Internship Project

report on

EFFECTIVE HEART DISEASE PREDICTION

Submitted by:

Meghamala C	4BD18CS045
Supritha G R	4BD18CS108
Smitha B M	4BD18CS094
Sriraksha N S	4BD18CS101

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BAPUJI INSTITUTE OF ENGINEERING AND TECHNOLOGY

DAVANGERE

ACADEMIC YEAR:2021-2022

CONTENTS

CHAPTERS PAGE NO'S

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Heart disease found to be most occurred disease on people of all age groups for the past many years. It is found to be most major threat by the World Health Organization (WHO). It is confirmed that the major reason for the 41% of deaths happened in real world is the heart attacks, strokes and other circulatory diseases. This surveyed and proved by the European public health alliance. Thus it is more essential to predict the occurrence of heart disease in the accurate manner with the consideration of the various security threats. Heart disease consists different kind of symptoms before it happens based on human health. It makes more difficult to predict the heart disease accurately due to presence of various kind of symptoms. Manual involvement of doctors would make this process easier where they can manually assign weight values to the attributes based on their important level. This would make heart disease prediction process more flexible and convenient.

1.2 PURPOSE

The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis . Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING SYSTEM

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods. Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

2.2 PROPOSED SYSTEM

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. shows the entire process involved.

CHAPTER 3

THEORETICAL ANALYSIS

HARDWARE AND SOFTWARE DESIGNING

HARDWARE DESIGNING:

The hardware required for the development of this projectis:

1. Processor : Intel® CoreTM i5-9300H
2. Processor speed : 2.4GHz
3. RAM Size : 8 GB DDR
4. System Type :X64-based processor

SOFTWARE DESIGNING:

The software required for the development of this projectis:

1. Operating System : Windows 10(and other higherversion)
2. Cloud Computing Service : IBM Cloud Services

CHAPTER 4

EXPERIMENTAL ANALYSIS

ANALYSIS OR INVESTIGATION MADE WHILE WORKING

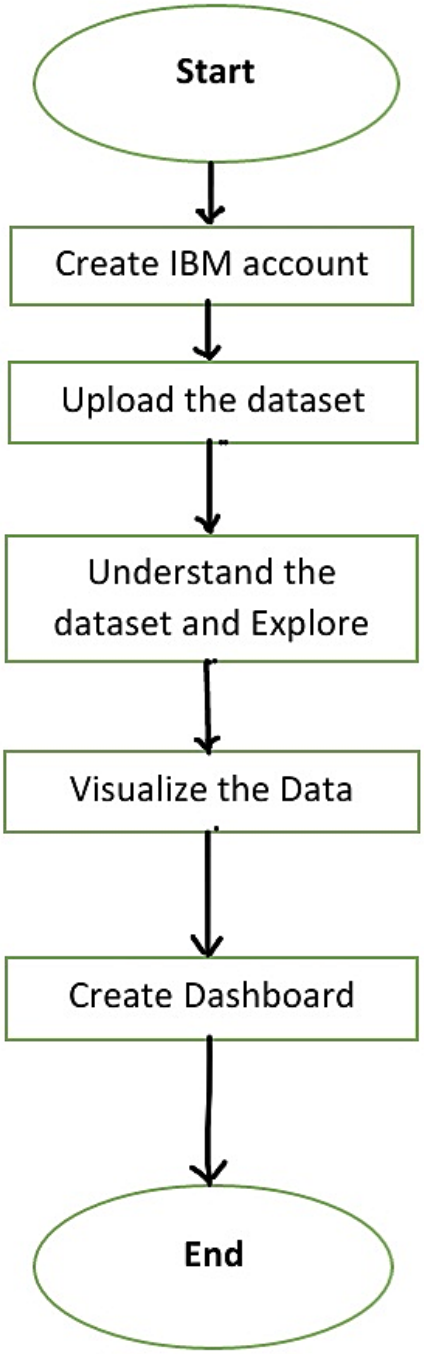
Risk factors are most prominent factors in the process of risk factor selection which needs to be focused well for the accurate and early prediction of heart disease. There are upto 300 risk factors have identified already which are reasonable for the coronary heart disease. Finding the presence of these risk factors would increase the prediction rate of heart disease. There are various research method has been introduced earlier by various researchers that focus on prediction of risk factors. Those research methods have been discussed here in this section. In authors performed risk factor presence analysis in the different kind of disease databases by adapting the non linear classifiers namely non linear support vector. Heart disease prediction techniques have been discussed in terms of their working procedure. The different metrics involved under those methodologies has been discussed in detail in this section. Authors introduced statistical methods for the prediction of coronary heart disease occurrence on patients. This research work would be adjusting the risk factors that are reasonable for the heart disease occurrence by finding the discrimination difference between them. Authors introduced firefly algorithm and the type 2 fuzzy logic system for the accurate prediction of heart disease.

ARCHITECTURE

The system architecture is like a blueprint of any object. It is a conceptual model to integrate between business logic and physical system in an organized way. It demonstrates the structure, view, behavior, features, and functionalities of the system. It is the way of portraying the desired

system in visualizing a way to well understand for people. The System architecture is the foundational orchestrate of a system that incorporated in its elements, their relationships of elements, and the science of its design (MITRE, 2019). HDPS is a web-based application that runs on the browser. This system is embodied in a web application. The web application architecture of the HDPS is to define the communication between applications, and database on the web. It also helps to know about other third-party application required like python's packages. It represents the represent the relationship between them and visualizes how they work together simultaneously. Architecture of HDPS has demonstrated the comprehensive structure of the system that easily understands all components and their relationship. The way to represent the system as a structural way can beneficial to understand the system for technical and non-technical users. The abstract architecture consists of a system, database, and interface design to visualize the better way of viewing. It describes the overall design of the system to easily understand the system elements, feature and functionalities, and behavior of the system.

FLOW CHART


Start
Create IBM account
Upload the dataset
Understand the dataset and Explore
Visualize the Data
Create Dashboard
End

CHAPTER 5

WORKFLOW

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease

1. Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination f 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features [9]. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis.

Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems.

In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

Naive Bayes

NaÃ¯ve Bayes algorithm is based on the Bayes rule[]. The independence between the attributes

of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds.

2. Classification

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques [12]. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analysed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further.

Random Forest

Algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

Decision Tree

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to records attribute

3. RESULT AND ANALYSIS

The results obtained by applying Random Forest, Decision Tee, Naive Bayes and Logistic Regression.

CHAPTER 6

RESULT

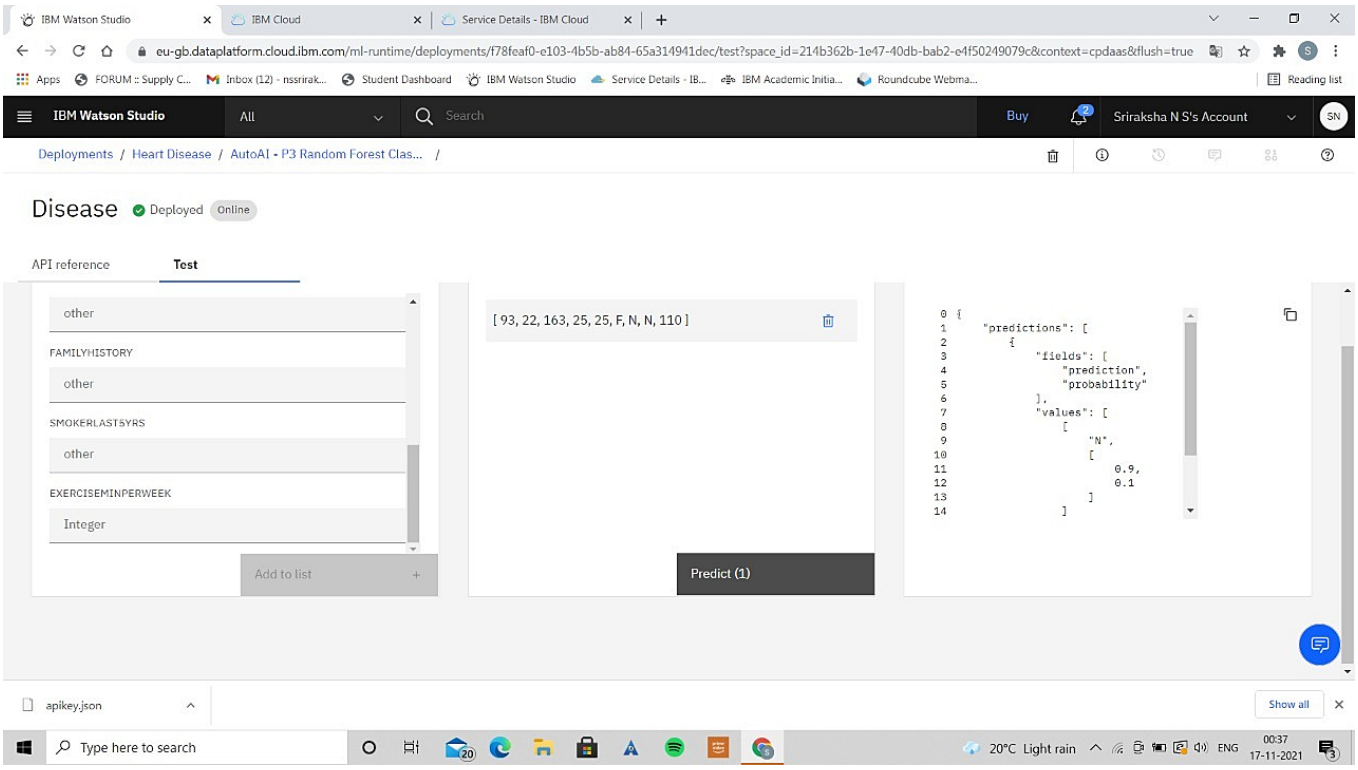


Fig : IBM Deployment , Predicting Heart Failure

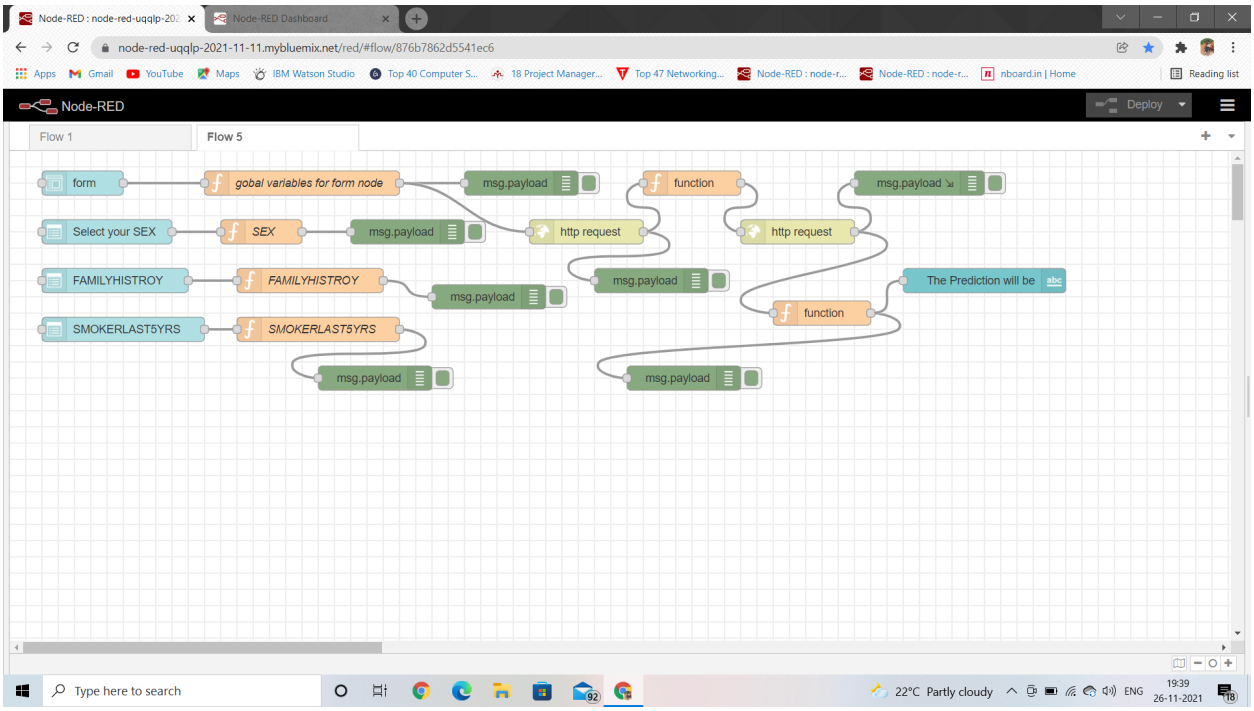


Fig : Node - Red Creation

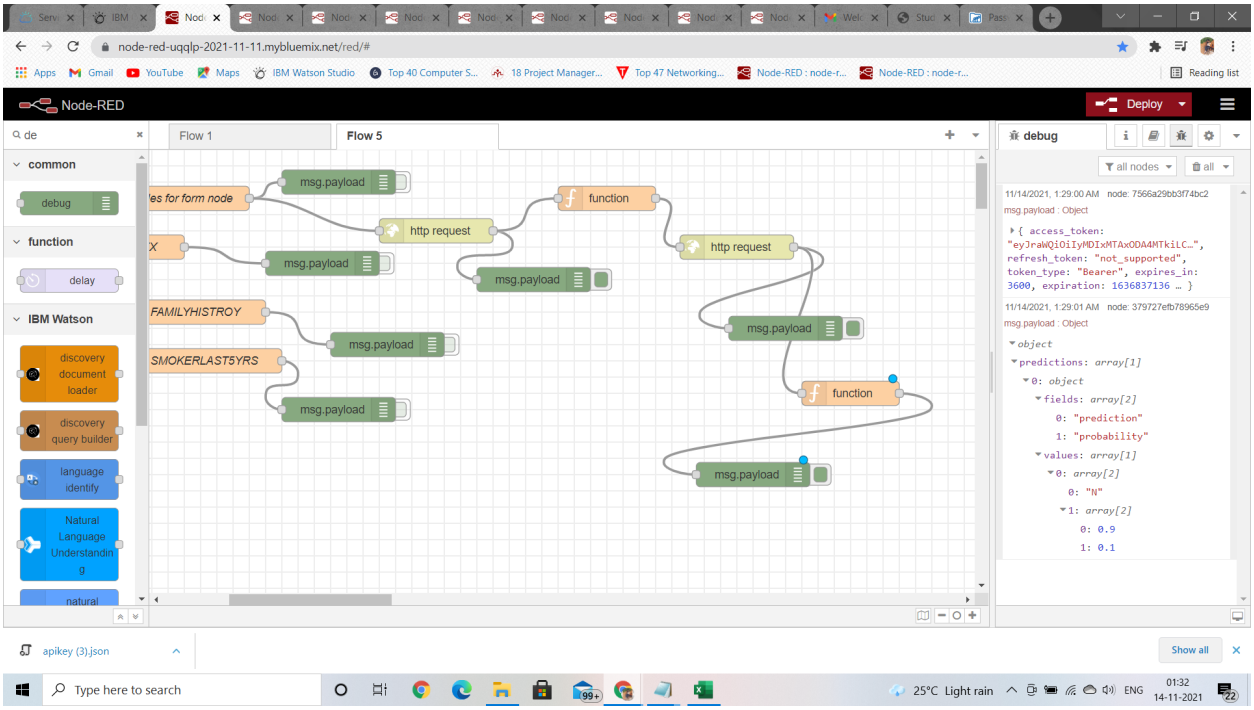


Fig : Output of access token along with the path of the object

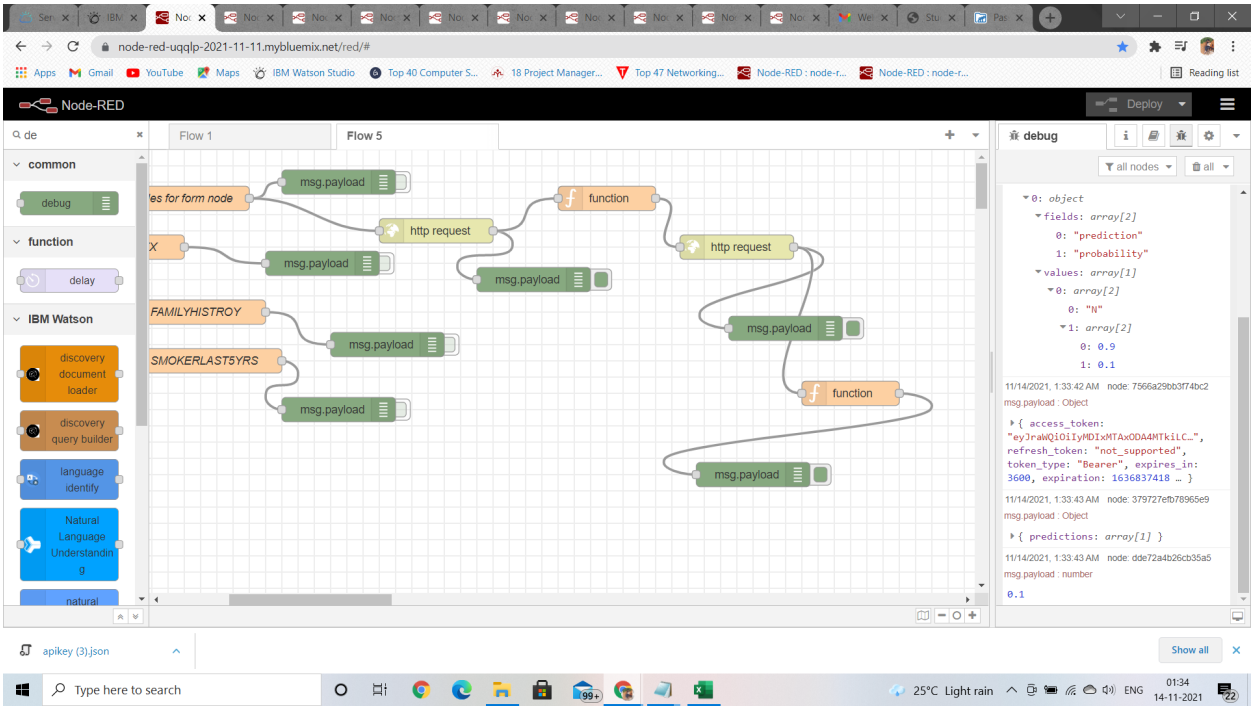


Fig : Output of the Prediction Column i.e Heart Failure

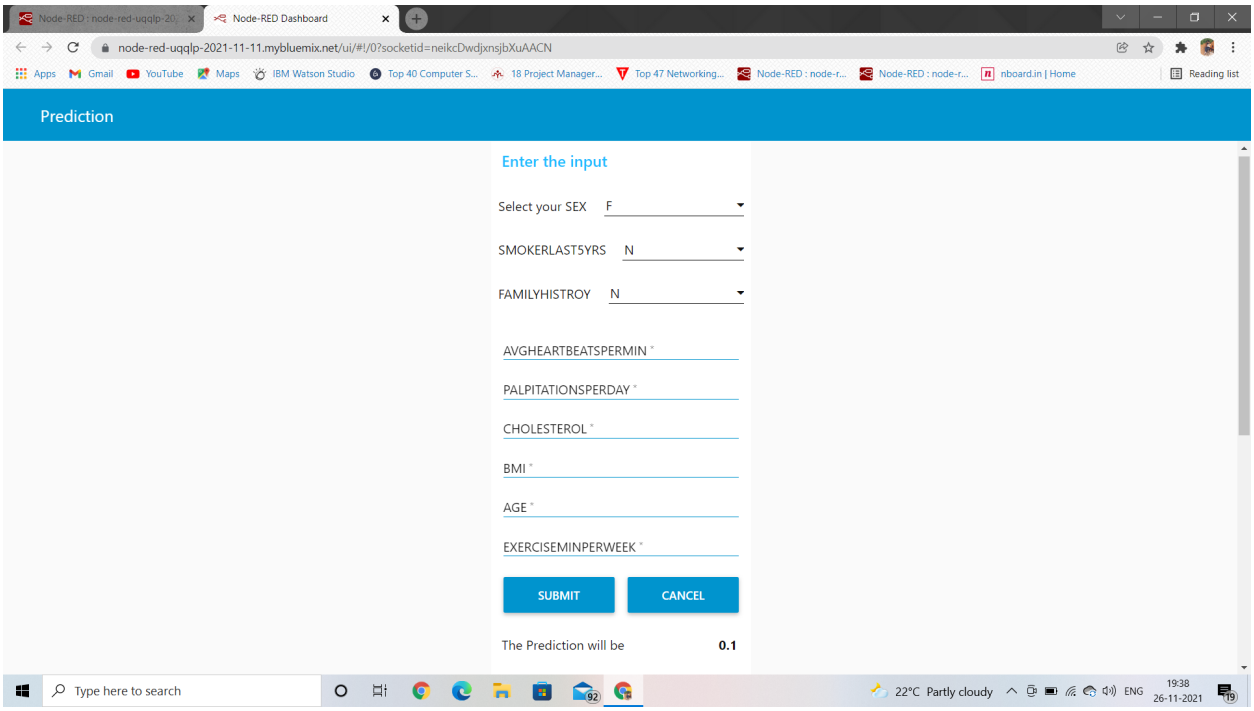


Fig :Final Output of the prediction Column

CHAPTER 7

ADVANTAGES AND DISADVANTAGES

Advantages

- User can search for doctor’s help at any point of time.
- User can talk about their Heart Disease and get instant diagnosis.
- Doctors get more clients online.
- Very useful in case of emergency.

Disadvantages

- Accuracy Issues: A computerized system alone does not ensure accuracy, and the warehouse data is only as good as the data entry that created it.
- The system is not fully automated, it needs data from user for full diagnosis.

CHAPTER 8

APPLICATIONS

The healthcare environment is generally perceived as being ‘information rich’yet ‘knowledge poor’. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) are used. This is an extension of naive Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. Discovery of hidden patterns and relationships often goes unexploited. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, eg patterns, relationships between medical factors related to heart disease, to be established.

CHAPTER 9

CONCLUSION

Heart disease prediction is the most concerned research issue focused by the various researchers to perform early prediction of heart disease. There is multiple research methods has been introduced earlier for prediction of heart disease in the accurate manner. In this research analysis paper, analysis multiple existing research methodologies has been discussed in terms of their working procedure and the performance metrics utilized. This analysis work also shows overview about their working procedure along with merits and demerits of those research methodologies. This work provides the discussion about the heart disease prediction in sectionalized way. Those are novel feature selection techniques; risk factors identification and prediction procedures. These research methods are defined and indicated with their faults and advantages. The overall review of the research methods has been conducted in the matlab simulation environment from which it can be found that improvement of the research methodologies. This evaluation has been carried out based on merits and demerits of research methods that have been discussed.

CHAPTER 10

FUTURESCOPE

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

CHAPTER 11

BIBILOGRAPHY

- https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
- https://www.seguetech.com/benefits-adhering-software-develpomentmethodolgy-concepts/
- https://www.researchgate.net/publication/331589020_Heart_Disease_Prediction System