

HEALTHCARE PREDICTOR USING ML ALGORITHM

A FINAL YEAR PROJECT

Submitted by

MALASREE RALLAPALLI (9916004214)

NAGALAKSHMI PABBISSETTY (9916004107)

VAMSHIKRISHNA BANDARI (9916004019)

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING



SCHOOL OF COMPUTING

COMPUTER SCIENCE AND ENGINEERING

KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION

KRISHNANKOIL - 626 126

Academic Year-Even 2019-2020



School of Computing
Department of Computer Science and Engineering
Project Summary

Project Title	HEALTHCARE PREDICTOR USING ML ALGORITHM		
Project Team Members (Name with Register No)	Malasree Rallapalli Nagalakshmi Pabbisetty Vamshikrishna Bandari	9916004214 9916004107 9916004019	
Guide Name/Designation	Dr K KARTHEEBAN (Associate Professor, Department of Computer Science and Engineering)		
Program Concentration Area	Prediction and Precautions		
Technical Requirements	Hardware requirements: 1. 4 GB RAM with I3 core processor. 2. Hard Disk of 160GB. Software requirements: 1. Anaconda 5.3 2. MySQL 6.0 3. Windows 10 OS		
Engineering standards and realistic constraints in these areas			
Area	Codes & Standards / Realistic Constraints		Tick ✓
Economical	Healthcare Predictor is considered to be economical constraint than direct way of diagnosing the diseases in hospital. It is economical as it does not involve any costly tools except the system configuration.		✓
Social	The diseases the project involved and predicted are the most anticipated diseases around the world. In addition to this it also creates awareness on these diseases that people have. Hence it follows Social Constraint.		✓
Health and Safety	This project follows health and safety constraint by the means of involving people in the project to ensure their good health and safety.		✓
Sustainability	This project uses open source software including Python, scikit-learn, Flask framework and MySQL. The core part is to deploy a model to predict diseases using free technologies that being recommending users to go through the project finds its place in every student/ researcher online presence. Hence the project is sustainable.		✓

REALISTIC CONSTRAINTS

ECONOMIC CONSTRAINTS

The project budget is adequate, allocated appropriately in terms of quality, safety, functionality and performance. The cost is less due to usage of more open source software's like python, scikit-learn, MySQL. The individuals can reduce the medical expenses spent for diagnosis, doctor appointment, travelling etc. The project doesn't have any significant delays or additional costs for the users.

SOCIAL CONSTRAINTS

This project construction work could include people's involvement. The diseases we considered in this project are more frequent occurring diseases all over the world. This helps individuals to become familiar about the drugs, medicines they are consuming for prevention of disease. It creates social awareness among individuals related to particular diseases. The people can have better perspective view about precautions related to diseases. The individuals can save time by these trained models. Thus, it considers social constraints.

HEALTH AND SAFETY

The project is related to medical sector so it includes all health and safety constraints. This project helps the individuals to have a better perspective idea about the medicines they consume to prevent a particular disease and to have a better health, they are provided with safety and an ease diagnosis and analysis approach. The project demonstrates the consistency of software safety constraints, specification. The requirement elicitation is maintained in a safety manner and condition.

SUSTAINABILITY

The project has high degree of security related to user details .it is built on python framework which are open source applications. This helps user for managing time, budget, quality. It provides good end user output and initiates a gateway for future scope and research. So, whenever there is new update in the features used it freely apt to it. Thus, making the application easily available, usable and sustainable

ENGINEERING STANDARDS

This project complies to **2755-2017** - IEEE Guide for Terms and Concepts in Intelligent Process. This standard is intended to provide a set of definitions established by and for the community involved with *software-based intelligent process automation (SBIPA)* so that when terminology is used, all understand the meaning. This IEEE standard has emerged recently. Because of the newness of this kind of automation capability, there are no common definitions of concepts, capabilities, terms, technology, types, etc. This standard is published for the purpose of promoting clarity and consistency in the use of Software Based Intelligent Process Automation (SBIPA) terminology. The definitions represent the consensus of a diverse panel of industry participants.

The health care predictor uses an intelligence to predict the diseases. It is a complete software package which is developed using user interface and a database named MySQL. Thus, it satisfies the engineering constraints of SBIPA - *software-based intelligent process automation* of **2755-2017** - IEEE Guide for Terms and Concepts in Intelligent Process.

DECLARATION BY THE STUDENT

We Mr. Vamshikrishna Bandari, Miss Nagalakshmi Pabbisetty & Miss Malasree Rallapalli students of **B. Tech (Semester – 8th) Computer Science and Engineering**, Roll No. **9916004019, 9916004107 & 9916004214** hereby declare that the final year project titled, **“Healthcare Predictor Using ML Algorithm”** submitted by us in partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Kalasalingam Academy of Research and Education, examination during the academic year 2019-2020,

This is the actual work carried by us under the guidance and supervision of **Dr K Kartheeban**, Associate Professor KARE. We further state that this work is original and not submitted anywhere else for any examination.

VAMSHIKRISHNA BANDARI
9916004019

NAGALAKSHMI PABBISSETTY
9916004107

MALASREE RALLAPALLI
9916004214

Signature of student



KALASALINGAM
ACADEMY OF RESEARCH & EDUCATION
(DEEMED TO BE UNIVERSITY)
Under sec. 3 of UGC Act 1956. Accredited by NAAC with "A" Grade



BONAFIDE CERTIFICATE

Certified that this project report titled **“Healthcare Predictor Using ML Algorithm”** is the Bonafide work of **“Vamshikrishna Bandari, Nagalakshmi Pabbisetty & Malasree Rallapalli”**, who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SUPERVISOR

Dr K KARTHEEBAN

Associate Professor

Computer Science and Engineering

Kalasalingam Academy of Research and
Education - Krishnankoil – 626126

HEAD OF THE DEPARTMENT

Dr A. FRANCIS SAVIOUR DEVARAJ

Professor

Computer Science and Engineering

Kalasalingam Academy of Research and
Education - Krishnankoil – 626126

Submitted to the project viva-voice examination held at Kalasalingam Academy of Research and Education, Krishnankoil on_____.

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

An exchange of ideas generates a new object to work in a better way. Apart from the ability labour and time devotion, guidance and co-operation are two pillars for the success of a project. Whenever a person is helped or co-operated by others, his heart is bound to pay gratitude to others. A satiation and pleasure that accompany the successful completion of task would be incomplete without the mention of the people who have made it possible and whose consent guidance and encouragement served as a guiding light for the completion of the study.

We express deep sense of gratitude to “Kalvivallal” Thiru. **T. Kalasalingam** B.com., Founder Chairman, “Ilayavallal” **Dr.K.Sridharan** Ph.D., Chancellor, **Dr.S.Shasi Anand**, Ph.D., Vice President (Academic) , **Mr.S.Arjun Kalasalingam** M.S., Vice President (Administration) , **Dr.R.Nagaraj**, Vice-Chancellor, **Dr.V.Vasudevan** Ph.D., Registrar, **Dr.P.Deepalakshmi** M.E., Ph.D., Dean (School of Computing) . And also, a special thanks to **Dr.A.Francis Saviourdevaraj**. Head of Department of CSE, Kalasalingam Academy of Research and Education for granting the permission and providing necessary facilities to carry out Project work.

We would like to express our special appreciation and profound thanks to our enthusiastic Project Guide **Dr. K Kartheeban.**, Associate Professor/ CSE of Kalasalingam Academy of Research and Education [KARE] and Project Coordinator **Dr S. Dhanasekaran** M.E., Ph.D., Associate Professor/ CSE of Kalasalingam Academy of Research and Education [KARE] for his inspiring guidance, constant encouragement with our work during all stages. I am extremely glad that I had a chance to do my Project under my Guide, who truly practices and appreciates deep thinking. I will be forever indebted to my Guide for all the time he has

spent with me in discussions. And during the most difficult times when writing this report, he gave me the moral support and the freedom I needed to move on.

Besides my Project guide, I would like to thank the rest of Class committee members and all faculty members and Non-Teaching staff for their insightful comments and encouragement. We would fail in my duty if we don't thank our parents who are pillars of our lives. Finally, we would express our gratitude to all those who directly and indirectly helped us in completing this project.

INDEX

CHAPTER NO.	TITLE	PAGE NO.
	LIST OF FIGURES AND TABLES	X
	LIST OF ACADEMIC REFERENCE COURSES	X1
	ABSTRACT	XII
1	INTRODUCTION	1
	1.1 OVERVIEW	
	1.2 HEART DISEASE	
	1.3 MALARIA	
	1.4 TYPHOID	
	1.5 DENGUE	
2	LITERATURE REVIEW	4
3	SYSTEM ANALYSIS	7
	3.1 EXISTING WORK	
	3.2 ISSUES IN EXISTING WORK	
	3.3 PROPOSED WORK	
4	SYSTEM DESIGN	11
	4.1 COMPONENTS	
	4.1.1 HARDWARE COMPONENTS	
	4.1.2 SOFTWARE COMPONENTS	
	4.1.3 COMPONENTS DESCRIPTION	
	4.2 SYSTEM IMPLEMENTATION	
	4.2.1 MACHINE LEARNING PROCESS	
	4.2.2 MySQL PROCESS	
	4.3 DESIGN SPECIFICATION, STANDARDS AND CONSTRAINT	
	4.4 ALTERNATIVE DESIGN	
5	RESULTS AND DISCUSSIONS	17
6	CONCLUSION AND RECOMMENDATIONS	25
	APPENDIX	26
	REFERENCE	33

LIST OF FIGURES & TABLES

FIGURE NO.	TITLE	PAGE NO.
1	Architecture Diagram	8
2	Use Case Diagram	9
3	Correlation Graph	17
4	Model Deployment Code Snippet	18
5	Confusion Matrixes	18
6	Accuracy Bar Graph	19
7	Accuracy Table	19
8	Hyper Parameter Tuning Code Snippet	20
9	User Login	21
10	Signup Pages	21
11	Model UI Pages	21
12	Predictions Pages	22
13	Doctors Appointment	22
14	List of Doctor's Page	22
15	User Registered Page	23
16	Adding Doctor Page	23
17	Doctor Appointments Page	23

LIST OF ACADEMIC REFERENCE COURSES

S NO.	COURSE CODE	COURSE NAME
1	CSE401	Object Oriented Software Development
2	CSE102	Programming Languages
3	CSE103	Data Structures
4	CSE206	Object Oriented Programming
5	CSEX005	Introduction to Machine Learning and Pattern Recognition
6	MAT222	Probability and Statistics
7	CSE303	Software Engineering
8	CSE305	Database Management Systems
9	CSE307	Artificial Intelligence
10	CSE327	Data Mining and Data Ware Housing
11	CSE439	Machine Learning Techniques
12	CSE402	Internet Programming
13	CSEX008	Data Science with R

ABSTRACT

Healthcare in India amidst the ongoing COVID-19 epidemic is really crucial and a daunting task ahead of us. Every citizen needs immediate access to proper health guidance for their health condition/situation including maintenance or improvement of health via the prevention, diagnosis, treatment of disease, illness, injury, and other physical and mental impairments in humans. Health care is generally delivered by health professionals (providers or practitioners) in allied health fields. Health care can be done in different stages it may include providing primary care, secondary care, and tertiary care, as well as in public health.

Our work on Healthcare Prediction system targets this specific issue by providing health support to the public through an online consultation platform. The system is loaded with data collected from various accredited sources possessing various symptoms, disease or illness. When the user register in the website it allows user to share their symptoms and issues according to that the system processes the data by using appropriate model and guesses the most accurate illness that could be associated with patient's symptoms. On making sure the problem is addressed, direct consultation to a doctor is facilitated with a detailed report if needed by the end user.

This area of research is much needed as the ratio of doctors to patients and the affordability to reach and consult a doctor keeps decreasing. Though there are many others who have jumped into this sector/field, they have failed to provide a fool proof system which we are trying to develop by incorporating large sum of reliable data.

Chapter - I

INTRODUCTION

1.1 OVERVIEW

For centuries, our world continues to believe that medical services are a basic need of all the people. According to the world bank, Indian government has spent only 1.17 % of GDP towards healthcare. People are evolving more towards solutions which are more reliable, fast and sustainable in terms of cost and resources. The breath-taking pace of change in the way health care is financed and delivered has brought challenges and new activities to all participants in the healthcare system. Healthcare system needs to revamp itself to such solution set. Healthcare services are more important to maintain good health by improving health via the diagnosis, and treatment of disease, injury, and other physical and mental impairments in human beings.

Health care is generally delivered by medical professionals (providers or specialists) in health-related fields. Dentistry, midwifery, nursing, medicine, optometry, audiology, pharmacy, psychology, occupational therapy, physical therapy and other health professions are all part of health care. Health care can be done in different stages it may include providing primary care, secondary care, and tertiary care, as well as in public health. Sometimes we have seen situations where someone belonging to us may need doctors help immediately, but they are not available due to some reason. In present existing system there are many problems like frequently we want to visit a doctor even when we come across normal symptoms, illness, injury. It's tedious task for the user to wait for the doctor's appointment long time. generally, people are not aware of type of drugs and medicines that are essential to use for a particular disease. Even people are not much aware about the type of diseases and illness that person get affected. Even for every minor reason also, we have to reach the hospital.

We proposing healthcare predictor system is an online consultation project which is an end user support system. It allows users to get instant guidance on their health issues through an intelligent online health care system. The system is fed with various symptoms, illness associated with the user. The intelligent system allows user to share their symptoms and issues. Once the symptoms or issues faced by the user are given, then the system processes user's symptoms to check for different illness, diseases that could be associated with it. We are going

to use some intelligent techniques to guess the most accurate illness that could be associated with patient's symptoms. In these online system ,users can get instant guidance and precautions on their health issues through an intelligent health care system by using SVM(support vector machine algorithm), Random forest algorithm, logistic regression model Unified Modelling Language (UML) is used as a dialect for indicating, visualising, constructing and archiving the curios of programming, non-programming frameworks and for business demonstration.

The system is fed with various symptoms and the disease/illness associated with those systems. The system allows user to share their symptoms and issues. Once the symptoms or issues faced by the user are given, then the system processes user's symptoms to check for different illness, diseases that could be associated with it using Machine learning algorithm, it guesses the most accurate illness that could be associated with patient's symptoms and issues.

If the healthcare predictor system is not able to provide or predict suitable results, it intimates the user about the type of illness/disease or disorder it feels with the associated user's symptoms /issues. If user's symptoms are unable to match exactly with various disease in our database, the system shows the diseases user might probably get attacked by considering his/her symptoms. These online systems have another benefit it provides the user with doctor address, contact number along with Feedback and administrator dashboard for system operations.

1.2 HEART DISEASE

Heart disease is the leading cause of death for men and women. The most common type of disease is coronary artery disease. The CAD affects the flow of blood to the heart. When the blood flow decreases it causes to the heart attack. The risk factors of heart disease are high blood cholesterol, smoking and high blood pressure. The symptoms of heart attack are chest pain, neck pain, heart burn etc.

1.3 MALARIA

Malaria disease caused by protozoa and people who suffer with high fever and shaking chills. The symptoms of malaria are divided into two categories, they are Uncomplicated and

severe malaria. The symptoms for uncomplicated are cold, sweating, fever and headache. The severe malaria symptoms are fever, chills, deep breathing etc.

1.4 TYPHOID

The Typhoid is a disease which is caused by salmonella typhi bacteria. The symptoms are stomach pain, weakness, fatigue, rash, headache and loss of appetite etc.

1.5 DENGUE

It is disease caused by a bite of mosquito which is infected with one of the four dengue viruses. The symptoms are pain behind the eye, joint pain, severe headache, muscle pain. The three stages in dengue are critical stage, febrile stage, convalescent stage. In which it appears two to five days after the onset of fever.

Chapter - II

LITERATURE REVIEW

Basically, Literature survey is the one of the most important step or stage in software development process. Before developing the tool or model, it is more important to determine the time factor, economy and the company strength. Once all these categories are fulfilled, then the next steps are to determine which operating system and the programming language can be used to implement and to developing the proper tool. when the programmers start developing the tool the programmers generally need a lot of external support. This support can be obtained from the senior programmers or from books or from websites. Before building the system, the above steps have to be considered for developing the good proposed system.

[1] Yanwei Xing, Jie Wang and Zhihong Zhao, “Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease”

From decades the prediction and the survival of Coronary heart disease (CHD) has been a challenging research problem for every medical society. The main goal of this paper is to develop the best data mining algorithms for predicting the survival of CHD patients based on sample 1000 cases. During these processes we carry out a clinical observation and a 6-month follow up to collect the information about the 1000 CHD cases. The survival of disease information of each case is obtained via a proper follow up. Based on this obtained data, we have employed three most popular data mining algorithms to develop, implement the most prediction models using the sample 502 cases. We used the 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. According to the results we obtained that the SVM is the best predictor with a performance 92.1 % accuracy on the holdout sample of artificial neural networks came out to be the second best with 91.0% accuracy and the third model decision trees models came out to be the worst of the three with 89.6% accuracy. This different comparative study using multiple prediction models for the prediction of CHD patients along with method named a 10-fold cross-validation will helps us with a proper insight into the relative prediction ability of different data.

[2] SAM CHAO, FAI WONG, “An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining”

Decision tree is one the important inductive learning algorithms that provides an efficient, accurate, practical method for the generalizing classification of the rules from the previous cases which have been already solved by the domain experts. This method is considered to be one of the attractive algorithms for many real-life problems, applications, mostly due to the method interpretability. According to the Recent studies, many researches have been reported to provide a decision trees with the incremental learning ability, which is used to address the learning task, method with a flow of training instances. However, there are very minimum literatures exist discussing about the learning algorithms with the incremental learning ability according to the new attributes. In this paper, i+Learning (Intelligent, Incremental and Interactive Learning) theory is proposed to complement the traditional incremental decision tree learning algorithms by concerning new available attributes in addition to the new incoming instances. According to the experimental results the i+Learning method reveals to offer the promise of making appropriate decision trees that are more accurate, flexible, powerful and valuable paradigm, especially in the field of medical data mining research and community.

[3] M. Ilayaraja, T. Meyyappan, “Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm”

The data mining techniques plays a vital role in the process of analysing, extracting a huge amount of data from different sources and then summarizing it into the useful information. This obtained information can be transformed into the knowledge obtained from the historical patterns and future analysis, trends. Here Data mining plays a crucial and significant role in the field of information technology. Health care sector these days generates a huge amount of data about the patient’s details, symptoms, issues, hospitals resources, diagnosis methods, electronic patients’ records, etc. The data mining techniques are more useful to make the medicinal decisions in the way to cure diseases. The healthcare sector, industry generally collects a huge amount of data which is unfortunately, cannot be "mined" to discover the hidden patterns, information for proper decision making. This discovered knowledge can be used by the healthcare administrators, industry to improve and provide the quality of service. In this paper the authors developed a method, model to identify the

frequency of diseases in particular area in a given time period with the aid of association rule based on the Apriori data mining technique.

[4] My Chau Tu AND Dongil Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms"

Medical data mining has gained more popularity due to the data mining topic of late. Especially, the heart disease diagnosing is one of the most important issues faced. Many researchers working to develop the more intelligent medical decision support systems to help the physicians. In this paper, we propose the use of decision tree C4.5 algorithm, bagging with decision tree C4.5 algorithm and bagging with Naïve Bayes algorithm to identify the heart disease of a patient and then compare the correctness and effectiveness, correction rate among them. The data we study and used is generally obtained from patients having coronary artery disease

[5] Ranjit Abraham, Jay B.Simha, Iyengar, "A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier"

Naive Bayes classifier another important model which became more popular as a probability-based classification method despite of its assumption that the class label has conditionally mutually independent attributes. This paper is a study into discretization techniques to improve the classification accuracy of Naive Bayes with respect to medical datasets. From the experimental results obtained to suggest on an average, the minimum description length (MDL) discretization of the Naive Bayes classifier is one of the best performers compared to other variants of Naive Bayes as well as some popular non-Naive Bayes statistical classifiers.

Note: It is the basis for every research article as it provides us immense knowledge, plethora of perspectives and recent trends that helps us to decide the scope and functions of any research. Additionally, it enlightens us on areas which are a dead end and areas which should be focussed upon to publish a quality work. Here we referred the academic subject are *Software Engineering* and *Object Oriented Software Development* with course code are *CSE303* and *CSE401*. It helps us to done the right way to perform literature survey.

Chapter – III

SYSTEM ANALYSIS

3.1 EXISTING WORK

In present existing system users face many issues like frequently they have to visit a doctor even when we come across normal symptoms, illness, injury. It's a tedious task for the user to wait for the doctor's appointment long time. Generally, people are not aware of the type of drugs and medicines that are essential to use for a particular disease. Even people are not much aware about the type of diseases and illness that a person gets affected. Even for every minor reason also, we have to reach the hospital.

The healthcare predictor system is an online consultation project which is an end user support system. The proposed system allows users to get instant guidance on their health issues through an intelligent online health care system. The system is fed with various symptoms, illness associated with the user. The intelligent system allows the user to share their symptoms and issues. Once the symptoms or issues faced by the user are given, then the system processes the user's symptoms to check for different illnesses, diseases that could be associated with it. We are going to use some intelligent techniques to guess the most accurate illness that could be associated with the patient's symptoms.

In these online systems, users can get instant guidance and precautions on their health issues through an intelligent health care system by using SVM (support vector machine algorithm), Random forest algorithm, logistic regression model. In this proposed system, Admin can include new disease, symptoms/issues details into the database. Based on the type of disease and symptom, issues the data mining algorithm works. In this, Admin can view various diseases and their symptoms which are stored in the database. This system will provide appropriate guidance/precautions when the user specifies the symptoms, illness/issues encountered.

3.2 ISSUES IN EXISTING WORK

- The system is not fully automated;
- It needs doctors for full diagnosis.

- We have to go to hospital for every small reason.
- Disease diagnosis time is more.
- User can't discuss their illness to doctors immediately.
- User need to wait for doctor until doctor come.

3.3 PROPOSED WORK

The proposed system allows users to get instant guidance on their health issues through an intelligent online health care system. The system is fed with various symptoms, illness associated with the user. The intelligent system allows user to share their symptoms and issues. Once the system receives the symptoms of user it processes user's symptoms to check different illness that might be associated with patient's symptoms. To predict the accurate result, we use intelligent techniques. The intelligent model predicts the most accurate illness/disease that might be associated with patient's symptoms /issues.

Here the model is trained on labelled data to predict whether the disease is present or not. The machine learning model used here is ensemble learning model. In that we considered a random forest or random decision tree. The model is trained with various types of symptoms

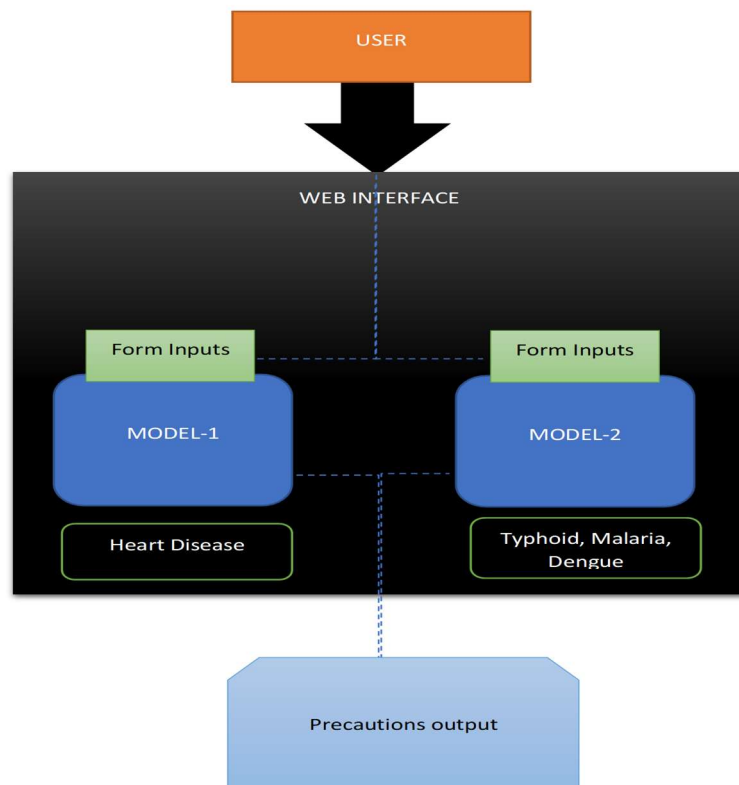


Figure 3. 1- Architecture diagram

that are associated with different diseases and used cross fold validation method to validate the model.

Coming to the User interface way the entire project is proposed in three modules. Namely – User, Admin and Doctor.

USER: In this module User can register regarding the basic details like username, password, email, phone, etc. which are mandatory to identify the disease. After that user can login into the website using username and password. If any appropriate data is encountered the system throws a printed message like incorrect password, please fill the rows with correct details etc. Once the user completes the login process the User is allowed to fill the details regarding the symptoms/issues. According to that the System will Predict the particular disease by providing with a proper precaution, instructions to follow for curing the disease. User can have different options to search for example Dengue Fever, Malaria, typhoid, heart attack. According to user symptoms the system will Predict. The disease by providing proper precautions. User have another benefit he/her can search for doctor to talk about their illness/injury and will be provided with instant diagnosis.

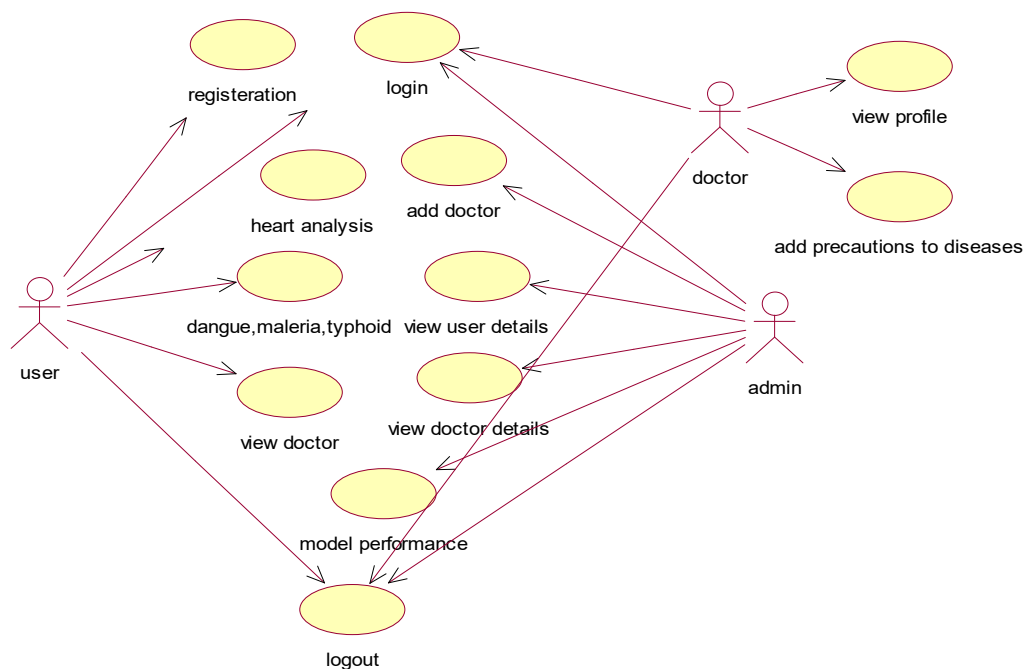


Figure 3. 2- Use Case Diagram

ADMIN: In this module admin can login into the website by providing the username and password. Admin can add information about doctor and can view user, doctor details whenever needed. Admin can Upload the Independent and Dependent Train and Test Data sets

by Selecting the different algorithms namely Random forest Algorithm, Support vector machine algorithm. This algorithm when trained on the labelled dataset gives the accuracy results.

DOCTOR: In this module doctor can login by providing details like username and password. Doctor can view the profile and can suggest the precautions necessary for a disease.

Advantages:

- Healthcare Prediction system helps in reducing the cost of medical tests as users has to spend huge amount of money on diagnostics centres.
- User can search for doctor's help at any point of time and instantly.
- They can clarify their doubts regarding the symptoms that prevails in their body in an instance.
- Disease diagnosis time is less.
- User can talk about their illness openly to machine model.
- Doctors get more clients through online.
- People will become familiar with medicines and drugs that they are consuming.
- User will be clear about the precautions for particular type of diseases.
- Users get more knowledge when they use these intelligent models.
- Users can obtain doctor's appointment easily.

Note: System Analysis is a point to clear with the idea proposing that must be comparatively have high advantageous than the existing one. Here we learned to make our idea more efficient than the existing ones. Here we referred the academic subject are Introduction to Machine Learning and Pattern Recognition, Artificial Intelligence, Internet Programming, Object Oriented Software Development and *Software Engineering* with course code are CSEX005, CSE307, CSE402, CSE401 and CSE303. These courses help us to demonstrate the proposed work and helps to understand the domain knowledge.

Chapter – IV

SYSTEM DESIGN

4.1 COMPONENTS

4.1.1 HARDWARE COMPONENTS:

Processor	- I3/Intel Processor+
RAM	- 4GB (min)
Hard Disk	- 160GB
Key Board	- Standard Windows Keyboard
Mouse	- Two or Three Button Mouse
Monitor	- SVGA

4.1.2 SOFTWARE COMPONENTS:

Operating System	- Windows 10 OS
Environment	- Anaconda 5.3
Front End	- HTML5, CSS3, Bootstrap4
Scripts	- Python 3.5+, JavaScript 1.8. 5,
Libraries	- NumPy 1.16. 0, Pandas 1.0.3, Matplotlib 3.2.1, Seaborn 0.10.1, Scikit-Learn 0.21.
Database	- My SQL 6.0

4.1.3 COMPONENTS DESCRIPTION:

Intel Core i3 Processor: Core i3 microprocessors are considered performance processors, but generally sit on the low side of the scale. It sits below the Core i5 and well below the Core i7 families, making it fairly decent for multitasking. All Core i3 processors up to are dual-core. It provides output faster and with higher quality. As this project uses Anaconda, Chrome and MySQL server upon i3 processor it really works the best and produces output instantly.

Anaconda: Anaconda is a free and open-source distribution. It supports the Python and R programming languages for scientific computing. It aims to simplify package management. Anaconda Edition With over 19+ million users worldwide. It is the

industry standard for developing, testing, and training on a single machine, enabling individual data scientists to work on it. It supports the quickly download of 7,500+ Python/R data science packages. In this project we use almost most of the machine learning tools like Scikit-Learn, Pandas, NumPy and SciPy which is pre-installed in anaconda and PyMySQL to deal with MySQL server. Anaconda also supports rich libraries for data visualization to print graphs and tables like matplotlib, seaborn and broken. In our scenario we used seaborn to set the graph and matplotlib to print the graphs and charts.

Python: Python is a general purpose and high-level programming language. You can use Python for developing desktop GUI applications, websites, Machine learning applications and web applications. Also, Python, as a high-level programming language, allows us to focus on core functionality of the application by taking care of common programming works. The syntax of python is very simple, clean and easy to learn with compare to other languages. So, in this Health care predictor we use Python as a language to implement both machine learning work and website work with the help of flask framework.

Scikit-Learn: Scikit-learn known as Sk-learn. Sk-learn an open source Python library that implements a range of machine learning works like pre-processing, cross-validation and tuning with its rich algorithms using a unified interface on the data provided and accumulated in the system to generate various models. In our project we use Pandas module to read the csv file which contains comma separated values and convert that into a pandas DataFrame. Scikit-Learn's simple and efficient algorithms helps to split the data set to test and train samples for compiling under various algorithms to improve the accuracy by opting the best supervised model. The SVM, Random forest, statistical modelling, decision tree analysis and logical regression clustering can be carried out in one place making it simple and robust for our project.

MySQL: MySQL is a Relational Database Management System (RDBMS) based on SQL – Structured Query Language. This application is used for a wide range of purposes, including data warehousing, e-commerce, and logging applications. The most common use for MySQL is for the purpose of database. Healthcare Predictor is provided with large amount of data about the diseases, symptoms, treatments, precautions, advices and so on are stored in a database to correlate, categorise, validate,

visualise and provide inputs for the machine learning models to guess the disease. This is taken care by MySQL, a free open source de-facto standard database system. It also helps us to manage the information accumulated by the end users. It is highly compatible with all operating systems to make it more user friendly and simple.

4.2 SYSTEM IMPLEMENTATION

4.2.1 MACHINE LEARNING PROCESS:

Step-1:

- Download and Install Python and Anaconda from the Browser.
- Open Anaconda prompt or anaconda navigator.
- Choose environment then open up Spyder application and Jupyter Notebook.

Step-2:

- Rename the file name as comparison.py and save in the working directory.
- Import all necessary libraries.
- Load csv files (datasets) in Data Frames using Pandas.

Step-3:

- If any possible outliers or missing values means remove them using Pandas library as a process data handling and wrangling.
- Took out the feature attributes in X data frame and Label attribute in Y DataFrame.
- If necessary, reshape the Data Frames.

Step-4:

- Split the data sets i.e., x and y Data Frames using Scikit-learn algorithm.
- Give some random value to shuffle the records similar fashion in X and Y data frames.
- Here we got four data frames named as x_train, y_train, x_test and y_test.

Step-5:

- Now load the machine learning classification algorithms.

- Call those algorithms with their objects by passing the x_train and y_train data frames.

Step-6:

- Now predict the values by passing new data set.
- Plot the confusion matrixes for algorithms.
- Calculate the accuracy by using metrics score algorithm in scikit learn for different machine learning classification algorithms.
- Plot bar graph for the accuracies.

Step-7:

- Now take the model with high accuracy.
- Tune its hyper parameters by using Machine Learning Tuning algorithm.
- Take the best set of parameters and train the model.
- Print the accuracy as the above-mentioned way.
- Now deploy that best algorithm on main file.

Step-8:

- Save and Close the script file.
- Close Spyder application and close the anaconda.

4.2.2 MySQL PROCESS:

Step-1:

- Download and Install MySQL Community Server from the Browser.
- Open MySQL workbench from command prompt or by clicking on GUI application.

Step-2:

- Create root and password to access MySQL server
- Create an environment to access the database.
- Open database connector page in MySQL workbench.

- Close the starting page or welcome page.

Step-3:

- Create a database name it as health_predictor.
- Now open new script file to perform operations like creating the tables and modifying the data in tables.

Step-4:

- Create tables for users, doctors, precautions and admins with respective data fields.
- Insert values into the tables.
- Use some procedures to work at run time like to increment the id and modifying the data.

Step-5:

- Save and Close the script file.
- Close the MySQL workbench.

4.3 DESIGN SPECIFICATION, STANDARDS AND CONSTRAINT

This project follows to the “**2755-2017** – IEEE” Guide for Terms and Concepts in Intelligent Process. This standard is intended to provide a set of definitions established by and for the community involved with *software-based intelligent process automation (SBIPA)* so that when terminology is used, all understand the meaning. Since, it is a highly interactive tool that machine learning with interface supports the best use for medical purpose. This also supports learning of the various diseases like heart disease, typhoid, malaria and dengue which supplements the medical practitioners in diagnostics purposes and give the best precautions. The project is based on few design constraints. It is economical as it does not involve any costly tools except the system configuration. This makes it cost efficient in comparison with the Real doctor appointment and diagnose as it needs several processes to do like booking an appointment, travelling to hospital etc... As the project deals with predicting the disease, where people frequently get these diseases which is a global threat to end some extent of people so, it tries to sort out the social issue by aiding the doctors in providing better perspective view of

giving precautions for predicted disease. In addition to this it also creates awareness on the diseases that people have. Thus, it considers social constraints. As the project is eco-friendly, it does no harm to the biological environment. The approach we used to implement our project is all open source free libraries. So, whenever there is new update in the features used it freely apt to it. Hence our project constraint to sustainable too.

4.4 ALTERNATIVE DESIGN

In general, Healthcare Prediction is maintained by medical professional manually. For this manual work a lot of process is followed to predict a disease. Instead we go with machine learning way to predict the disease. Our project is designed by using Python, MySQL and Anaconda. We can also use R and MATLAB to implement our project. Both are having rich libraries support of machine learning, data pre-processing, data wrangling and data visualization. As similar to python does. The environment is also same for R and Python. But the R and MATLAB are not enough developed to access database and to create websites from it. Though R and Python have same speed on machine learning tasks python have rich support of other useful frameworks which are needed in our project like Flask – a framework is to connect to the website and PyMySQL a library which is an open source and freely available to perform database related task. Moreover, python is very simple, clean and easy to learn. Python is one of the mostly worldwide used language. Therefore, these advantages make us to go with the python. Coming to the model design we use random forest approach which is an ensemble approach to take voting and give result instead of direct classification techniques like logistic regression – a probability approach, support vector machine – a hyperplane classification approach to classify the data labels and finally decision tree – dividing based on information gain and entropy. Out of these, Random forest is best to tune its parameters to give best accuracy.

Note: Implementation deals with the type of approach needed for different modules. It should include a way through which a project is set including the input, process and output design. Here we referred the academic subject are Introduction to Machine Learning and Pattern Recognition, Database Management Systems, Probability and Statistics, Data Science with R and Machine Learning Techniques with course code are CSEX005, CSE305, MAT222, CSEX008 and CSE439. These courses help us to apply the proposed work, perform mathematical operations and helps to understand the domain knowledge.

Chapter – V

RESULT AND DISCUSSIONS

DATA SET:

The data sets used in this project are collected from Kaggle website. This is the place where we can notice different types of projects, datasets, current ongoing works. There are multiple number of datasets with different information. In this we are involved in exploring the data how it is correlated with other variables in the data sets mainly with independent values. The below graph describes the correlation between attributes.

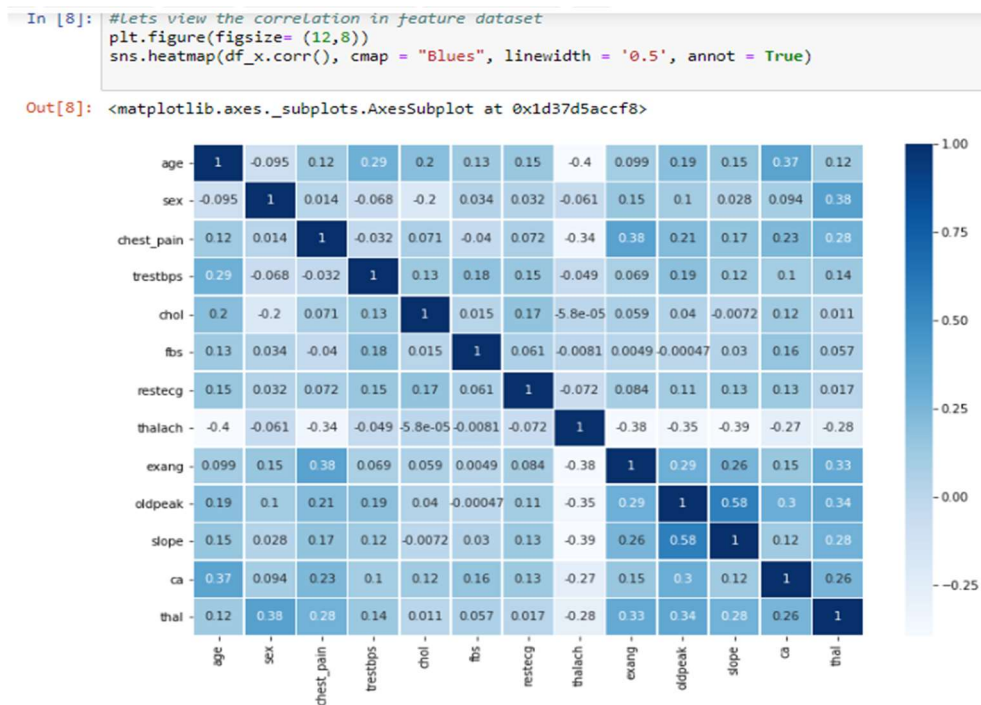


Figure 5. 1 – Correlation Graph

The above figure describes about correlation if the blue colour has more intensity then it has more correlated between x and y attributes. The diagonals are a pair of similar attributes so they are more correlated in the graph compared to others.

MODEL EVALUATION:

The dependent variable used here is categorical. So, we opt for supervised classification models. When it comes to classification, we have multiple machine learning models to deploy

```
In [54]: #lets select our model from random forest, LogisticRegression, SVC and decision tree classifier
```

```
from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.linear_model import LogisticRegression as LR
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier as DTC

from sklearn.model_selection import GridSearchCV, train_test_split, cross_val_score
from sklearn.metrics import confusion_matrix, accuracy_score
```

```
In [55]: x_train,x_test,y_train,y_test = train_test_split(df_x,df_y, test_size = 0.2)
y_test1 = y_test.copy()
```

```
In [56]: clf = RFC(n_estimators=10)
clf1 = LR()
clf2 = SVC()
clf3 = DTC()
```

```
In [57]: clf.fit(x_train,y_train)
pred = clf.predict(x_test)

clf1.fit(x_train,y_train)
pred1 = clf1.predict(x_test)

clf2.fit(x_train,y_train)
pred2 = clf2.predict(x_test)

clf3.fit(x_train,y_train)
pred3 = clf3.predict(x_test)
```

Figure 5. 2 – Models Deployment Code Snippet

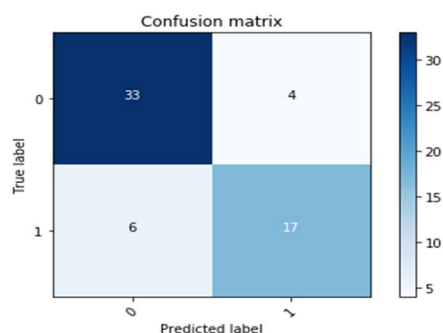
in Scikit-learn. We used logistic regression, support vector machine, decision tree and ensemble learning method namely random decision tree. Despite of all these four models we choose to use random forest which gives more accuracy on testing data set. The accuracy and the confusion matrixes are shown below in pictorial way.

The confusion matrixes that produced from different models are shown below with the help of matplotlib library.

```
In [58]: print("Random Forest")
cm = confusion_matrix(pred, y_test)

plot_confusion_matrix(cm, classes = [0,1])

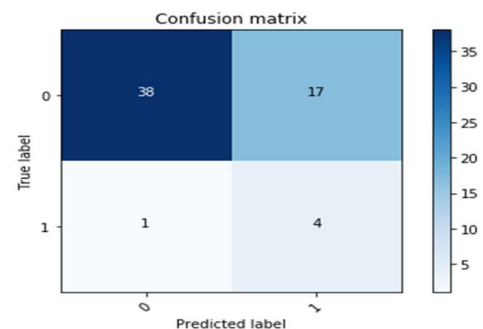
Random Forest
Confusion matrix, without normalization
```



```
In [60]: print("svm")
cm = confusion_matrix(pred2, y_test)

plot_confusion_matrix(cm, classes = [0,1])

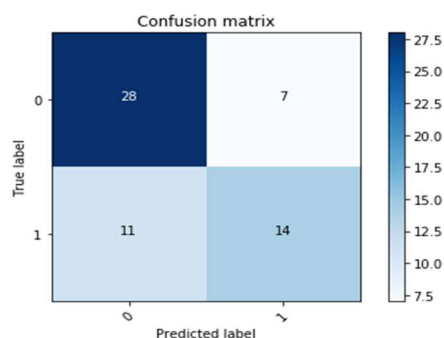
svm
Confusion matrix, without normalization
```



```
In [61]: print("Decison tree")
cm = confusion_matrix(pred3, y_test)

plot_confusion_matrix(cm, classes = [0,1])

Decison tree
Confusion matrix, without normalization
```



```
In [59]: print("logistic regression")
cm = confusion_matrix(pred1, y_test)

plot_confusion_matrix(cm, classes = [0,1])

logistic regression
Confusion matrix, without normalization
```

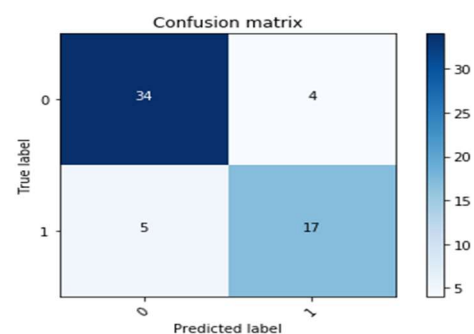


Figure 5. 3 – Confusion matrixes for 4 Models

From the above graphs, graph confusion matrix works fine with the random forest and logistic regression which is almost similar to the dataset. The accuracy is plotted in bar graph and also showed in tabular format before we start tuning the hyperparameters on the same dataset.

```
In [85]: plt.legend("Accuracy for different models")
names = ['random forest', 'Logistic regression', 'svm', 'decision tree']
plt.xticks(rotation=90)
plt.bar(pd.Series(names),pd.Series(li),color=["green","red","yellow","blue"])

Out[85]: <BarContainer object of 4 artists>
```

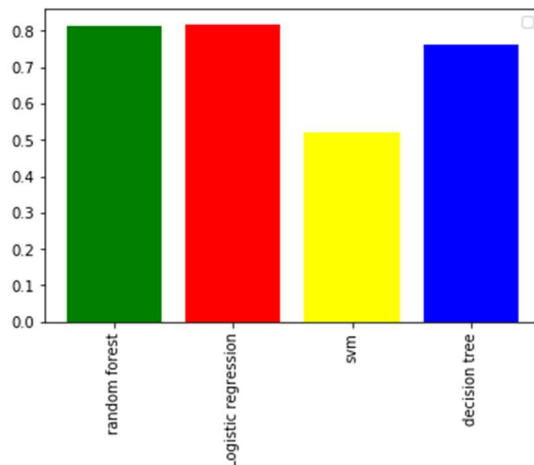


Figure 5. 4 – Accuracy Bar Graph

Hence it is clear that random forest which is indicated in green colour and logistic regression which is indicated in red colour are almost works fine for data set with similar accuracy.

Table 5. 1 – Accuracy Table

S. No.	Model Name	Accuracy score
1	Random Forest Decision Tree	82.12%
2	Logistic Regression	82.12%
3	Support Vector Machine	54.03%
4	Decision Tree	76.58%

HYPERPARAMETERS TUNING:

When we consider the random forest as our model and tunes its hyper parameters by using GridSearchCV - a algorithm in scikit-learn library that used to perform all models on data set by varying its hyperparameters that provides and gives the best accuracy model with its hyperparameters.

```
In [37]: param = {"max_depth":[5,10,15],
                 "n_estimators":[10,15,20,25,30,40,50],
                 "max_features":[2,3,4],
                 "min_samples_leaf":[3,4,5],
                 "min_samples_split":[2,3,4,5],
                 "bootstrap":[True] }

grid = GridSearchCV(estimator = clf, param_grid = param, cv = 7, verbose = 3)

grid.fit(x_train,y_train)

cv_keys = ("mean_test_score","std_test_score","params")

for k,_ in enumerate(grid.cv_results_["mean_test_score"]):
    print(f'{grid.cv_results_[cv_keys[0]][k]} , {grid.cv_results_[cv_keys[1]][k]}, {grid.cv_results_[cv_keys[2]][k]}")
```

```
In [38]: print(f"Best score : {grid.best_score}")
        print(f"Best Param : {grid.best_params}")

Best score : 0.8516949152542372
Best Param : {'bootstrap': True, 'max_depth': 5, 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 4, 'n_estimators': 25}

clf = RFC(bootstrap= True, max_depth= 5, max_features= 3, min_samples_leaf= 4, min_samples_split= 4, n_estimators=25)
clf.fit(x_train,y_train)
pred = clf.predict(x_test)

accuracy = accuracy_score(pred,y_test)

print(f"Accuracy on Test dataset = {round(accuracy,2)}")
```

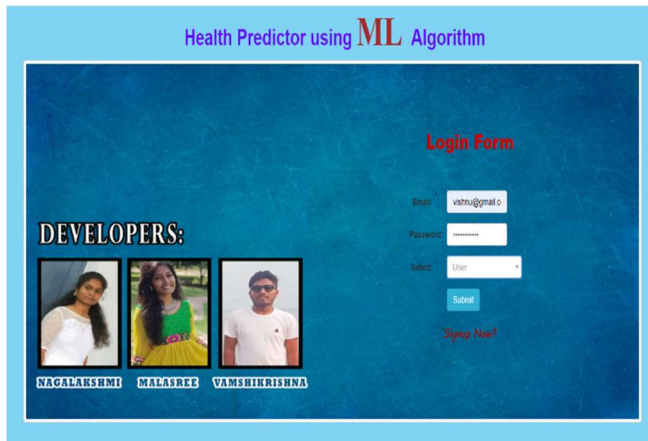
Figure 5. 5 - Hyperparameter Tuning Code Snippet

The best hyperparameters is provided by GridSearchCV are bootstrap is True, maximum depth of nodes are five, the maximum features to take are 3, minimum samples at leaf are 4, split on samples test is 4 and number of estimators to use are 25. By providing these it gives an accuracy on validation and training set is 85.169% and on test data is 83.92% accuracy.

USER INTERFACE:

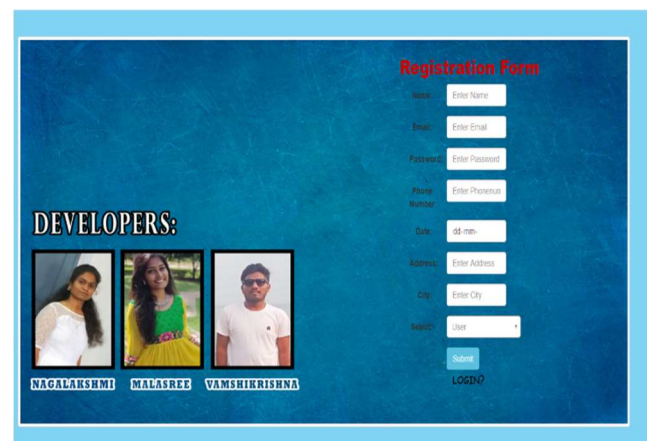
The entire project is implemented in three modules with User friendly interfaces using web technologies. The interfaces are User, Admin and Doctor.

User: For every interface there must be two things in common to provide security to the users. The login page and the sign in page.



The login page features a blue background with a heart icon. At the top, it says "Health Predictor using ML Algorithm". Below this, there's a "Login Form" with fields for Email (vstru@gmail.o), Password (masked), and a dropdown for Select (User). A "Submit" button and a "Sign Up Now?" link are at the bottom. On the left, under "DEVELOPERS:", there are three profile pictures of Nagalakshmi, Malasree, and Vamsikrishna.

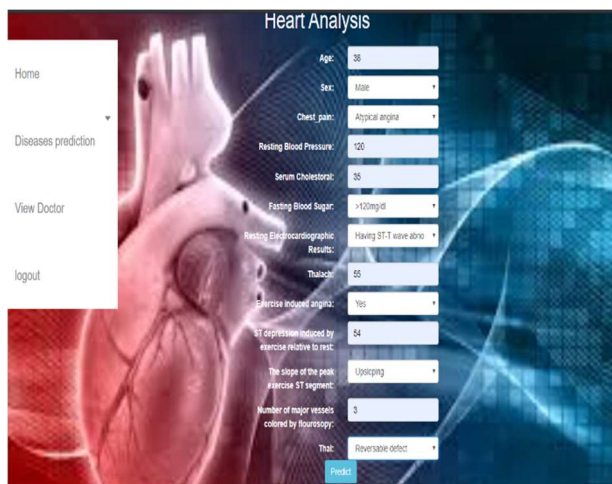
Figure 5. 6- Login Page



The registration page has a blue background with a heart icon. It features a "Registration Form" with fields for Name, Email, Password, Phone Number, Date, Address, City, and a dropdown for Select. A "Submit" button and a "Login?" link are at the bottom. On the left, under "DEVELOPERS:", there are three profile pictures of Nagalakshmi, Malasree, and Vamsikrishna.

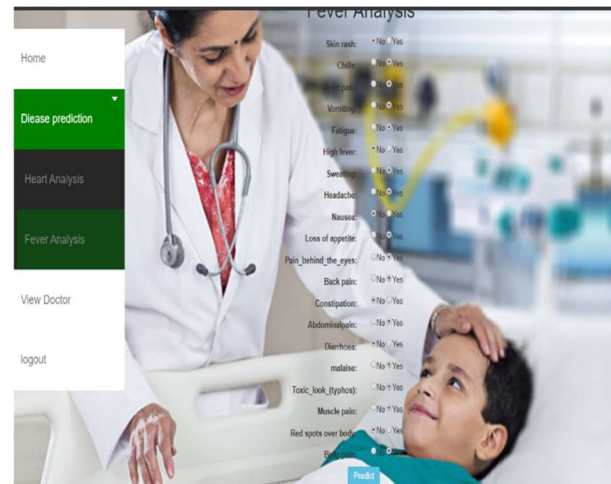
Figure 5. 7 - User Signup Page

When user enter after signup, he needs to login to enter the working modules. Where he can actually see two modules in prediction. One is for heart analysis and another is for normal health issues like dengue, typhoid, malaria.



The heart analysis page has a blue background with a heart icon. It features a "Heart Analysis" form with fields for Age, Sex, Chest pain, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting Electrocardiographic Results, Throat, Exercise induced angina, ST depression induced by exercise relative to rest, The slope of the peak exercise ST segment, Number of major vessels colored by fluoroscopy, and Thab. A "Predict" button is at the bottom. On the left, there's a sidebar with links: Home, Diseases prediction, View Doctor, and logout.

Figure 5. 8 - User Heart Prediction Page



The fever analysis page has a blue background with a heart icon. It features a "Fever Analysis" form with fields for Skin rash, Chills, Fatigue, High fever, Swollen, Headache, Nausea, Loss of appetite, Pain behind the eyes, Back pain, Constipation, Abdominal pain, Diarrhea, Malaise, Toxic look, Muscle pain, Red spots over body, and Rash. A "Predict" button is at the bottom. On the left, there's a sidebar with links: Home, Disease prediction, Heart Analysis, Fever Analysis, View Doctor, and logout.

Figure 5. 9 - User Other Disease Prediction Page

The above two diagram describes two random forest models prediction two datasets. The user needs to fill the above form for their respective field. Once they enter all fields and submit the form by clicking on predict means the model will predict as per fields data and give the precautions for the disease predicted as shown below.

Figure 5. 10 - Heart Disease Precaution Page

Figure 5. 11 - Fever Precautions Page

After prediction of disease the user can book the appointment by clicking on the button named book appointment shown after precautions to the doctors registered in our site. After pressing the button, a new page opens to book the appointment. He can cross check the doctors registered in our site and he can call to them by noting the doctor's details as shown below. After done his work he can logout from the interface by clicking on logout menu icon.

Figure 5. 12 - User Appointment Page

Name	Email	Mobile Number	Hospital name	Hospital address	City	Landmark	Age	Specialist	Gender
poornima	poornima@gmail.com	9847256071	venkateswara	trappell, 123	trappell	temple	25	cardiologist	female
Dillo	dillo@gmail.com	980774455	Apollo	ibangalore	bangalore	Near SPRT	35	CARDIOLOGY	male
Durga	durga@gmail.com	9876541234	medicare	mainup road	vasanagal	hustup	45	General	male

Figure 5. 13 - Doctor Details Page

Admin: The admin needs to login to enter to his interface. Once his credentials match with database pre-entered credentials he/she allowed to view the registered users and doctors. Here the admin has the access to add new doctors with their respective details. Below images shows the addition of new doctor to database and view of users registered. The view of doctors registered is same to user and doctor. Then he has an option to logout from interface.

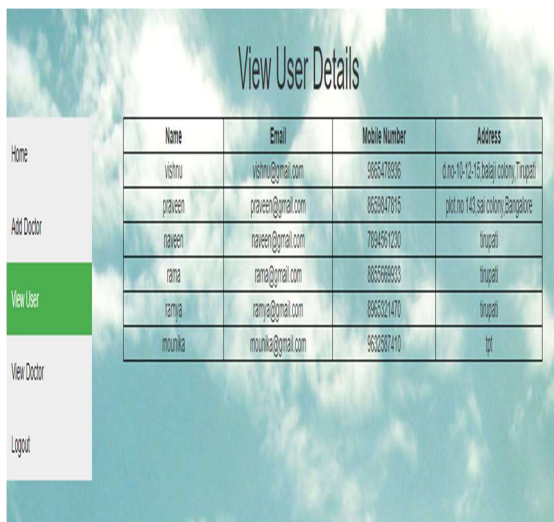


Figure 5. 15 - User Details Page

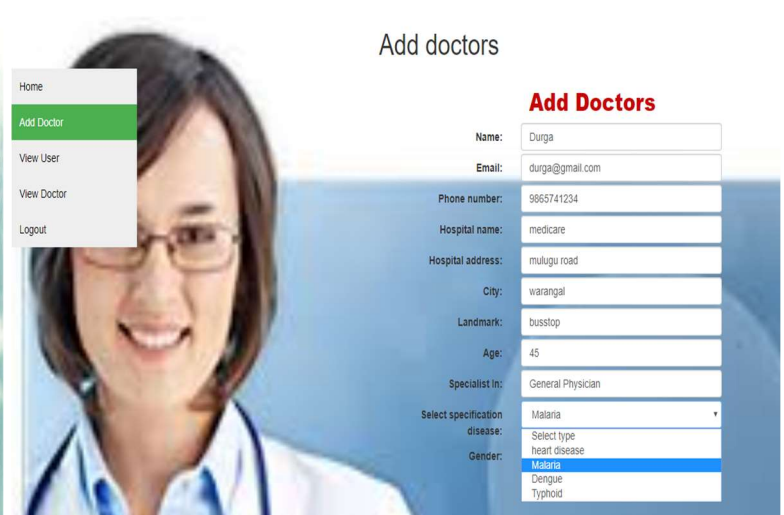


Figure 5. 16 - Doctors Adding Page

Doctor: The doctor after entering his credentials which is given by admin of site, he/she can enter into site. The doctor can cross check his appointment's and he can give precautions to the users who booked his/her appointment. Below image shows one of doctor and his users' appointments.

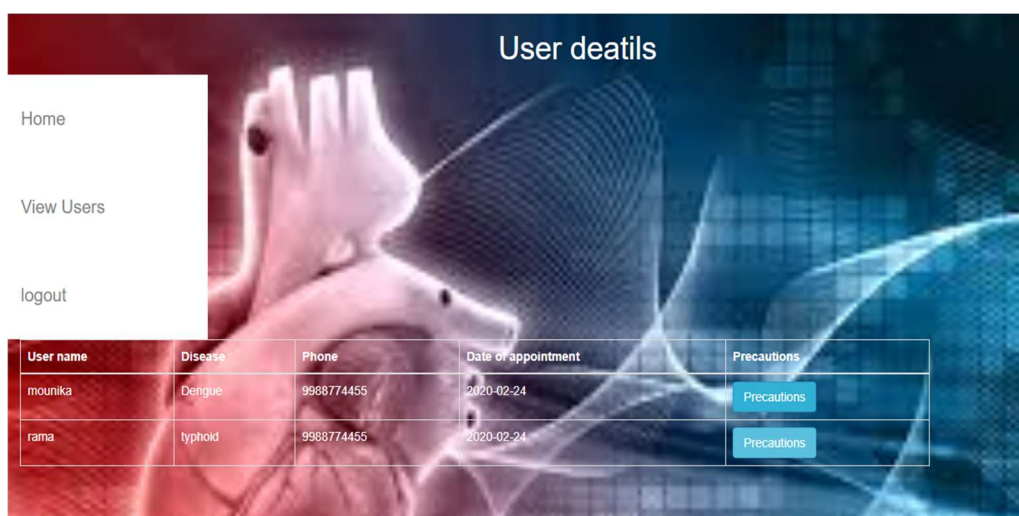


Figure 5. 17 - Doctors Appointment Page

The doctor has an option to enter precautions to the respective user by clicking on the field in table shown above. Once the doctor done with precautions, he can logout from this interface by clicking on logout button.

Note: Result Analysis is the process through which we understand the efficiency and notice the possibilities of important. Here, various datasets are correlated, different models are validated to find the best possible accuracy level and to help tune the hyper-parameters to optimise the accuracy. The final output or the front end of the web which the user interacts with the system should be designed such that it's simple, elegant, clear, consistent by contributing to ease of use and making important information obvious. Here we referred the academic subject are Programming Languages, Introduction to Machine Learning and Pattern Recognition, Artificial Intelligence, Probability and Statistics, Data Science with R, Internet Programming and Machine Learning Techniques with course code are CSE102, CSEX005, CSE307, MAT222, CSEX008, CSE402 and CSE439. These courses help us to implement the coding work in languages like Python, HTML, CSS, JavaScript

Chapter – VI

CONCLUSION AND RECOMMENDATIONS

The system would drastically reduce the human effort, and it reduces the cost, time constraint in terms of human resources and expertise by increasing the diagnostic accuracy. Generally, the prediction of disease using Data Mining applications is a challenging and risky task as the data found to be noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring the knowledge of the medical data which is quite interesting.

The current work generally helps to open up a new research area. The models that we used to predict the disease based on user symptoms can be further extended by considering different types of issues/diseases. There are many possible improvements that can be explored to improve the scalability and accuracy of the prediction system. Currently we have developed a generalized system. In future we can use these intelligent systems with different data sets containing different symptoms and diseases to train the model. The performance and the accuracy of the health care can be improved significantly by handling numerous class labels, data sets in the prediction process. This will lead to a positive impact in the field of medical research. So, when the model is trained with different datasets, we can extract powerful insights because the datasets are extremely enormous. Generally, the dimensionality of the heart disease database is high and thus the identification, selection of significant attributes for better accuracy and diagnosis of heart disease are the challenges for future research.

APPENDIX

MACHINE LEARNING CODE SNIPPET:

```
# import necessary files
import pymysql
import time
from datetime import datetime
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import itertools

from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.metrics import confusion_matrix, accuracy_score

# connected to the database to store logins information.
db = pymysql.connect(host='localhost', user='root', password='password', db='health_predictor1')
cursor = db.cursor()

...

# load data files which are pre-processed
x_train = pd.read_csv('x_train.csv').drop("Unnamed: 0", axis=1)
print(x_train)
y_train = pd.read_csv('y_train.csv').drop("Unnamed: 0", axis=1)
print(y_train)

...

# call the model and store in model object.
clf = RFC(random_state=21, bootstrap=True, max_depth=5, max_features=3,
min_samples_leaf=3, min_samples_split=2,
n_estimators=25)
clf.fit(x_train, y_train.values.ravel())
# predict the output by taking the inputs from user.
list2 = pd.DataFrame(np.array(list1).reshape(1, 13), columns=x_train.columns)
pred = clf.predict(list2)
predict = pred[0]

...
```

#hyperparameters tuning of random forest by using gridsearchCV:

```
param = {"max_depth":[5,10,15],
        "n_estimators":[10,15,20,25,30,40,50],
        "max_features":[2,3,4],
        "min_samples_leaf":[3,4,5],
        "min_samples_split":[2,3,4,5],
        "bootstrap":[True]}

grid = GridSearchCV(estimator = clf, param_grid = param, cv = 7, verbose = 3)
grid.fit(x_train,y_train)

cv_keys = ("mean_test_score","std_test_score","params")
for k,_ in enumerate(grid.cv_results_["mean_test_score"]):
    print(f'{grid.cv_results_[cv_keys[0]][k]} , {grid.cv_results_[cv_keys[1]][k]},
    {grid.cv_results_[cv_keys[2]][k]}")
```

#printing best hyperparameters with good accuracy

```
print(f'Best score : {grid.best_score_}')
print(f'Best Param : {grid.best_params_}')
accuracy = accuracy_score(pred,y_test)
print(f'Accuracy on Test dataset = {round(accuracy,2)}')
...
```

#defining function for confusion matrix

```
def plot_confusion_matrix(cm, classes, normalize=False, title='Confusion matrix',
cmap=plt.cm.Blues):

    plt.imshow(cm, interpolation='nearest', cmap=cmap)

    plt.title(title)

    plt.colorbar()

    tick_marks = np.arange(len(classes))

    plt.xticks(tick_marks, classes, rotation=45)

    plt.yticks(tick_marks, classes)

    if normalize:
```

```

    cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
    print("Normalized confusion matrix")
else:
    print('Confusion matrix, without normalization')
thresh = cm.max() / 2.
for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
    plt.text(j, i, cm[i, j],
             horizontalalignment="center",
             color="white" if cm[i, j] > thresh else "black")
plt.tight_layout()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
...

#calling confusion matrix
cm = confusion_matrix(pred,y_test,labels)
plot_confusion_matrix(cm,classes =[1,0] )
...

```

MYSQL CODE SNIPPET:

```

-- Table structure for table `doctor_reg`
CREATE TABLE `doctor_reg` (
  `id` int(11) NOT NULL,
  `name` varchar(50) NOT NULL,
  `email` varchar(50) NOT NULL,
  `password` varchar(15) NOT NULL,
  `phnumber` varchar(50) NOT NULL,
  `hname` varchar(50) NOT NULL,
  `haddress` varchar(50) NOT NULL,
  `city` varchar(50) NOT NULL,

```



```

`landmark` varchar(50) NOT NULL,
`age` varchar(50) NOT NULL,
`specailist` varchar(50) NOT NULL,
`specification` varchar(50) NOT NULL,
`gender` varchar(50) NOT NULL,
`date_of_reg` varchar(50) NOT NULL,
`image` varchar(50) NOT NULL
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

--

-- Dumping data for table `doctor_reg`
INSERT INTO `doctor_reg` (`id`, `name`, `email`, `password`, `phnumber`, `hname`, `haddress`,
`city`, `landmark`, `age`, `specailist`, `specification`, `gender`, `date_of_reg`, `image`) VALUES
(2, 'poornima', 'poornima@gmail.com', 'doctor@1234', '4561239871', 'venkaateswara', 'tirupati, 123',
'tirupati', 'temple', '29', 'cardiologist', '', 'female', '2019-03-12 11:50:28.071682', ''),
(3, 'Dilip', 'dilip@gmail.com', 'doctor@1234', '9988774455', 'Apollo', 'bangalore', 'bangalore', 'Near
GRT', '35', 'CARDIOLOGY', 'heart disease', 'male', '2020-02-24 14:22:17.934436', ''),
(4, 'Durga', 'durga@gmail.com', 'doctor@1234', '9876541234', 'medicare', 'mulugu road', 'warangal',
'busstop', '45', 'General', 'malaria', 'male', '2020-02-24 14:22:17.934436', '');

--

-- Table structure for table `paitent_precautions`
CREATE TABLE `paitent_precautions` (
`id` int(15) NOT NULL,
`username` varchar(50) NOT NULL,
`disease` varchar(20) NOT NULL,
`precautions` varchar(1000) NOT NULL,
`patient_a_id` int(10) NOT NULL,
`user_email` varchar(50) NOT NULL
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

--

```

-- Dumping data for table `paitent_precautions`

```
INSERT INTO `paitent_precautions` (`id`, `username`, `disease`, `precautions`, `patient_a_id`, `user_email`) VALUES
```

```
(1, 'user', 'Dengue', 'r\n  hi good', 42, 'user1234@gmail.com'),
```

```
(2, 'user', 'Dengue', 'r\n  take a full bed rest\r\n', 42, 'user1234@gmail.com'),
```

```
(3, 'user', 'Dengue', 'r\n  health is good', 42, 'user1234@gmail.com');
```

--

-- Table structure for table `reg`

```
CREATE TABLE `reg` (
```

```
  `reg_id` varchar(50) NOT NULL,
```

```
  `username` varchar(50) NOT NULL,
```

```
  `email` varchar(50) NOT NULL,
```

```
  `password` varchar(50) NOT NULL,
```

```
  `phone_number` varchar(50) NOT NULL,
```

```
  `dob` varchar(50) NOT NULL,
```

```
  `address` varchar(50) NOT NULL,
```

```
  `city` varchar(50) NOT NULL,
```

```
  `user_type` varchar(50) NOT NULL,
```

```
  `reg_timedate` varchar(50) NOT NULL
```

```
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

--

-- Dumping data for table `reg`

```
INSERT INTO `reg` (`reg_id`, `username`, `email`, `password`, `phone_number`, `dob`, `address`, `city`, `user_type`, `reg_timedate`) VALUES
```

```
('Admin1546577899017', 'vamshikrishna', 'krishnavamshi.12321@gmail.com', 'vk@1234', '8919765647', '11/18/1993', 'H.no 3-13/b, warangal', 'Telangana', 'admin', 'Fri Jan 4 10:28:19 2019'),
```

```
('User1546578023622', 'vishnu', 'vishnu@gmail.com', 'vishnu@1234', '9865478936', '9/22/1994', 'd.no-10-12-15,balaji colony,Tirupati', 'Tirupati', 'user', 'Fri Jan 4 10:30:23 2019'),
```

```
('User1546578117244', 'praveen', 'praveen@gmail.com', 'hari@1234', '8659847815', '10/6/1992', 'plot.no 143,sai colony,Bangalore', 'Bangalore', 'user', 'Fri Jan 4 10:31:57 2019'),
```

```
('User1561696148372', 'naveen', 'naveen@gmail.com', 'naveen@1234', '7894561230', '2019-03-27', 'tirupati', 'tirupati', 'user', 'Fri Jun 28 09:59:08 2019'),
```

```
(('User1561781245697', 'rama', 'rama@gmail.com', 'rama@1234', '8855669933', '2019-06-30',
'tirupati', 'tirupati', 'user', 'Sat Jun 29 09:37:25 2019'),

('User1566810473558', 'ramya', 'ramya@gmail.com', 'ramya@1234', '8965321470', '2019-08-26',
'tirupati', 'tirupati', 'user', 'Mon Aug 26 14:37:53 2019'),

('Admin1569215765192', 'malasree', 'malasree@gmail.com', 'malasree@1234', '8523697410', '2019-
09-23', 'tpt', 'tpt', 'admin', 'Mon Sep 23 10:46:05 2019'),

('Admin1569215765193', 'nagalakshmi', 'nagalakshmi@gmail.com', 'nagalakshmi@1234',
'8524597420', '2019-09-23', 'anathapur', 'anathapur', 'admin', 'Mon Sep 23 10:46:05 2019'),

('User1582374948659', 'mounika', 'mounika@gmail.com', 'mounika@123', '9632587410', '2020-02-
22', 'tpt', 'tpt', 'user', 'Sat Feb 22 18:05:48 2020');
```

--

-- Table structure for table `user_appointment`

```
CREATE TABLE `user_appointment` (
  `a_id` int(11) NOT NULL,
  `doctor_name` varchar(50) NOT NULL,
  `doctor_email` varchar(50) NOT NULL,
  `user_name` varchar(50) NOT NULL,
  `disease` varchar(50) NOT NULL,
  `pnumber` varchar(50) NOT NULL,
  `date_of_appointment` varchar(50) NOT NULL,
  `user_email` varchar(50) NOT NULL
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

--

-- Dumping data for table `user_appointment`

```
INSERT INTO `user_appointment` (`a_id`, `doctor_name`, `doctor_email`, `user_name`, `disease`,
`pnumber`, `date_of_appointment`, `user_email`) VALUES
(1, 'poornima', '', 'vishnu', 'heart', '9632587410', '2019-03-09', ''),
(2, 'poornima', '', 'ramya', 'heart', '9632587410', '2019-03-09', ''),
(3, 'Durga', '', 'praveen', 'malaria', '9632587410', '2019-03-11', ''),
(4, 'Durga', '', 'naveen', 'dengue', '9632587410', '2019-03-11', ''),
(5, 'Dilip', '', 'naveen', 'dengue', '1238', '2019-03-21', ''),
(6, 'Durga', '', 'mounika', 'heart-disease', '1238', '', ''),
```

```
(7, 'Dilip', 'dilip@gmail.com', 'mounika', 'Dengue', '9988774455', '2020-02-24',  
'user1234@gmail.com'),  
(8, 'Dilip', 'dilip@gmail.com', 'rama', 'typhoid', '9988774455', '2020-02-24', 'user1234@gmail.com');  
--
```

Note: For every conundrum that arises, we have a logical solution. Likewise, our system targets the information/symptoms that's given as input by users to simulate various logic tress with our algorithm to propose the possible disease that he/she might possess. So, the best algorithm is the one which can accommodate various datasets to simulate different decision trees that guesses the disease more accurately. Here we referred the academic subject are Programming Languages, Introduction to Machine Learning and Pattern Recognition, Artificial Intelligence, Data Science with R, Data Structures, Database Management Systems, Object Oriented Programming and Machine Learning Techniques with course code are CSE102, CSEX005, CSE307, CSEX008, CSE103, CSE305, CSE206 and *CSE439*. These courses help us to implement the coding work in languages like Python and MySQL

REFERENCES

- [1] 2007 International Conference on Convergence Information Technology (ICCIT 2007) - Yanwei Xing, Jie Wang and Zhihong Zhao, “Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease”.
- [2] Machine Learning and Cybernetics, 2009 International Conference on, Volume: 3 - SAM CHAO, FAI WONG, “An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining”.
- [3] 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22 -M. Ilayaraja, T. Meyyappan, “Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm”.
- [4] Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC 2009, Chengdu, China, 12-14 December, 2009 - My Chau Tu AND Dongil Shin, “A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms”.
- [5] 9th International Conference on Information Technology (ICIT'06) - Ranjit Abraham, Jay B.Simha, Iyengar, “A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier”.
- [6] Shaikh Abdul Hannan, A.V. Mane, R. R. Manza, and R. J. Ramteke, Dec 2010, “Prediction of Heart Disease Medical Prescription using Radial Basis Function”, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .
- [7] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010, “Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network”, International Conference on Computer and Communication Technology (ICCCT), DOI:10.1109/ICCCT.2010.5640377, 17-19.
- [8] http://www.heart.org/HEARTORG/Conditions/HeartAttack/WarningSignsofaHeartAttack/Warning-Signs-of-aHeartAttack_UCM_002039_Article.jsp#.WNpKgPI97IU.
- [9] www.who.int/cardiovascular_diseases/en/.
- [10] <http://food.ndtv.com/health/world-heart-day-2015-heart-disease-in-india-is-agrowing-concern-ansari-1224160>.
- [11] https://en.wikipedia.org/wiki/Cardiovascular_disease.

- [12] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, “HDPS: Heart Disease Prediction System”,Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.
- [13] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, “Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map”,IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.
- [14] "Health Topics: Health Systems". www.who.int. WHO World Health Organization. Retrieved 2013-11-24.
- [15] "Health at a Glance 2013 - OECD Indicators" (PDF). OECD. 2013-11-21. pp. 5, 39, 46, 48. Retrieved 2013-11-24.
- [16] "OECD.StatExtracts, Health, Health Status, Life expectancy, Total population at birth, 2011" (online statistics). stats.oecd.org/. OECD's iLibrary. 2013. Retrieved 2013-11-24.
- [17] World Health Organization. Anniversary of smallpox eradication. Geneva, 18 June 2010.
- [18] United States Department of Labor. Employment and Training Administration: Health care. Retrieved June 24, 2011.