.18: doctor chat view

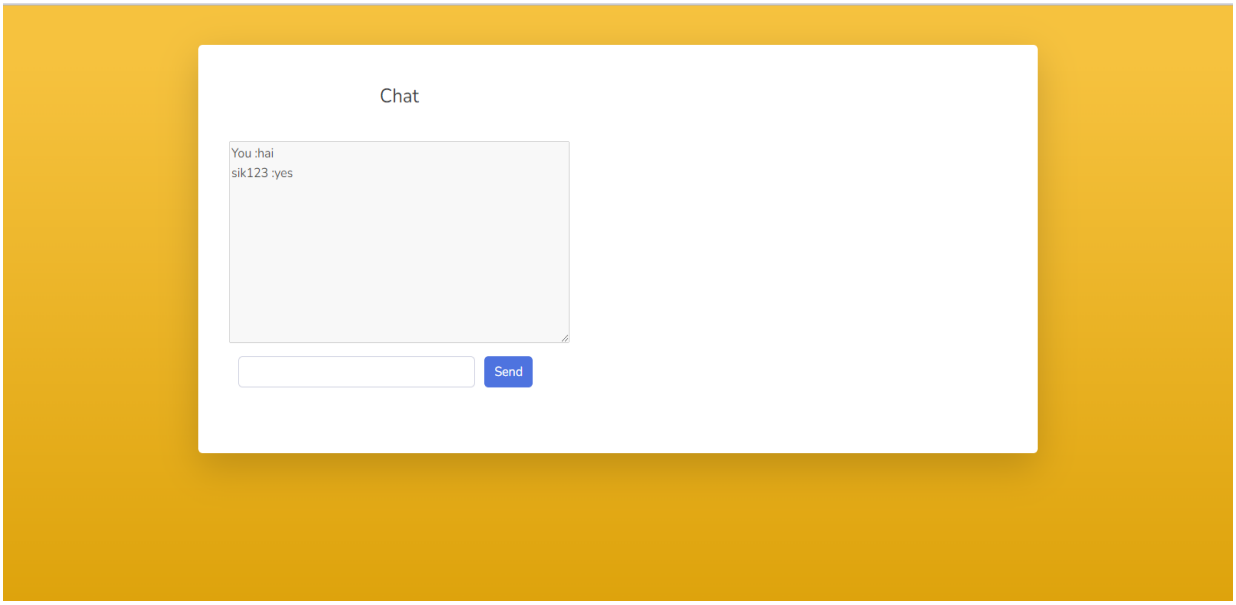


Figure 7.19: chat of a particular patient

Chapter 8

REFERENCES

1.Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

2.Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

3.Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

4.Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

5.Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.