Health Prediction Analysis using Data Mining

Submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Engineering

by

Miss. Rane Kritika Ashok

Roll No. 28

Mr. Salve Ashish Ravindra

Roll No. 30

Miss. Sawant Ashwini Dhananjay

Roll No. 31

Under the Supervision of

Prof. J. P. Patil

DEPARTMENT OF INFORMATION TECHNOLOGY

KONKAN GYANPEETH COLLEGE OF ENGINEERING

KARJAT-410201

April 2019

Certificate

This is to certify that the project entitled **Health Prediction Analysis using Data Mining** is a bonafide work of**Miss. Rane Kritika Ashok (Roll No. 30), Mr. Salve Ashish Ravindra (Roll No. 33), Miss. Sawant Ashwini Dhananjay (Roll No. 34)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Undergraduate** in **Department Of Information Technology**.

Prof. J. P. Patil
Guide
Department of Information Technology

Prof. J. P. Patil
Head of Department
Department of Information Technology

Dr. Madhukar J. Lengare
Principal
Konkan Gyanpeeth College of Engineering

Project Report Approval for B.E.

This thesis / dissertation/project report entitled **Health Prediction Analysis using Data Mining** by **Miss. Rane Kritika Ashok (Roll No. 30), Mr. Salve Ashish Ravindra (Roll No. 33), Miss. Sawant Ashwini Dhananjay (Roll No. 34)** is approved for the degree of **Department of Information Technology**.

Examiners

1.....

2.....

Date.

Place.

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/-source in Our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature
Rane Kritika Ashok
(Roll No. 30)

Signature
Salve Ashish Ravindra
(Roll No. 33)

Signature
Sawant Ashwini Dhananjay
(Roll No. 34)

Date.

Abstract

As we know that health care industry is completely based on assumptions, which after get tested and verified via various tests and patient have to be depend on the doctors knowledge on that topic . so we made a system that uses data mining techniques to predict the health of a person based on various medical test results. so we can predict the health of that person based on that analysis performed by the system.The system currently design only for heart issues, for that we had used Statlog (Heart) Data Set from UCI Machine Learning Repository it includes attributes like age, sex, chest pain type, cholesterol, sugar, outcomes,etc.for training the system. we only need to passed few general inputs in order to generate the prediction and the prediction results from all algorithms are they merged together by calculating there mean value that value shows the actual outcome of the prediction process which entirely works in background .

Acknowledgements

Success is nourished under the combination of perfect guidance, care blessing. Acknowledgement is the best way to convey. We express deep sense of gratitude brightness to the outstanding permutations associated with success. Last few years spend in this estimated institution has molded us into confident and aspiring Engineers.

We express our sense of gratitude towards our project guide Prof. J. P. Patil. It is because of his valuable guidance, analytical approach and encouragement that we could learn and work on the project. We will always cherish the great experience to work under their enthusiastic guidance.

We are also grateful to our principle Dr. M.J. Lengare who not only supporting us in our project but has also encouraging for every creative activity.

We extend our special thanks to all teaching and non-teaching staff, friends and well-wishers who directly or indirectly contributing for the success of our maiden mission. Finally, how can we forget our parents whose loving support and faith in us remains our prime source of inspiration.

Lastly we would like to thank all those who directly and indirectly helping to complete this project. We would also like to acknowledge with much appreciation the crucial role of the staff of Information Technology Department, who gave the permission to use the all required software/hardware and the necessary resources to completing to the project.

Contents

List of Figures

List of Tables

Abbreviations

KDD	**K**nowlage **D**iscovry in**D**atabase
EHR	**E**lectronic **H**ealth **R**ecord
KNN	**K** **N**earest **N**eighbors
DSS	**D**ecision **S**uport **S**ystem

Chapter 1

INTRODUCTION

1.1 Introduction

There was a time when data were not readily available. As data become more abundant, however limitations in computational capabilities prevented the practical application of mathematical models. Consequently, data mining tools are now being used for clinical data. The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc. Data mining is a process of selecting, exlporing and modelling large amounts of data. This process has become an increasingly pervasive activity in all areas medical science research. Data mining has resulted in tha discovery of useful hidden patterns from massive databases. By using data mining techniques finally physicians need to know how quickly identify and diagnose potential cases.

1.2 Data Mining

Data mining is a process of selecting, exploring and modelling large amounts of data. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. The major steps involved in a data mining process are:

Extract, transform and load data into a data warehouse.

Store and manage data in a multidimensional databases.

Provide data access to business analysts using application software.

Present analysed data in easily understandable forms, such as graphs.

KDD Process

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

Major KDD application areas include marketing, healthcare sector, fraud detection, telecommunication and manufacturing.

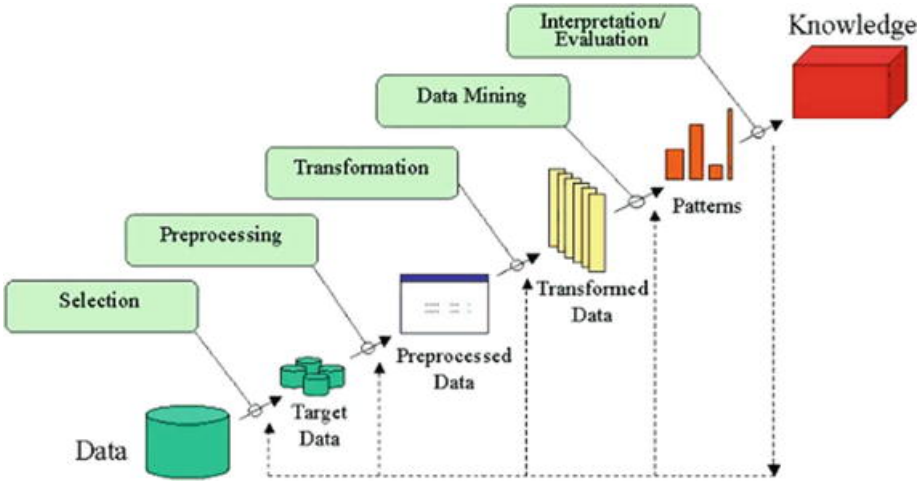


FIGURE 1.1: KDD Process

- Data Cleaning: Remove noise and inconsistent data Data Integration: Where multiple data sources may be combined. Data Selection: Where the data relevant to the analysis task are retrieved from the database.

- Data Transformation: Where data are transfoemed and consolidated into forms appropriated for mining abd performing summery or aggreagation operation.
- Data Mining: An essential process where intelligent methods are applied to extract data patterns.
- Pattern Evaluation: to identify truly interesting patterns representing knowledge based on interesting measure.
- Knowledge Representation: Where visualization and knowledge representation technique are used to present mined knowledge to users.

What Is The Use Of Data Mining In Healthcare?

The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc.

1.3 Objectives

I. Predictive Analytics And Preventive Measures Prevention is always better than cure. For the healthcare industry, it also happens to save a lot of money.
II. The Ultimate EHR System One of the biggest dreams of all is a fully digital and unprecedented comprehensive electronic health record (EHR).
III. Disease Modelling and Mapping One of the flashiest uses of data science in the past few years has been in tracking (and finding ways to halt or prevent) diseases.
IV. Reduce Fraud And Enhance Security Some studies have shown that this particular industry is 200% more likely to experience data breaches than any other industry. The reason is simple: personal data is extremely valuable and profitable on the black markets. And any breach would have dramatic consequences.
V.Personalized Medicine the pharmaceutical companies can make combined medicine based on health analysis performed by the system

1.4 Motivation

Rural Areas get relatively less healthcare facilities and also doctor availability is very poor but there might be some people who have relatively sufficient knowledge about pharmaceuticals and they can treat the patient in urgent basis by understanding what may happen to the patient. So we need something that predict what is happened to the patient in less time so we can save patient's life.So our main motto of doing this project is to help the needy people who did not get proper medical attention in short time.

1.5 Purpose, Scope, and Applicability

Purpose, Scope and Applicability: The description of Purpose, Scope, and Applicability are given below:

1.5.1 Purpose

There's a huge need for Data Mining in healthcare as well, due to rising costs in nations like the United States. As a McKinsey report [4] states, "After more than 20 years of steady increases, healthcare expenses now represent 17.6 percent of GDP —nearly 600 billion more than the expected benchmark for a nation of the United States size and wealth."
In other words, costs are much higher than they should be, and they have been rising for the past 20 years. Clearly, we need some smart, data-driven thinking in this area. And current incentives are changing as well: many insurance companies are switching from fee-for-service plans (which reward using expensive and sometimes unnecessary treatments and treating large amounts of patients quickly) to plans that prioritize patient outcomes

1.5.2 Scope

- **Hospital management**

  Data mining has been used very successfully in aiding the prevention and early detection of medical insurance fraud. The ability to detect anomalous behavior based

on purchase, usage and other transactional behavior information has made data mining a key tool in variety of organizations to detect fraudulent claims, inappropriate prescriptions and other abnormal behavioral patterns. Another key area where data mining based fraud detection is useful is detection and prediction of faults in medical devices.

- **Healthcare management** Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amounts of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making

- **Pharmaceutical industry** Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular treatment option Based on patients' profile, history, physical examination, diagnosis and utilizing previous treatment patterns, new treatment plans can be effectively suggested

- **Personalized treatment planning** Healthcare organizations make customer relationship management decisions, Physicians identify effective treatments and best practices, and Patients receive better and more affordable healthcare services

- **Prediction of diseases** Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular treatment option Based on patients' profile, history, physical examination, diagnosis and utilizing previous treatment patterns, new treatment plans can be effectively suggested .

- **Monitor patient's vital signs and many more....**

1.5.3 Applicability

"Health administration or healthcare administration is the field relating to leadership, management, and administration of hospitals, hospital networks, and health care systems."* It is actually a broad area that could encompass:

Healthcare Informatics –

Medical Device Industry

Pharmaceutical Industry –

Hospital Management –

System Biology and many more. . . .

1.6 Organisation of Report

In this section we summarize to contains of all chapters present in this report.
In introduction we studied about basic over view of the project in which we learned various outcomes of the project. how the project is going to the implemented what are the request require to develop the project etc.
In literature survey we will see various presents system the which are previously developed by various experts and we had compared system which are previously design.
in survey of technologies we will see various technologies which can be use to complete the system.
In requirements and analysis we will analyse various requirements pre-quest for developing the system also see various functional requirements software requirements and hardware requirements of the system.
In system design we see basic overall design of the system.
In the implementation chapter we will see how the system is implemented.This chapter is also contain implementation strategies and test strategies
In Results we will see test results which we have been obtained by testing various algorithms and methods.
In final chapters we will see a conclusion of the project it also describes limitations and future scope of the system.

Chapter 2

LITERATURE SURVEY

In this chapter we study about previous research done on the Health Prediction Analysis using Data Miningin this we had studied about following papers published by some experts

2.1 Data mining application in health care sector[01]

Author Name:- M.Summanth

The author was aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of disease data mining applications but it is challenging.It need human efforts to improve accuracy.

Application in healthcare sector:-

1. Treatment effectiveness
2. Healthcare management
3. Customer relationship management
4. Medical device industry
5. Pharmaceutical industry
6. Hospital management
7. System biology

2.2 Hybrid Approach for Heart Disease Detection Using Clustering and ANN [02]

Author Name:-Tejasweeta Dixit ,reshmi gore,prerana akade

Heart disease is dominant caused death in developed countries and main contributors to disease strain in develop countries. Data mining the extraction of hidden predictive from large database is powerful new technology with great probability to help companies focus on the most important in info in their data warehouse. By using various data mining techniques such as clustering algorithm,decision tree,classifiers, we can predict heart diseases.

2.3 Application of big data in medical science brings in revolution in managing health care of humans [03]

Author Name:-Dr.Gagandeep Jagdev,Sukhpreet Singh

By enabling research to identify compounds with higher likelihood of success big data can help reduce the cost and the time to market for diagnosing the diseased.

Role played by Big Data:-

1. Personalised treatment planning
2. Assisted diagnosis
3. Fraud direction
4. Monitor patient vital signs
5. Digitisation of data

Issues in Big Data:-

1. Security
2. Authorization
3. Non repudiation

2.4 Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset[05]

Author Name:- Tapas Ranjan Baitharua , Subhendu Kumar Panib

As In the paper , The authors conducted various experiment on Liver Disorder Dataset by using WEKA Tool and they had found the impact of liver disorder on the predictive performance of different classifiers. after they analyzed quantitative data they found that we find that the general concept of improved predictive performance of all above classifiers but Naive Bayes performance is not significant.

2.5 Comparison

Characteristics	Paper 1	Paper 2	Paper3
Title	Application of big data in medical science Brings revolution in managing health Care of humans	Data mining applications in healthcare sector	Hybrid approach for heart disease detection using clustering and ANN
Domain	Big data	Data mining	Clustering and ANN
Task Performed	• Personalized treatment planning • Assisted diagnosis • Fraud detection • Monitor patient vital signs • Digitization of data	• Treatment management • Healthcare management • Customer relationship management • Fraud and abuse • Medical device industry • Pharmaceutical industry • System biology • Hospital management	Prediction Of heart Disease
Technology Used	• First stage: mapping • Intermediate stages: shuffling • Final stage: reducing	--na--	• Clustering • *Neural networks navie bayes* • *Decision tree* • *K- nearest neighbor*
Provide Information About	• Various disease • Patient treatment • Hospital management info	• Patient treatment	Heart disease
Descriptive Or Predictive	Descriptive and predictive	Predictive	Predictive
Objectives	By enabling researchers to identify Compounds with a higher likelihood of success, big data Can help reduce the cost and the time to market for new Drugs. Also, by integrating learning from medical data in the Early stages of development, researchers will now be able to Customize drugs to suit aggregated patient profiles. Currently, information privacy concerns are the	The prediction of diseases using data Mining applications is a challenging task but it drastically Reduces the human effort and increases the diagnostic Accuracy. Developing efficient data mining tools for an Application could reduce the cost and time constraint in Terms of human resources and expertise.	The overall objective of this paper is to study different data Mining techniques available to predict the heart disease and to Compare them to find the best method of prediction. We can Use naive bayes, decision tree, classification algorithms to Detect heart disease, but as these are having some of the Limitations we prefer to use ann and k-means algorithms.

FIGURE 2.1: Comparison Table

Chapter 3

SURVEY OF METHODOLOGIES

In this chapter we study various Methodologies which we had used to implement project following are the methodologies with its descriptions, requirements working principles formulae and example.

3.1 Decision tree

3.1.1 Introduction

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

An example of a decision tree can be explained using above binary tree. Let's say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a lot of pizzas'? And the leaves, which are outcomes like either 'fit', or 'unfit'. In this case this was a binary classification problem (a yes no type problem).

3.1.2 Types of Decision Trees:

Classification trees (Yes/No types) What we've seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.

Regression trees (Continuous data types) Here the decision or the outcome variable is Continuous, e.g. a number like 123.

3.1.3 Working

Now that we know what a Decision Tree is, we'll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called as ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3.

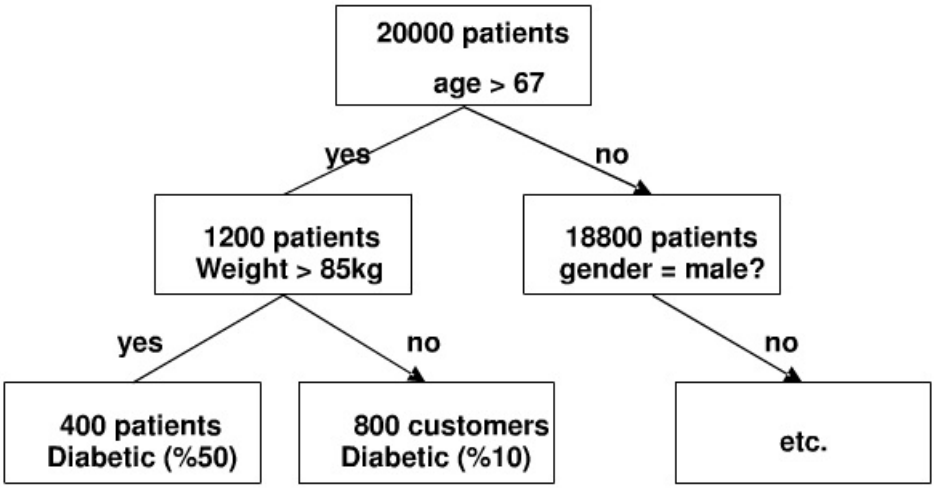


FIGURE 3.1: Dtree Example

3.2 Clustering

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. A

cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

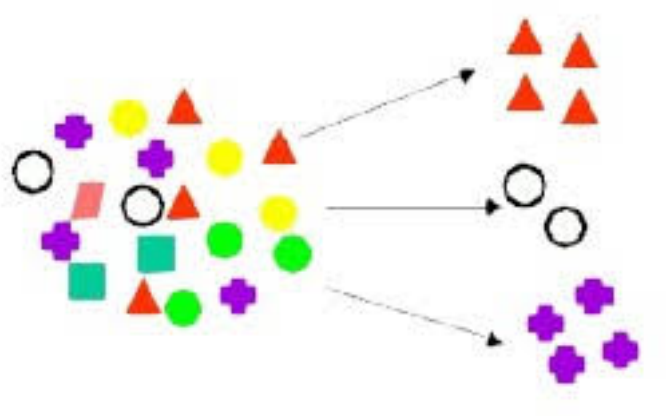


FIGURE 3.2: Clustering Example

3.2.1 K nearest neighbors Algorithm

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

3.2.2 Algorithm

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1

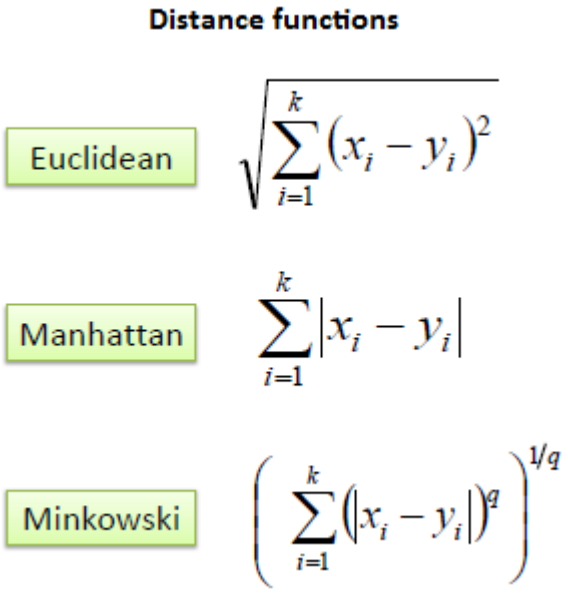


FIGURE 3.3: Distance Functions

when there is a mixture of numerical and categorical variables in the dataset. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN

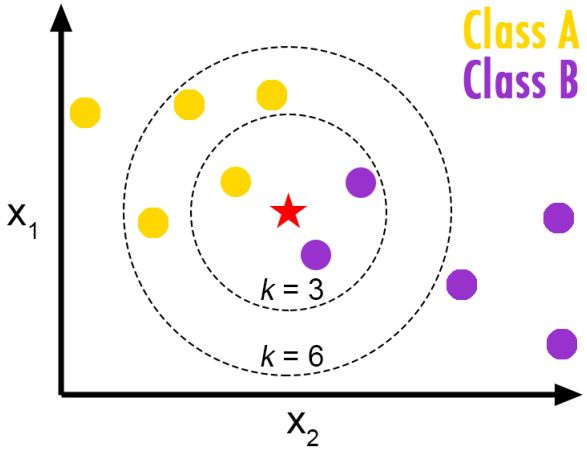


FIGURE 3.4: Example Of KNN

3.3 Logistic Regression

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

• A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)

• Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.
On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

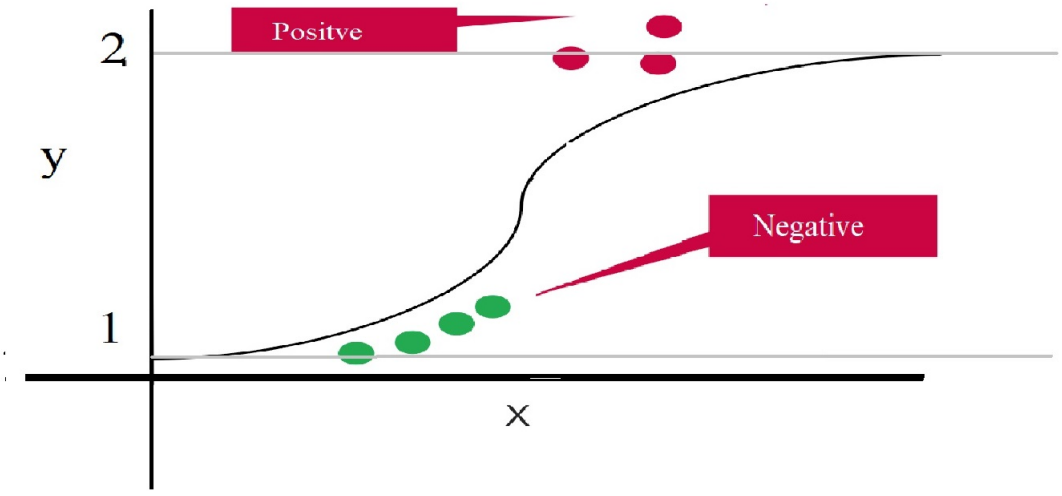


FIGURE 3.5: Logistic Regration

3.4 Bayes Theorem

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) is a theorem with two distinct interpretations. In the Bayesian interpretation, it expresses how a subjective degree of belief should rationally change to account for evidence. In the frequentist interpretation, it relates inverse representations of the probabilities concerning two events. In the Bayesian interpretation, Bayes' theorem is fundamental to Bayesian statistics, and has applications in fields including science, engineering, medicine and law. The application of Bayes' theorem to update beliefs is called Bayesian inference.

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

FIGURE 3.6: Bayes Theorem

Bayes' theorem is named for Thomas Bayes , who first suggested using the theorem to update beliefs. However, his work was published posthumously. His ideas gained limited exposure until they were independently rediscovered and further developed by Laplace, who first published the modern formulation in his 1812 Théorie analytique des probabilités. Until the second half of the 20th century, the Bayesian interpretation attracted widespread dissent from the mathematics community who generally held frequentist views, rejecting Bayesianism as unscientific. However, it is now widely accepted. This may have been due to the development of computing, which enabled the successful application of Bayesianism to many complex problems.

Chapter 4

REQUIREMENTS AND ANALYSIS

In this chapter we study about various requirements for the system for working and predicting results

4.1 Problem Definition

persons health is a critical factor which does not treated by an non professional if the professional is non available then the treatment can not be done. medical reports where to complex to understand by normal people so normal people does not able to interpret what is happening with the persons health.Health care industries is a huge business so it may happen that the doctors intentionally miss leads the patients in order to make more earnings.so we need some kind of an automatic system to avoid such problems.

4.2 Requirements Specification

we required lots of data like persons personal information, emergency contact information, email address, allergies ,height ,weight etc.for creating EHR.for prediction we also required health related reports containing blood suger level,serum cholestoral in mg/dl,thal,number

of major vessels effected ,maximum heart rate achieved ,resting electrocardiographic results in values 0,1,2 etc.

4.3 Planning and Scheduling

Planning and scheduling is a complicated part of software development. Planning, for our purposes, can be thought of as determining all the small tasks that must be carried out in order to accomplish the goal. Planning also takes into account, rules, known as constraints, which, control when certain tasks can or cannot happen. Scheduling can be thought of as determining whether adequate resources are available to carry out the plan. we had created an plan to complete this project in time.in that plan we had activities and their schedule.This plan helps us to monitor our project work.we represented that plan in gantt charts as shown bellow

FIGURE 4.1: Plan For Stage 1

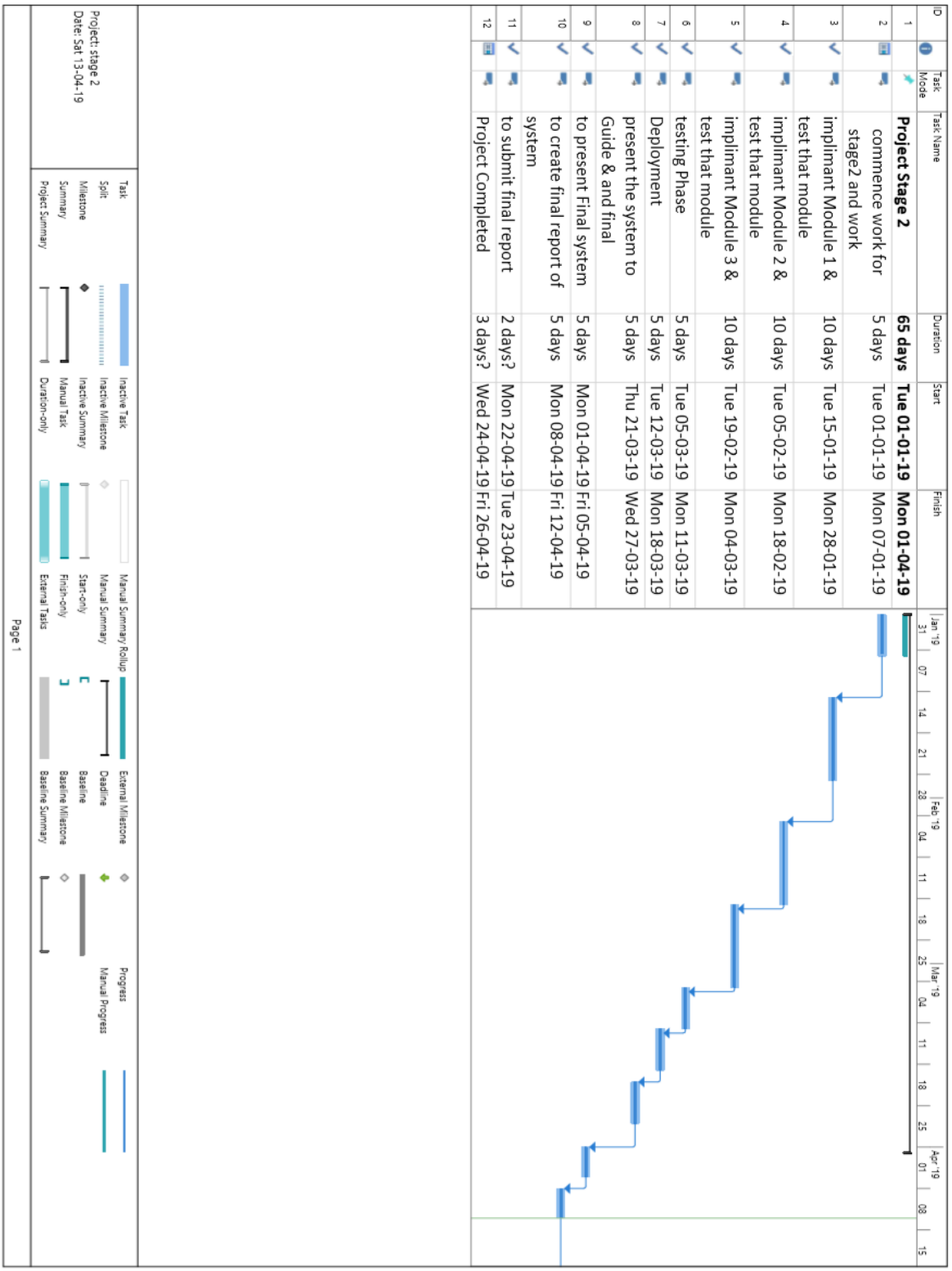


FIGURE 4.2: Plan For Stage 2

4.4 Software and Hardware Requirements

4.4.1 Hardware Requirement:

- 1.5 gigahertz (GHz) dual-core C.P.U.
- 4 GB RAM
- 1024x768 minimum screen resolution
- 10GB Of hard disk space

4.4.2 Software Requirements:

- Microsoft Windows 7+
- Xampp web server with mysql server.
- Text editor (notepad,notepad++)
- Anaconda.
- Microsoft Project
- Selenium
- Star Uml
- Web Browser

4.5 Conceptual Models

4.5.1 Data Flow Diagram

Level 0:-

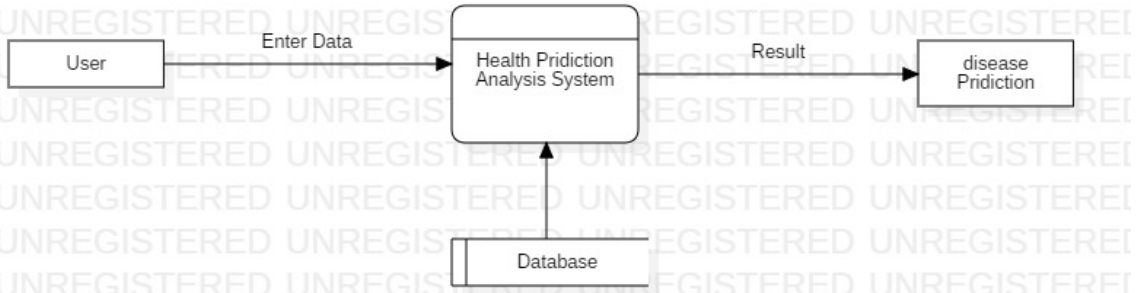


FIGURE 4.3: DFD Level 0

Level 1:-

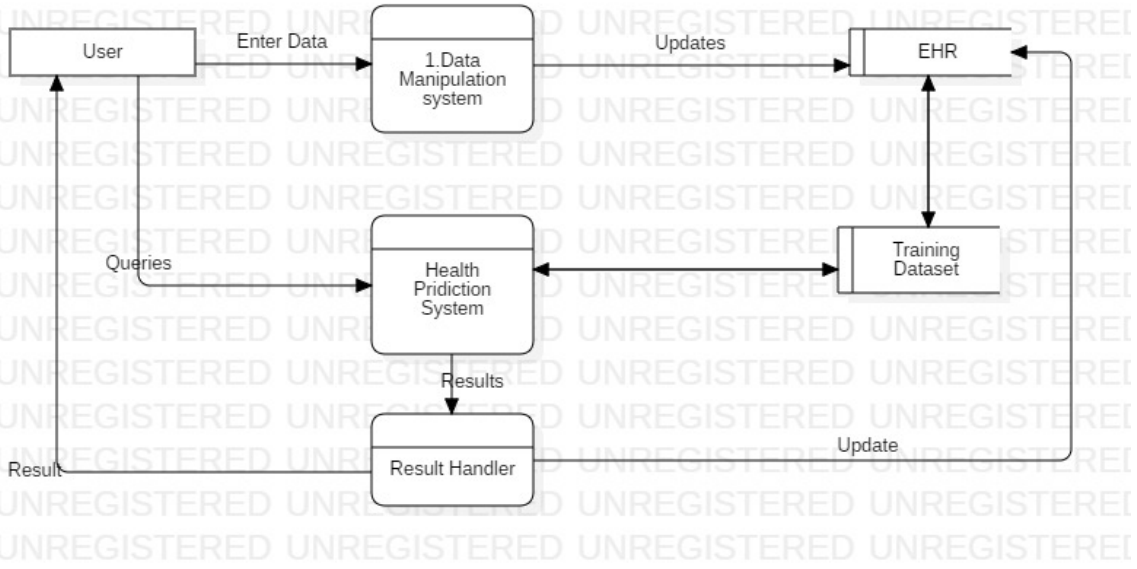


FIGURE 4.4: DFD Level1

Level 2:-

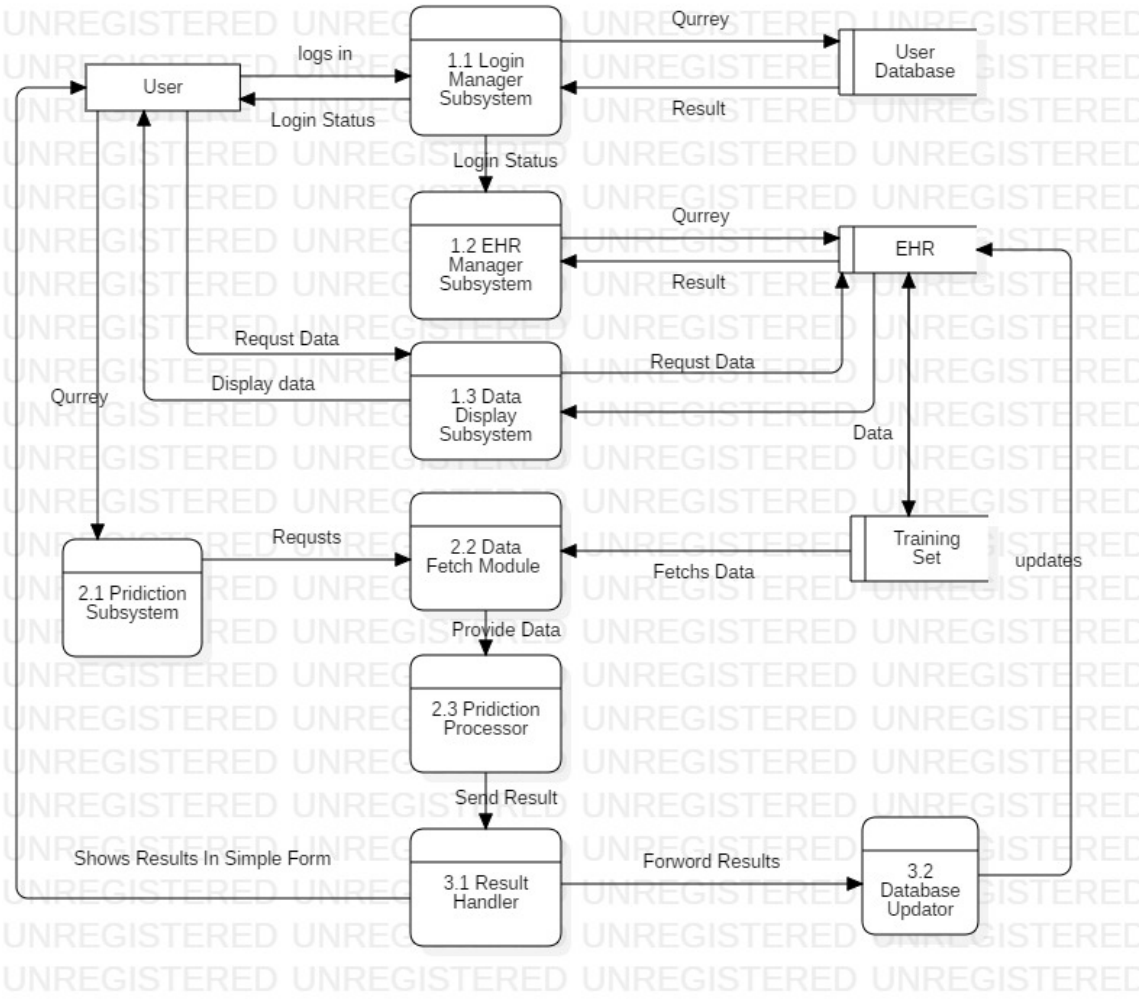


FIGURE 4.5: DFD Level 2

Chapter 5

SYSTEM DESIGN

In this chapter we study about various module , Various process diagrams of system which will help to give clear understanding of system.

5.1 Basic Modules

We had followed the divide and conquer theory, so we divided the overall problem into 3 parts and develop each part or module separately. When all modules are ready, we should integrate all the modules into one system.we briefly described all the modules and the functionality of these modules bellow.

- **User Authentication Module**
  This module will provide functionality of user authentication ,creation manages the credentials for user to authenticate.
- **Data Manipulation Module**
  This module will provide functionality of creation ,deletion,and modification of patient data.this module will also perform the data pre-processing since data provided by user may contain some missing field and noisy data which needs to be pre processed.

- **Prediction And Results Module**

  This module will Provide the prediction functionality for the system ,the prediction is done by using various data mining techniques.the prediction results where also handled by this module.

5.2 Data Design

On this section we will study about the data set and structure of data which we are obtaining and using for various orations.in which the training set is the database which is use for training.

5.2.1 Training Dataset

for training the system we had used the the data set present on UCI Machine Learning repository which is the repository which provides datasets for various purposes for free.from that we had used setLog heart dataset.which contains 13 attributes they are:

Attribute Information:

– 1. age
– 2. sex
– 3. chest pain type (4 values)
– 4. resting blood pressure
– 5. serum cholestoral in mg/dl
– 6. fasting blood sugar ¿ 120 mg/dl
– 7. resting electrocardiographic results (values 0,1,2)
– 8. maximum heart rate achieved
– 9. exercise induced angina
– 10. oldpeak = ST depression induced by exercise relative to rest
– 11. the slope of the peak exercise ST segment
– 12. number of major vessels (0-3) colored by flourosopy

– 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Attributes types

Real: 1,4,5,8,10,12

Ordered:11,

Binary: 2,6,9

Nominal:7,3,13

Variable to be predicted

Absence (1) or presence (2) of heart disease

where the rows represent the true values and the columns the predicted.

No missing values.

270 observations

5.2.2 Health Analysis Database

This database holds all the data in relational model i.e in tabular format.

5.2.2.1 EHR

EHR is the table which is used for storing health records, which contains attributes they are:

Name	Type	Description
Sr.	number	Serial No.,PatientId
Createdby	varchar	Creator of Record
Createdat	varchar	Tme of creation
Fname	varchar	Full Name
Uid	number	Aadhar no.
Addre	varchar	Address
Phone	number	Contact no.
Dob	varchar	Date of birth
Mt	varchar	Marital Status
Birpla	varchar	Birth place
Gender	varchar	Gender of patient
Emrname	varchar	Emergency contact name
Emrrel	varchar	Relation with Emr
Emradd	varchar	Address of Emr
Emrphone	number	Contact info of Emr
Insno	number	Insurance no
Exppodate	varchar	Expiry date of policy
Popro	varchar	Policy Provider
Insphone	number	contact no of policy provider
BG	varchar	Blood Group
Alg	varchar	Alergies
Did	number	Doctot Id
Dname	varchar	Doctor name
Dqal	varchar	Qualification of doctor
Dphone	number	Contact no of doctor
Dadd	varchar	Doctor Address
Hei	number	Height
Wei	number	Weight

5.2.2.2 Doctor

this table contains all the information required for login to the system for a Doctor which contains:-

Name	Type	Description
Username	varchar	Username foe accessing system
Email	varchar	Email Address
Password	varchar	Password hash
did	number	Doctor Id
dName	varchar	Doctor name
dadd	varchar	Doctor address
dPhone	number	doctor contact
dQal	varchar	Doctor Qualification

5.2.2.3 Users

this table contains all the information required for login to the system for a normal user.,contains:-

Name	Type	Description
Sr	number	Serial No., Patient Id
Username	varchar	User name for accessing system
Email	varchar	Email Address
Password	varchar	password hash

5.2.2.4 Pridi

this table contains prediction results generated by our system

5.2.3 Flow Diagram

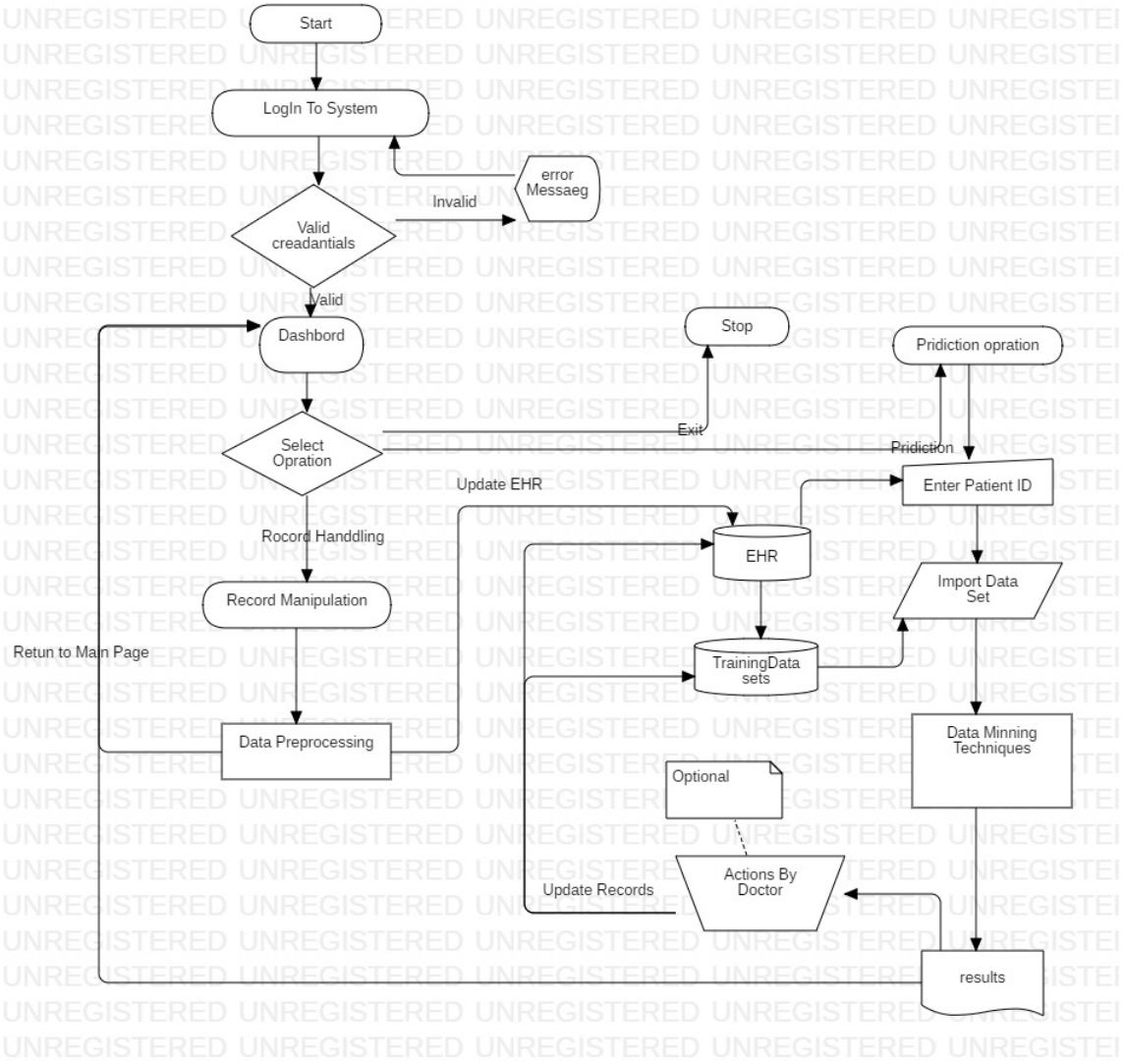


FIGURE 5.1: Flow Chart

5.2.4 Sequence Diagram

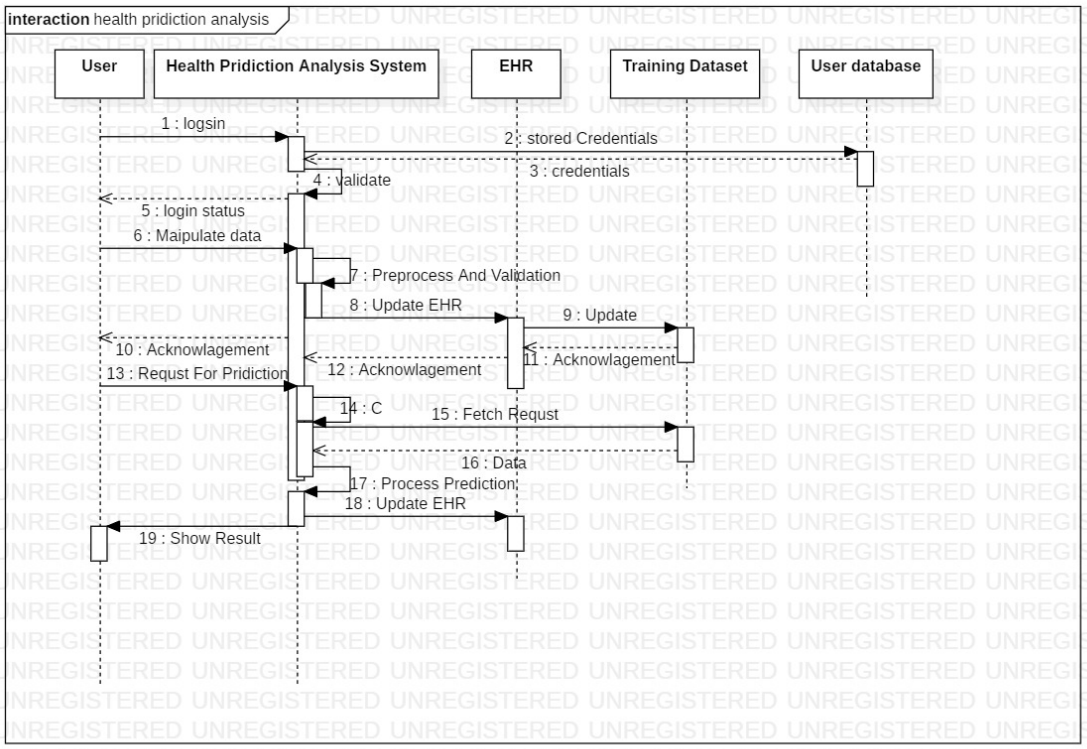


FIGURE 5.2: Sequence Diagram

5.3 User interface design

Login Module:-

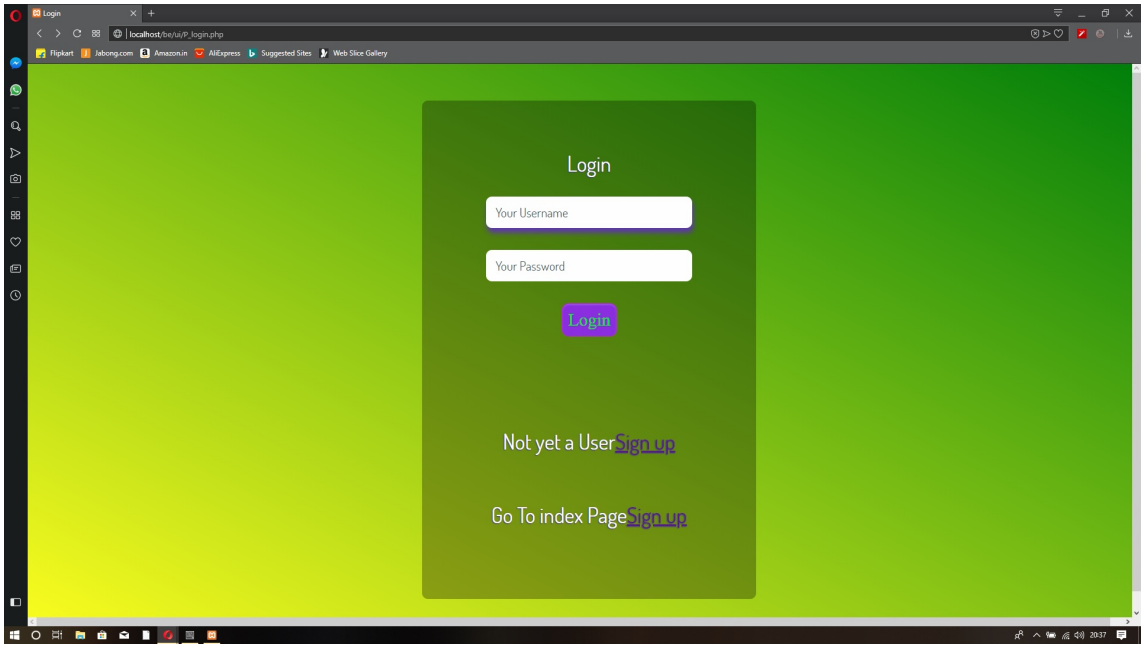


FIGURE 5.3: Login Module

This is an login module by using which the doctors or patients can access the system the doctor and patient both has separate login modules.

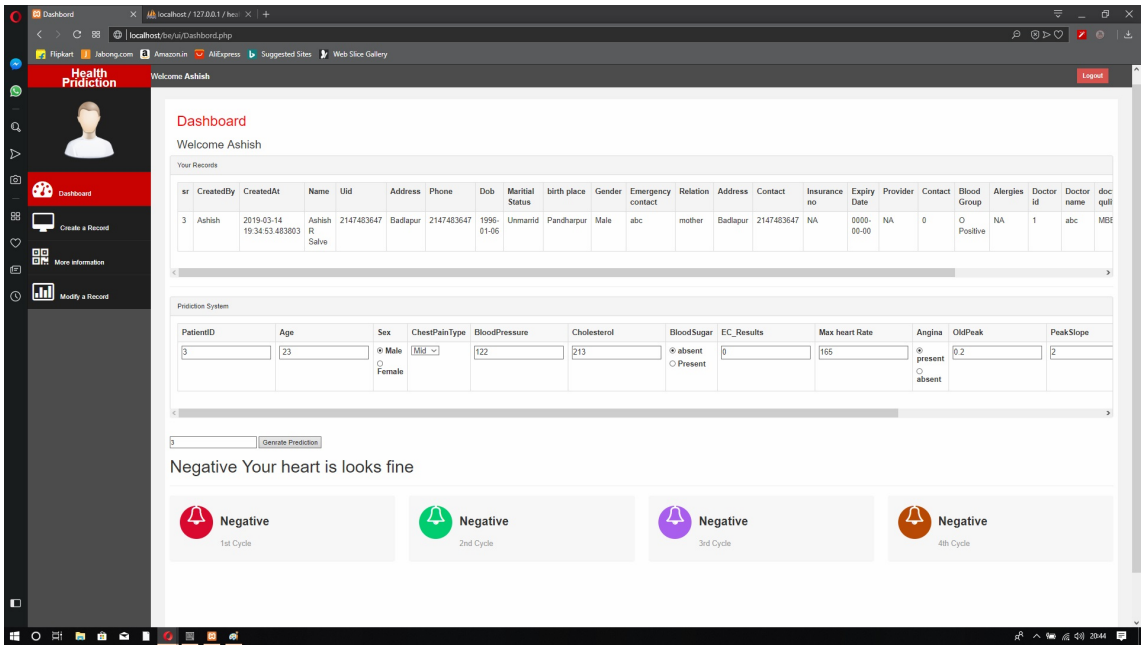


FIGURE 5.4: Patient Dashboard

Patient Dashboard:-

This is an dashboard for patient from which the patient can see their reports,EHR Information and he can also generate prediction and can view them.

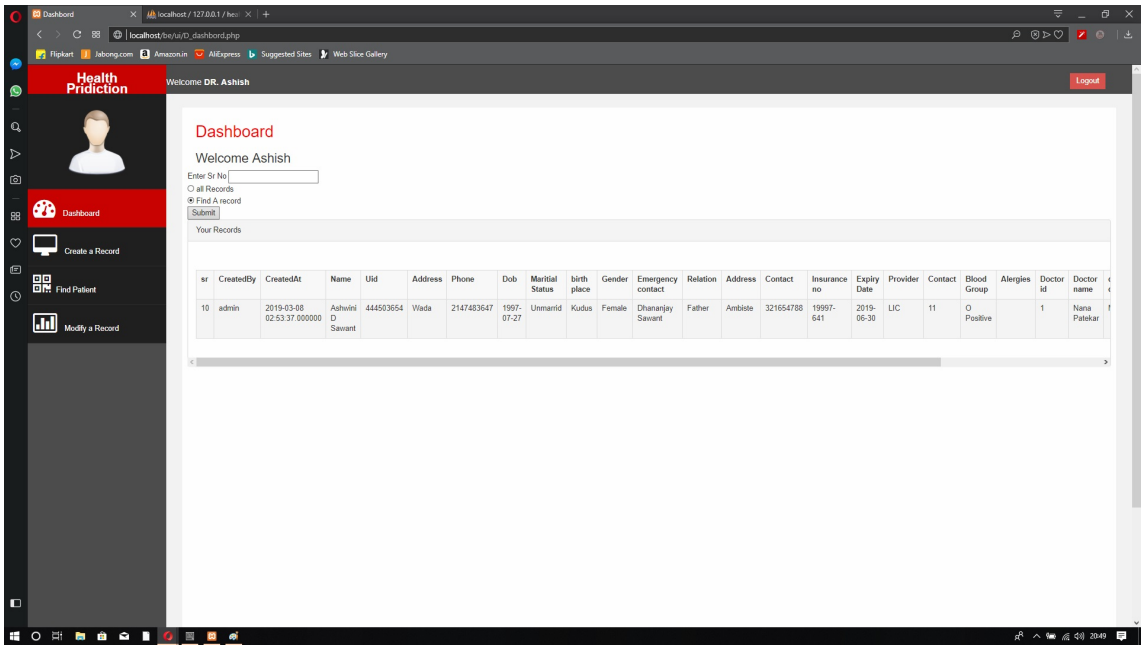


FIGURE 5.5: Doctor Dashboard

Doctor Dashboard:-

This is an dashboard for doctors from which the doctors can Find their patients,EHR Information of patient.

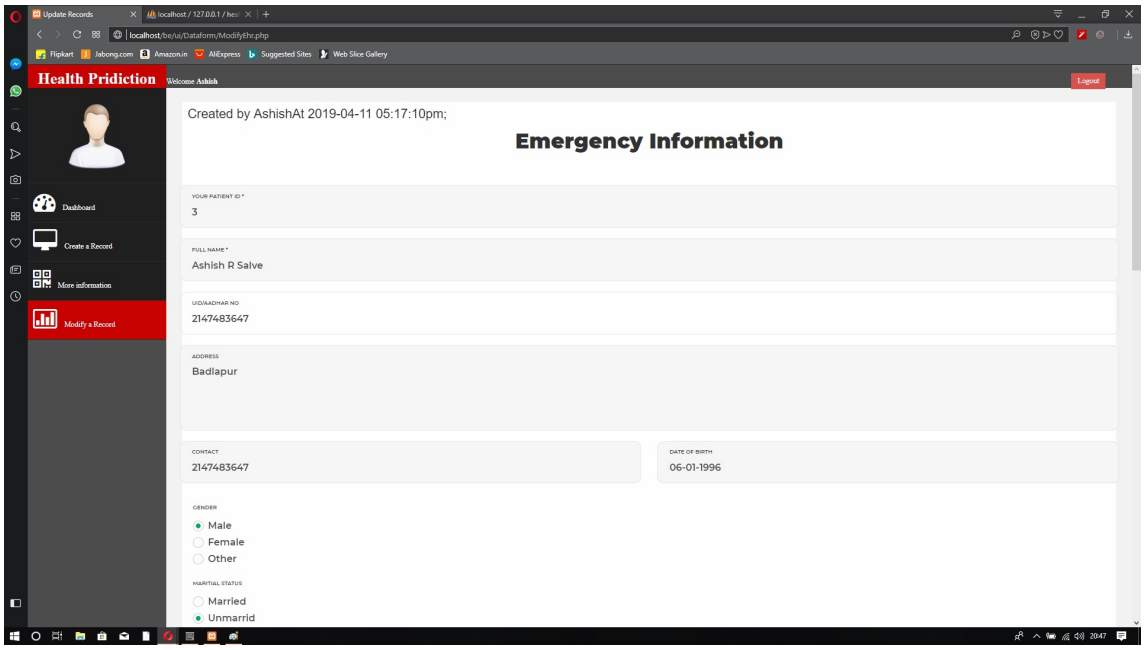


FIGURE 5.6: EHR Creation And Modification Module

EHR Creation And Modification Module:-

This is an form by which patient or doctor can create and modify the EHR information.

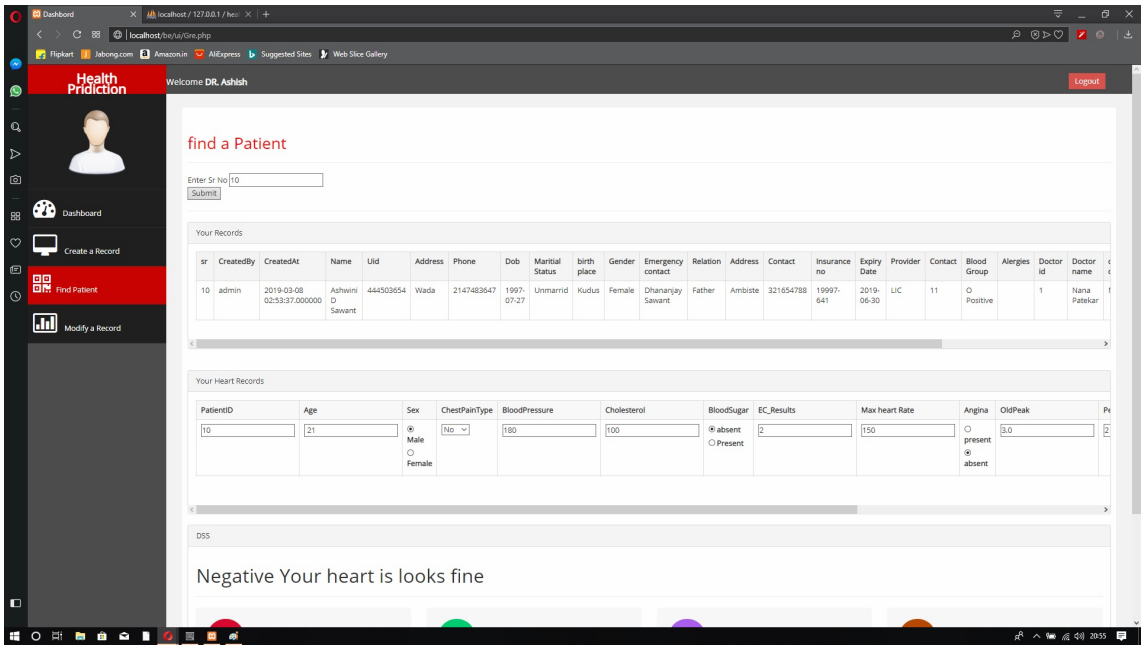


FIGURE 5.7: Find Patient Module

Find Patient Module:-

This module allows Doctors with proper patient id to see the EHR information of patient in emergency situations also he can generate prediction which will act as DSS.

5.4 Security Issues

1. The system stores data(except passwords) in unencrypted format which is easily understandable at server side.

2. Anyone can create multiple EHRs in unrestricted manner.

3. Doctors can see other doctors patient record unrestrictedly.

Chapter 6

IMPLEMENTATION AND TESTING

In This Chapter ,we will study about how we had implemented this system using various methodology and technologies and how we had tested them using various testing methods

6.1 Implementation Approaches

In This section ,we study about how we had implemented this project using various methodology and technologies

6.1.1 Programming Languages

6.1.1.1 HTML CSS

We had used **HTML**(the Hypertext Markup Language) and **CSS** (Cascading Style Sheets) along with bootstrap framework for making user interface for entire system HTML and CSS are one of the core technologies for building this project where HTML provides the structure of the page CSS handles layout designs.we had also used bootstrap templates for making our application more attractive and user-friendly the templates we had used are given bellow

Binary Admin from www.freecss.com

Contact Form 5 From www.uicookies.com

6.1.1.2 PHP

PHP stands for Hypertext Preprocessor and is a server-side programming language. PHP is a general-purpose scripting language that is especially suited to server-side web development,we had used PHP for Two purposes For making the code intract with the sql database and to execute queries on shell of server which executes the python scripts which generates results for pridictions

6.1.1.3 Python

we had used python for generating results and storing that result in database.and we had also used Scikit-learn ,which is probably the most useful library for data mining in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for data mining and statistical modelling including classification, regression, clustering and dimensionality reduction.

Python Libraries

Scikit-learn:-

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy

Pandas:-

Pandas is a popular Python package for data science, and with good reason: it offers powerful, expressive and flexible data structures that make data manipulation and analysis easy, among many other things. The DataFrame is one of these structures.Pandas also allows us to read data from verity of file formats.

MySQLdb:-

MySQLdb is an interface for connecting to a MySQL database server from Python.

6.1.1.4 SQL

SQL is used to communicate with a database.) it is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database.

6.2 Coding

The code for generating prediction and updating results into data base along with a php code which allows to pass user input as command line argument to python code is given bellow:-

Python Code For Prediction:-

```
import pandas as pd
import numpy as np
import MySQLdb
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB, BernoulliNB, MultinomialNB
import sys
sys.argv[1]=3
mysql_cn= MySQLdb.connect(host='localhost', port=3306,user='demo',
passwd='123',db='healthanalysis')
df = pd.read_sql('select * from heattrain', con=mysql_cn)
dar = pd.read_sql('select * from priid where PatientID=%s'
%(sys.argv[1]), con=mysql_cn)

id =(sys.argv[1])
#da = dar.drop('sr',axis=1)
da = dar.drop('PatientID',axis=1)
X = df.drop('Outcome',axis=1)
Xtr = X.drop('PatientID',axis=1)
Ytr = df['Outcome']

#Logestic Regration
logmodel = LogisticRegression()
logmodel.fit(Xtr,Ytr)
t1 = logmodel.predict(da)
if t1==1:
     t1 = 1
else:
     t1=2
#print (t1)
#RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=100)
rfc.fit(Xtr,Ytr)
t2= rfc.predict(da)
if t2==1:
     t2 = 1
else:
     t2=2
#print (t2)
#KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(Xtr,Ytr)
t3 = knn.predict(da)
if t3==1:
```

```
else:
     t3=2
#print (t3)
#naive bayes
gnb = GaussianNB()
gnb.fit(Xtr,Ytr)
t4 = gnb.predict(da)
if t4==1:
     t4 = 1
else:
     t4=2
print (t4)
#calculating result
ma=[t1,t2,t3,t4]
tre=pd.Series(data=ma).mean()
print(tre)
#updating result to db
val = {'Pid':id,'ta':t1,'tb':t2,'tc':t3,'td':t4,'tr':tre}
mycursor = mysql_cn.cursor()
sql="Delete from predi where PatientID = %(Pid)s"
mycursor.execute(sql,val)
sql = "INSERT INTO predi (PatientID,t1,t2,t3,t4,result) VALUES
(%(Pid)s,%(ta)s,%(tb)s,%(tc)s,%(td)s,%(tr)s)"
mycursor.execute(sql,val)
mysql_cn.commit()
mysql_cn.close()
print(val)
```

Code For Passing Command Line Arguments to Python Using PHP

```
<?php
if (isset($_POST['prid'])){
$a=$_POST['pre'];
$py="python";
$loc=" C:\\xampp\\htdocs\\be\\t1.py ";
$cmd=$py.$loc.$a;
echo "$cmd";
shell_exec($cmd);
}
?>
```

6.3 Testing Approach

We had tested the system By using Following Approach which includes various testing techniques listed bellow

Testing strategies

Unit Testing: Unit testing deals with testing a unit or module as a whole. This would test the interaction of many functions but, do confine the test within one module.

Integrated Testing: Brings all the modules together into a special testing environment, then checks for errors, bugs and interoperability. It deals with tests for the entire application. Application limits and features are tested here.
Automated Testing:Automation testing is an Automatic technique where the we writes scripts and uses suitable software to test the software. It is basically an automation process of a manual process.Automation testing also used to test the application from load, performance and stress point of view.

6.3.1 Unit Testing

Unit testing includes testing of each fields in the user log in forms, user's registration form,EHR form. This forms includes fields a username,email,password. each field includes fields as username,email,password.
Each field will be validates for its correctness,errors will be noted and will passed on to the developers for the correction. Users profile page have field like fullname,email,mobile no. address, age,gender.unit testing incudes whether the fields are accepting data as needed and shows error if data is filled incorrect. Unit testing also includes testing of buttons ike the navigation buttons on the homepage.
It checks whetger relative web page opens up when clicking on respective button.Record Search box should show error if wrong or unavailable data is given as input to the search query.

6.3.2 Integrated Testing

In Integration Testing we had the tested the system as a whole ,as a single structure.the main intention behind that was to test the linking between the different modules and they are functioning correctly and they where redirecting the user to proper location.in this we tested the links present in the patient module where working fine all functions which are integrated are working properly.the session is working properly or not.we also validated the code is passing the proper arguments to prediction module or not.

6.3.3 Test Cases

Test Case 1: *Test For testing multiple Inputs to Prediction Form*

Selenium output:-

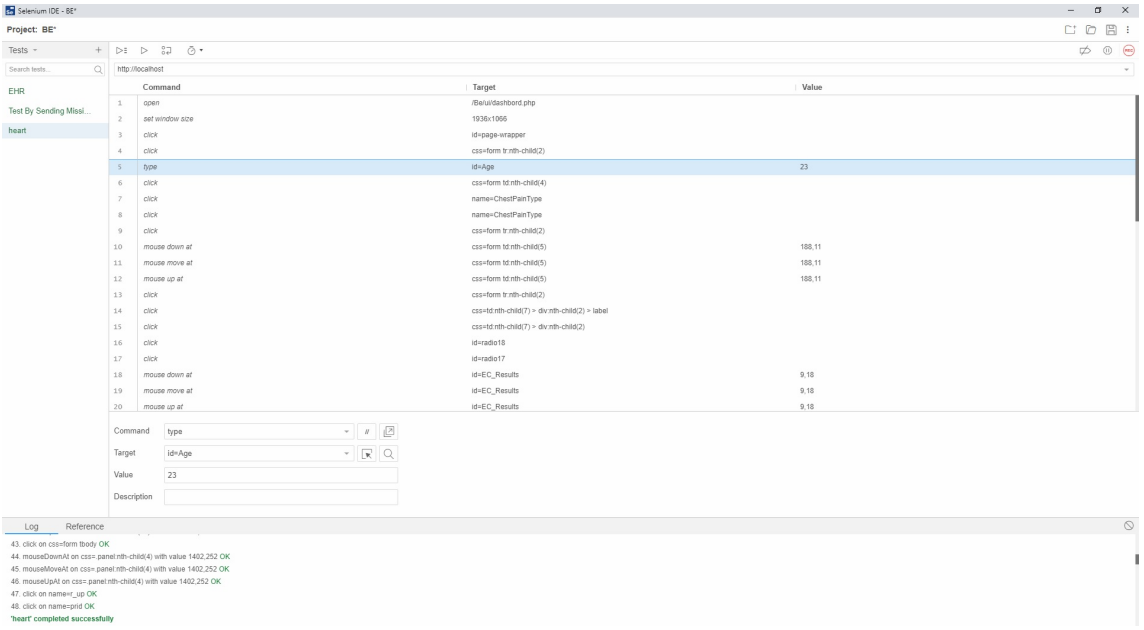


FIGURE 6.1: Selinum Output for Prediction Form

Log file From selinum

1. open on /Be/ui/dashbord.php OK
2. setWindowSize on 1936x1066 OK
3. click on id=page-wrapper OK

4. click on css=form tr:nth-child(2) OK
5. type on id=Age with value 23 OK
6. click on css=form td:nth-child(4) OK
7. click on name=ChestPainType OK
8. click on name=ChestPainType OK
9. click on css=form tr:nth-child(2) OK
10. mouseDownAt on css=form td:nth-child(5) with value 188,11 OK
11. mouseMoveAt on css=form td:nth-child(5) with value 188,11 OK
12. mouseUpAt on css=form td:nth-child(5) with value 188,11 OK
13. click on css=form tr:nth-child(2) OK
14. click on css=td:nth-child(7) $div : nth - child(2)$ $labelOK$

15.$clickoncss = td : nth - child(7)$ $div : nth - child(2)OK$

16.$clickonid = radio18OK$

17.$clickonid = radio17OK$

18.$mouseDownAtonid = EC_Resultswithvalue9, 18OK$

19.$mouseMoveAtonid = EC_Resultswithvalue9, 18OK$

20.$mouseUpAtonid = EC_Resultswithvalue9, 18OK$

21.$clickonid = EC_ResultsOK$

22.$mouseDownAtoncss = td : nth - child(7) > div : nth - child(1)withvalue63, 22OK$

23.$mouseMoveAtoncss = td : nth - child(7) > div : nth - child(1)withvalue63, 22OK$

24.$mouseUpAtoncss = td : nth - child(7) > div : nth - child(1)withvalue63, 22OK$

25.$clickoncss = formtr : nth - child(2)OK$

26.$clickoncss = formtr : nth - child(2)OK$

27.$mouseDownAtoncss = .panel : nth - child(4)withvalue1194, 285OK$

28.$mouseMoveAtoncss = .panel : nth - child(4)withvalue1194, 285OK$

29.$mouseUpAtoncss = .panel : nth - child(4)withvalue1194, 285OK$

30.$clickonid = radio19OK$

31.$mouseDownAtoncss = formtd : nth - child(10)withvalue64, 21OK$

32.$mouseMoveAtoncss = formtd : nth - child(10)withvalue64, 21OK$

33.$mouseUpAtoncss = formtd : nth - child(10)withvalue64, 21OK$

34.$clickoncss = formtr : nth - child(2)OK$

35.$clickoncss = formtr : nth - child(2)OK$

36.$mouseDownAtoncss = .panel : nth - child(4)withvalue1403, 255OK$

37.$mouseMoveAtoncss = .panel : nth - child(4)withvalue1403, 255OK$

38.$mouseUpAtoncss = .panel : nth - child(4)withvalue1403, 255OK$

39.$clickoncss = formtd : nth - child(13)OK$

40.$mouseDownAtoncss = formth : nth - child(13)withvalue126, 34OK$

41.$mouseMoveAtoncss = formth : nth - child(13)withvalue126, 34OK$

42.$mouseUpAtoncss = formth : nth - child(13)withvalue126, 34OK$

43.$clickoncss = formtbodyOK$

44.$mouseDownAtoncss = .panel : nth - child(4)withvalue1402, 252OK$

45.$mouseMoveAtoncss = .panel : nth - child(4)withvalue1402, 252OK$

46.$mouseUpAtoncss = .panel : nth - child(4)withvalue1402, 252OK$

47.$clickonname = r_upOK$

48.$clickonname = pridOK$

'heart' completed successfully

Test Case 2:Test For Testing The EHR Forms

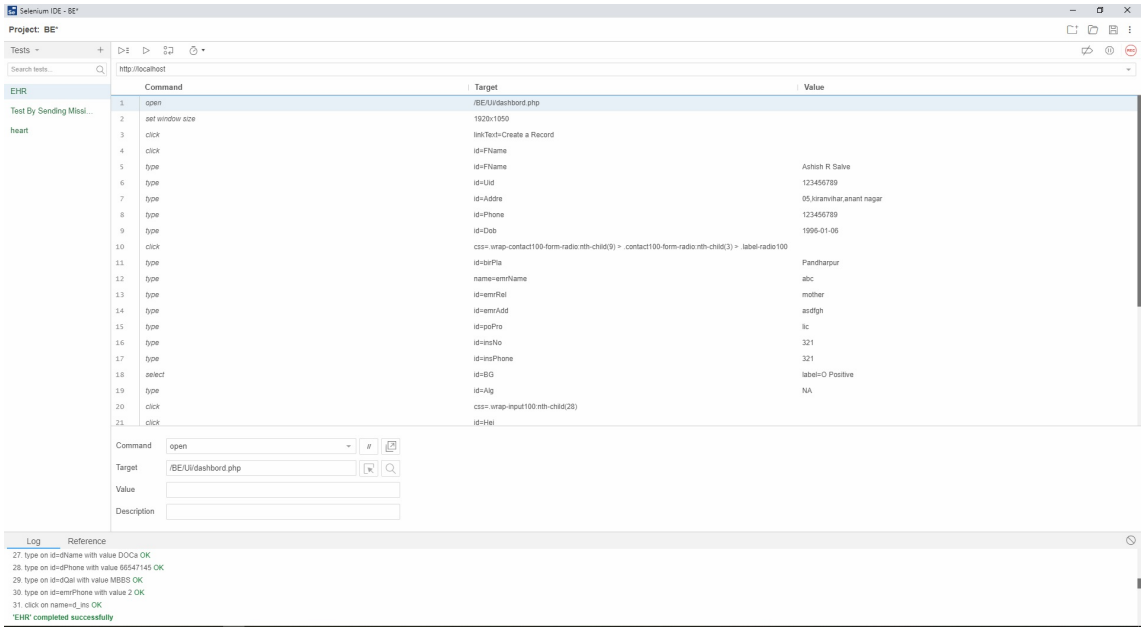


FIGURE 6.2: Selinum Output For EHR Form

Log file From selinum

Running 'EHR' 1. open on /BE/Ui/dashbord.php OK

2. setWindowSize on 1920x1050 OK

3. click on linkText=Create a Record OK

4. click on id=FName OK

5. type on id=FName with value Ashish R Salve OK

6. type on id=Uid with value 123456789 OK

7. type on id=Addre with value 05,kiranvihar,anant nagar OK

8. type on id=Phone with value 123456789 OK

9. type on id=Dob with value 1996-01-06 OK

10. click on css=.wrap-contact100-form-radio:nth-child(9) ¿ .contact100-form-radio:nth-child(3) ¿ .label-radio100 OK

11. type on id=birPla with value Pandharpur OK

12. type on name=emrName with value abc OK

13. type on id=emrRel with value mother OK

14. type on id=emrAdd with value asdfgh OK

15. type on id=poPro with value lic OK

16. type on id=insNo with value 321 OK

17. type on id=insPhone with value 321 OK

18. select on id=BG with value label=O Positive OK

19. type on id=Alg with value NA OK

20. click on css=.wrap-input100:nth-child(28) OK

21. click on id=Hei OK

22. type on id=Hei with value 185 OK

23. click on id=Wei OK

24. type on id=Wei with value 85 OK

25. click on id=did OK

26. type on id=did with value 1 OK

27. type on id=dName with value DOCa OK

28. type on id=dPhone with value 66547145 OK

29. type on id=dQal with value MBBS OK

30. type on id=emrPhone with value 2 OK

31. click on name=d$_{i}nsOK$

$'EHR'completedsuccessfully$

Test Case 3:-

Test ID:T3

Test Input:-999999999 (For Patient ID In DSS or Prediction System)

Expected Output:Not Present

Actual Output: –

Result:-Failed

Test Case 4:-

Test ID:T4

Test Input:-20/02/2020 (For Date Of Birth In EHR)

Expected Output:Error

Actual Output: Date Got Updated

Result:-Failed

6.4 Modifications and Improvements

we had found many small and large bugs in the systems from which we are mentioning some major bugs bellow:-

1. The prediction data of New user do not get updated in database, to solve this we had used 2 quarries at once 1st we had deleted the old data and then we updated new data to database

2.Complex User Interface Problem has been solved by using templates for dashboard and forms.

3.Date was not getting updated into database so we had observed that we where updating the date along with timestamps so it was mismatching the type.so we had changed the datatype of date to "Varchar2"

Chapter 7

RESULTS AND DISCUSSION

7.1 Test Reports

7.1.1 Test Results Of Python

We can compare various algorithms performance by using confusion matrix.For finding the confusion matrix and accuracy of the trained model we used same dataset named as heart_ disease in that there were few parameters regarding the heart disease and the final outcome which indicates actual result i.e. whether the patient had a heart disease or not so, by using this data we had trained a model using python sklearn library. For calculating confusion matrix, we required some training data and test data for that We split this Dataset into two different datasets, one for the independent Attribute x, and one for the Outcome Attribute y (which is the last column). We'll now split the dataset x into two separate sets xTrain and xTest. Similarly, we'll split the dataset y into two sets as well yTrain and yTest. we have split the dataset in a 70– 30 ratio, by specifying the test size parameter to 0.3. Then for generating the following Confusion matrix We had used "classification_ report()" functions from sklearn library with parameters Ground truth (correct) target values i.e. yTest and Obtained Prediction value by using respective algorithms. Then we find the accuracy by using accuracy_ score() function with same parameters as Confusion matrix Following are the Output of tests on various algorithms which we had implemented along with Confusion matrix and calculated accuracy.

```
------------------------- Logestic Regration------------------------------
--------------confusion matrix--------------------------
precision    recall  f1-score   support

0.81      0.95      0.88        41
0.94      0.78      0.85        40
 avg / total       0.88      0.86      0.86        81
 accuracy  of logistic regration is
0.8641975308641975
------------------- RandomForestClassifier--------------------------------
--------------confusion matrix--------------------------
precision    recall  f1-score   support

0.80      0.90      0.85        41
0.89      0.78      0.83        40
 avg / total       0.84      0.84      0.84        81
 accuracy of random forest  is
0.8395061728395061
---------------------- KNeighborsClassifier--------------------------------
--------------confusion matrix--------------------------
precision    recall  f1-score   support

0.62      0.76      0.68        41
0.68      0.53      0.59        40
 avg / total       0.65      0.64      0.64        81
 accuracy of KNN is 0.6419753086419753
------------------------------- Navi Bayes---------------------------------
--------------confusion matrix--------------------------
precision    recall  f1-score   support

0.79      0.93      0.85        41
0.91      0.75      0.82        40
 avg / total       0.85      0.84      0.84        81

accuracy of Navi bayes is
0.8395061728395061
```

7.1.2 Test Results Of Rapidminner

we had also tested same dataset with same data in rapidminner by using turbo prep function present in rapidminner and we had documented result below be had also compared that results with python results shown as follows:-

Model	Accuracy using Rapid Minner	Accuracy using python
Naive Bayes	0.9	0.83
Logistic Regression	0.8	0.86
Decision Tree	0.8	0.81
KNN		0.64

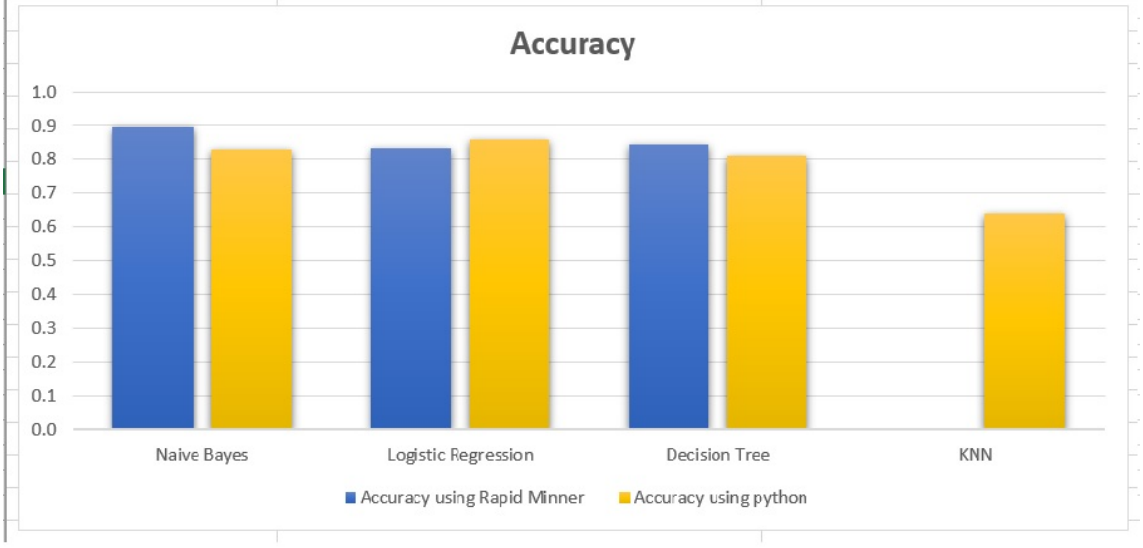


FIGURE 7.1: Result Comparison

In this we had observed that each algorithm has its own unique property on which the accuracy is determined. So the system is providing prediction based on mean of all results from various algorithms which helps to get more accurate and appropriate result which also help to make the application more robust.

7.2 User Documentation

- **Patient**

  patient just need to register to the system by simply clicking registered now button after registration user have to create EHR records by filing the form. Then the user can add or see its heart details and if he wishes to see the prediction he can simply click on generate prediction button present on bottom of the page he can also modified EHR record from modification tab.

- **Doctor**

  Doctors can registers the system same as patient but they have to mention there qualification and there contact numbers after creation of account doctor can crate modify EHR record created by them. They also have option to see all records or to find the particular records. Doctor can use the decision support system for verifies the decision made by them.

- **Find Records**

  This allows the third party to view EHR information created by any patient in case of emergency.

Chapter 8

CONCLUSIONS

8.1 Conclusion

The system can get great importance for area of healthcare, and it represents comprehensive process that demands thorough understanding of needs of the healthcare organizations. Knowledge gained with the use of this system can be used to make successful decisions that will improve success of healthcare organization and health of the patients. Our system requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results. Our system, once started, represents continuous cycle of knowledge discovery. currently the system was able to do only prediction of heart issues present on the person or not. in future we can add more health relates issues to be handled by the system so further development is needed in order to achieve system that is able to performed analysis of any kind of health issues.

8.2 Limitations of the System

- Privacy at risk

  since data is available to anyone who is authenticated or it may happen that the data is unintentionally got visible to an unauthorized user and he may miss-use your data.

- Requires large training set
  the system requires large no of data in training set to predict accurately,, more the data the more accurate prediction is

- Data manipulation is difficult and time consuming
  since it require large no of data filed or parameters, the system has to be able to manage it.for managing that it require more time and much more complex dataset.

8.3 Future Scope of the Project

In future the system can be used as:-

- As an mobile based application.

- In future the system can works with medical electronics to keep monitoring and analysing patients health in real time.

- In future the system can work As a nation wide health database attached with patients UID(adhar).

- In future the system can work with voice assistants and AI.So users doesn't need to always keep track of their reports since the AI can automatically does it for user and voice assistance can also provide day to day medical feeds to user

Bibliography

Conference and Journal:

[1] M. Sumanth. *Data Mining Applications in Healthcare Sector.*

[2] Neha Chikshe, Tejasweeta Dixit, Rashmi Gore Prerana Akade (2016). Hybrid Approach for Heart Disease Detection Using Clustering and ANN. *IJRITCC,JAN 2016 Volume 4 Issue 1.*

[3] Dr. Gagandeep Jagdev, (2015). Application Of Big Data In Medical Science Brings Revolution In Managing Health Care Of Humans. *IJEEE,JAN 2015 Volume 2 SPl. Issue 1*,

[4] Basel Kayyali, David Knott, and Steve Van Kuiken, (2013). The big-data revolution in US health care: Accelerating value and innovation *McKinsey Company April 2013*,

[5] Tapas Ranjan Baitharua, Subhendu Kumar Panib, (2016). Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset. *International Conference on Computational Modeling and Security (CMS 2016)* ,

[6] Wullianallur Raghupathi and Viju Raghupathi, (2014). Big Data Analytics in Healthcare:promise and potential *Health information science and system* ,

[7] Hian cbye Kob and Gerald Tan Data mining application in healthcare *Journal for information Management,vol 19.no.2*

[8] Sellappan Palaniappan , Rafiah Awang, (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques *IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8*

[9] Boris Milovic1, Milan Milovic , (2012). PREDICTION AND DECISION MAKING IN HEALTH CARE USING DATA MINING *Kuwait Chapter of Arabian Journal of Business and Management Review Vol. 1, No.12*

[10] Jyoti Soni , Ujma Ansari and team (2012). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction *International Journal of Computer Applications*

Books:

areth James, Daniela Witten ,Trevor Hastie , Robert Tibshirani ,"An Introduction to Statistical Learning" By Springer Publications

Websites:

1. www.ijritcc.com/index.php/ijritcc/article/view/1718, accessed on 17/08/18

2. www.slideshare.net/madallapallisumanth/data mininginhealthcaresector, accessed on 17/08/18

3. www.issuu.com/ijeeeapm/docs/id77, accessed on 17/08/18

4.www.immagic.com/eLibrary/ARCHIVES/GENERAL/WIKIPEDI/W1120615B.pdf, accessed on 31/08/18

5. www.Wikipedia.com, accessed on 05/10/18

6. www.saedsayad.com, accessed on 10/10/18

7. www.scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion$_matrix.html$, *accessed on* 9/01/19

8.*www.python.org*, *accessed on* 15/01/19

9.*www.archive.ics.uci.edu/ml/datasets/Heart + Disease*, *accessed on* 28/01/19

Publications

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

HEALTH PREDICTION ANALYSIS USING DATA MINING

Kritika Ashok Rane, Ashish Ravindra Salve
kritika2rane@gmail.com, a.r.salve@live.in
Ashwini Dhananjay Sawant
sawantashwini02@gmail.com
Department of Information Technology Konkan Gyanpeeth college of Engineering, Karjat Maharashtra, India

ABSTRACT
Data mining techniques are used for a variety of applications. In healthcare industry, datamining plays an important role in predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of tests can be reduced. This reduced test plays an important role in time and performance. This report analyses data mining techniques which can be used for predicting different types of diseases. This report reviewed the research papers which mainly concentrate on predicting various disease.

INTRODUCTION

There was a time when data were not readily available. As data become more abundant ,however limitations in computational capabilities prevented the practical application of mathematical models. Consequently, data mining tools are now being used for clinical data. The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc. Data mining is a process of selecting, exploring and modelling large amounts of data. This process has become an increasingly pervasive activity in all areas medical science research. Data mining has resulted in tha discovery of useful hidden patterns from massive databases. By using data mining techniques finally physicians need to know how quickly identify and diagnose potential cases.

Data Mining:-

Data mining is a process of selecting, exploring and modelling large amounts of data. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. The major steps involved in a data mining process are: Extract, transform and load data into a data warehouse. Store and manage data in a multidimensional databases. Provide data access to business analysts using application software. Present analysed data in easily understandable forms, such as graphs.

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

What Is The Use Of Data Mining In Healthcare?

The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc

Objectives:-

1. Predictive Analytics And Preventive Measures
2. The Ultimate HER
3. Disease Modelling and Mapping
4. Reduce Fraud And Enhance Security
5. Personalized Medicine

LITERATURE SURVEY

In the paper *"Data mining application in health care sector" by M.Sumanth* he to aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of disease data mining applications but it is challenging It need human efforts and improve accuracy. He also mention the application of data mining in healthcare sector such as Treatment effectiveness, Healthcare management, Customer relationship management, Medical device industry, Pharmaceutical industry, Hospital management, System biology.

In "Hybrid Approach for Heart Disease Detection Using Clustering and ANN " by Tejasweeta Dixit, Reshmi Gore, Prerana Akade they stated that the Heart disease is dominant caused death in developed countries and main contributors to disease strain in develop countries. Data mining the extraction of hidden predictive from large database is powerful new technology with great probability to help companies focus on the most important in info in their data warehouse. By using various data mining techniques such as clustering algorithem, decision tree, classifiers, we can predict heart diseases.

As stated in *"Application of big data in medical science brings in revolution in managing health care of humans" by Dr. Gagandeep Jagdev, Sukhpreet Singh* By enabling research to identify compounds with higher likelihood of success big data can help reduce the cost and the time to market for diagnosing the diseased. They also mentioned the roles played by big data in medical science such as Personalized treatment planning, Assisted diagnosis, Fraud direction ,Monitor patient vital signs, Digitization of disease.

As In the report *"Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset"* by *Tapas Ranjan Baitharua, Subhendu Kumar Panib* they conducted experiment on Liver Disorder Dataset by using WEKA Tool and they had found the impact of liver disorder on the predictive performance of different classifiers. After they analyzed quantitative data they found that the general concept of improved predictive performance of tested classifiers but Naive Bayes performance is not significant.

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

IMPLEMENTATION METHODOLOGIES

1. K-nearest neighbors algorithm:-

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN

2. Logistic Regression :-

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

3. Random forest :-

Random forests provide an improvement over bagged trees by way of a random forest small tweak that decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose m $\approx\sqrt{p}$—that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (4 out of the 13 for the Heart data).

SYSTEM DESIGN

We had followed the divide and conquer theory, so we divided the overall problem into 3 parts and develop each part or module separately. When all modules are ready, we should integrate all the modules into one system. We briefly described all the modules and the functionality of these modules bellow:

- **User Authentication Module**

This module will provide functionality of user authentication, creation manages the credentials for user to authenticate.

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

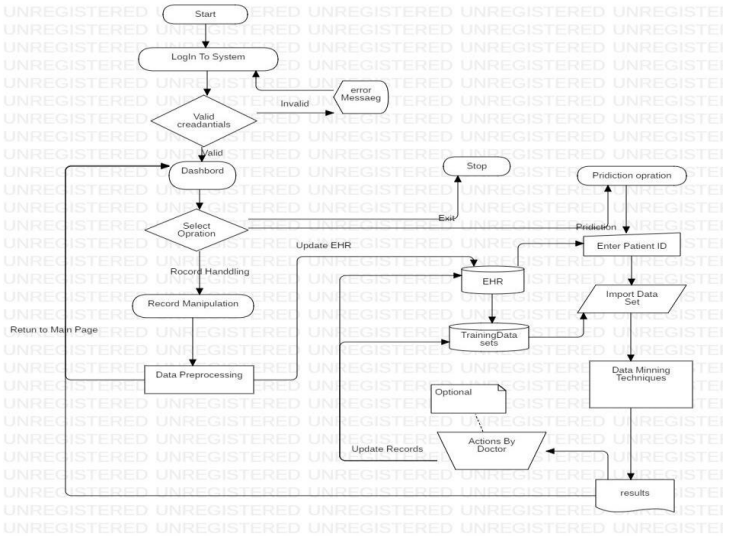
- **Data Manipulation Module**

This module will provide functionality of creation, deletion, and modification of patient data. This module will also perform the data pre-processing since data provided by user may contain some missing field and noisy data which needs to be pre-processed.

- **Prediction And Results Module**

This module will Provide the prediction functionality for the system ,the prediction is done by using various data mining techniques. The prediction results were also handled by this module.

Flow of System



CALCULATION REPORTS FOR VARIOUS ALGORITHM TESTED

We used same dataset for training of all models named as heart_disease, in that dataset there were few parameters regarding the heart disease and the final outcome which indicates actual result i.e. whether the patient had a heart disease or not so, by using this data we had trained a model using python & sklearn library.

For calculating confusion matrix, we required some training data and test data for that We split this Dataset into two different datasets, one for the independent Attributex, and one for the Outcome Attribute y (which is the last column). We'll now split the dataset x into two separate setsxTrain and xTest. Similarly, we'll split the dataset y into two sets as well yTrain and yTest. we have split the dataset in a 70– 30 ratio, by specifying the test size parameter to 0.3

Then for generating the following Confusion matrix We had used classification_report() functions from sklearn library with parameters Ground truth (correct) target values i.e. yTest and Obtained Prediction value by using respective algorithms. Then we find the accuracy by using accuracy_score() function with same parameters as Confusion matrix

Following are the Output of tests on various algorithms which we had implemented. The output contains confusion matrix and calculated accuracy

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

```
-------------------logistic regression --------------
confusion matrix
precision    recall  f1-score   support
1       0.81      0.75      0.78        40
2       0.77      0.83      0.80        41
avg / total       0.79      0.79      0.79        81

accuracy of logistic regression is
0.7901234567901234

----------------Random Forest Classifier---------------
confusion matrix
precision    recall  f1-score   support
1       0.77      0.75      0.76        40
2       0.76      0.78      0.77        41
avg / total       0.77      0.77      0.77        81

accuracy of random forest is
0.7654320987654321

------------- K Neighbours Classifier-----------------------
confusion matrix
precision    recall  f1-score   support
1       0.61      0.75      0.67        40
2       0.69      0.54      0.60        41
avg / total       0.65      0.64      0.64        81
accuracy of KNN is
0.6419753086419753
```

LIMITATIONS OF THE SYSTEM

1. Privacy at risk since data is available to anyone who is authenticated or it may happen that the data is unintentionally got visible to an unauthorized user and he may miss-use your data.
2. Requires large training set the system requires large no of data in training set to predict accurately,, more the data the more accurate prediction is
3. Data manipulation is difficult and time consuming since it require large no of data filed or parameters, the system has to be able to manage it.for managing that it require more time and much more complex dataset.

FUTURE SCOPE OF THE PROJECT

In future the system can be used as:-

1. As an mobile based application.
2. In future the system can work Can work with medical electronics to keep monitoring and analysing patients health in real time.
3. In future the system can work As a nation wide health database attached with patients UID(adhar).
4. In future the system can work with voice assistants and AI.So users doesn't need to always keep track of their reports since the AI can automatically does it for user and voice assistance can also provide day to day medical feeds to user

CONCLUSION

The system can get great importance for area of healthcare, and it represents comprehensive process that demands thorough understanding of needs of the

"Techno-Science-The Role of Entrepreneurial Development" ISBN: 978-93-88441-86-5

healthcare organizations. Knowledge gained with the use of the system can be used to make successful decisions that will improve success of healthcare organization and health of the patients. the system requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results.

ACKNOWLEDGMENT

Success is nourished under the combination of perfect guidance, care blessing. Acknowledgement is the best way to convey. We express deep sense of gratitude brightness to the outstanding permutations associated with success. Last few years spend in this estimated institution has molded us into confident and aspiring Engineers. We express our sense of gratitude towards our project guide Prof. J. P. Patil. It is because of his valuable guidance, analytical approach and encouragement that we could learn and work on the project. We will always cherish the great experience to work under their enthusiastic guidance.

REFERENCES

1. M. Sumanth. *Data Mining Applications in Healthcare Sector*.
2. Neha Chikshe, Tejasweeta Dixit, Rashmi Gore Prerana Akade (2016). Hybrid Approach for Heart Disease Detection Using Clustering and ANN. *IJRITCC,JAN 2016 Volume 4 Issue 1*.
3. Dr. Gagandeep Jagdev. Application Of Big Data In Medical Science Brings Revolution In Managing Health Care Of Humans. *IJEEE,JAN 2015 Volume 2 SPl. Issue 1*
4. Gareth James, Daniela Witten ,Trevor Hastie , Robert Tibshirani ,"An Introduction to Statistical Learning" ,Springer Publication
5. Tapas Ranjan Baitharua , Subhendu Kumar Panib "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset". *International Conference on Computational Modeling and Security (CMS 2016)*
6. S.K. Shinde & Uddagiri Chandrashekhar "Data Mining and business Intelligence ", Dreamtech Publication Websites:
    1. https://ijritcc.com/index.php/ijritcc/article/view/1718
    2. https://www.slideshare.net/madallapallisumanth/data-mininginhealthcaresector
    3. https://issuu.com/ijeeeapm/docs/id77
    4. https : //www.immagic.com/eLibrary/ARCHIV ES/GENERAL/W IKIP EDI/W 1120615B.pdf
    5. www.W ikipedia.com
    6. https://www.saedsayad.com
    7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
    8. https://www.python.org