HEART DISEASES PREDICTION SYSTEM

Project Stage ii report Submitted
In partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

By

Amit Panigrahi-2214390672

Under the Guidance of

Mrs S. D. CHAUDHARY

DEPARTMENT OF INFORMATION TECHNOLOGY

BHARATI VIDYAPEETH (DEEMED to be UNIVERSITY),
COLLEGE OF ENGINEERING, PUNE- 43 (2021-2022)

BHARATI VIDYAPEETH (DEEMED to be UNIVERSITY),
COLLEGE OF ENGINEERING, PUNE- 43

CERTIFICATE

This is to certify that the project report titled **HEART DISEASES PREDICTION SYSTEM**, has been carried out by **Amit Panigrahi** under the supervision of **Mrs. S. D. CHAUDHARY** in partial fulfillment of the degree of **BACHELOR OF TECHNOLOGY** in Information Technology of Bharati Vidyapeeth Deemed University, College of Engineering, Pune during the academic year 2021-2022.

Mrs. S. D. CHAUDHARY
(Project Guide)

Prof. Dr. S. B VANJALE
(Head)
(Dept. of Information)
Technology

Place: PUNE

Date: 13/07/2022

BHARATI VIDYAPEETH (DEEMED to be UNIVERSITY),

COLLEGE OF ENGINEERING, PUNE- 43

APPROVAL CERTIFICATE

This project report entitled **HEART DISEASE PREDICTION SYSTEM** by **(Amit Panigrahi)** is approved for the degree of **BACHELOR OF TECHNOLOGY**

Examiner Name & Sign

Guide's Name & Sign

Head of Department

Place: Pune

Date: 13/07/2022

Acknowledgement

Success is not only the hard work and innovation but also the inspiration and motivation. Completing this task was never one-effort. It is the result of invaluable contribution of number of individuals.

We would like to extend my sincere gratitude to the Principal **Dr. Vidula Sohoni**, Head of Department Computer Engineering, **Dr. S.B. Vanjale**, for nurturing a congenial yet competitive environment, which motivates all the students not only to pursue goals but also to elevate the humanitarian level.

Inspiration and guidance are invaluable in every aspect of life, which we have received from our respected project guide **Mrs. S. D. CHAUDHARY**, who gave me her careful and ardent guidance because of which we are able to complete this project. Mere words wouldn't suffice to express our gratitude to her untiring devotion.

We would also like to thank all the faculty members who directly or indirectly helped us from time to time with their invaluable inputs. Also, thanks to our family, friends who always co- operated us with their ideas and suggestions to make this project successful.

We would also like to thank Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci) who helped us in providing the free dataset.

Abstract

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (HDPS) is developed using Naives Bayes and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The HDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g. Relationships between medical factors related to heart disease and patterns, to be established. We have employed the multilayer perceptron neural network with backpropagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

Keywords:

Data Mining, K-nearest neighbor, Naïve Bayesian, Random forest, Decision Trees, Logistic regression, Support vector machine, Ensemble DM approach, Backpropagation, Disease Diagnosis

Table of Contents

Chapter 1

1.1 Background.

Among all fatal disease, heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease.

The term 'heart disease' includes the diverse diseases that affect heart. The number of people suffering from heart disease is on the rise (health topics, 2010).

The report from world health organization shows us a large number of people that die every year due to the heart disease all over the world. Heart disease is also stated as one of the greatest killers in Africa.

Data mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing. Of late, data mining has been applied successfully in healthcare fraud and detecting abuse cases.

1.2 Background of The Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

1.3 Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „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 the data for African genres**[2]**.

1.4 Objectives

1.4.1 Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set.

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

1.4.2 Specific Objectives.

- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
- Reduce the cost of medical tests.
-

1.5 Justification

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease.

1.6 Scope and Limitation

1.6.1 Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

1.6.2 Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

2.2 Tools

For application development, the following Software Requirements are:

Operating System: Windows 7 or any Linux Debian Distro.

Language: Python

Tools: Visual Studio Code, Microsoft Excel (Optional).

Technologies used: R, Unix, Shiny.

2.2.1 Software requirements:

Operating System	Any OS with clients to access the internet
Network	Wi-Fi Internet or cellular Network
Visual Code Studio	Create and design Data Flow and Context Diagram
Github	Versioning Control
Google Chrome	Medium to find reference to do system testing and run shinyApp

2.2.2 Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high

RAM: 1024 MB

Space on disk: minimum 100mb

For running the application: Device: Any device that can access the internet

Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i5-7th GEN processor (2.40GHZ, 2 Cores, 8GB RAM, running Windows 10 operating system with 64-bit Linux 4.31.0 kernel)

2.3 Literature Review

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.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed 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.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient's medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

2.4 Proposed Architecture

In this system we are implementing effective heart attack prediction system using Naïve Bayes algorithm. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is Naïve Bayes**.** After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

2.4.1 Naïve Bayes Classifier

Naïve Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which attribute value is independent of the values of other attributes. The Bayes theorem is as follows:

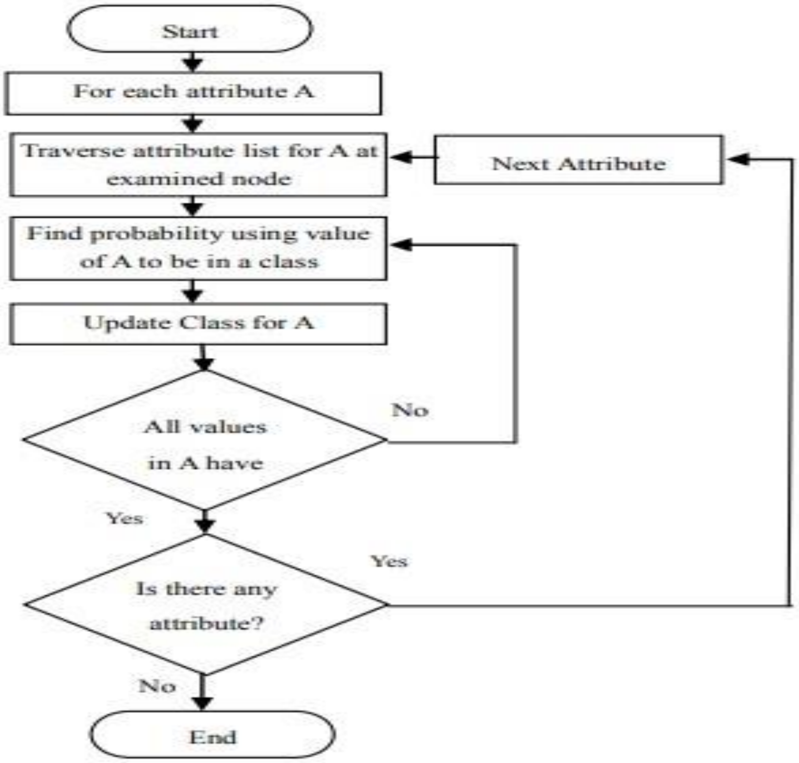
Let X= {x1, x2, ..., Xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as: P(H|X) = P(X|H) P(H) / P(X).4

Using Bayesian classifiers, the system will discover the concealed knowledge associated with diseases from historical records of the patients having heart disease. Bayesian classifiers predict the class membership probabilities, in a way that the probability of a given sample belongs to a particular class statistically. Bayesian classifier is based on Bayes' theorem. We can use Bayes theorem to determine the probability that a proposed diagnosis is correct, given the observation. A simple probabilistic, the naive Bayes classifier is used for classification based on which is based on Bayes' theorem.

According to naïve Bayesian classifier the occurrence or an occurrence of a particular feature of a class is considered as independent in the presence or absence of any other feature. When the dimension of the inputs is high and more efficient result is expected, the chief Naïve Bayes Classifier technique is applicable. The Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. For each input, it gives the possibility of attribute of the expectable state. Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value of a given class is independent of the values of other attributes. The advantage of using Naïve Bayes is that one can work with the Naïve Bayes model without. using any Bayesian methods. (Brownlee, 2016). P (Disease|symptom_1, symptom_2, … … . , symptom_n) **[3]**

P(Disease)P(symptom_1, … . . , symptom_n|Disease) =P(symptom_1, symptom_2, … . . Symptom_n).

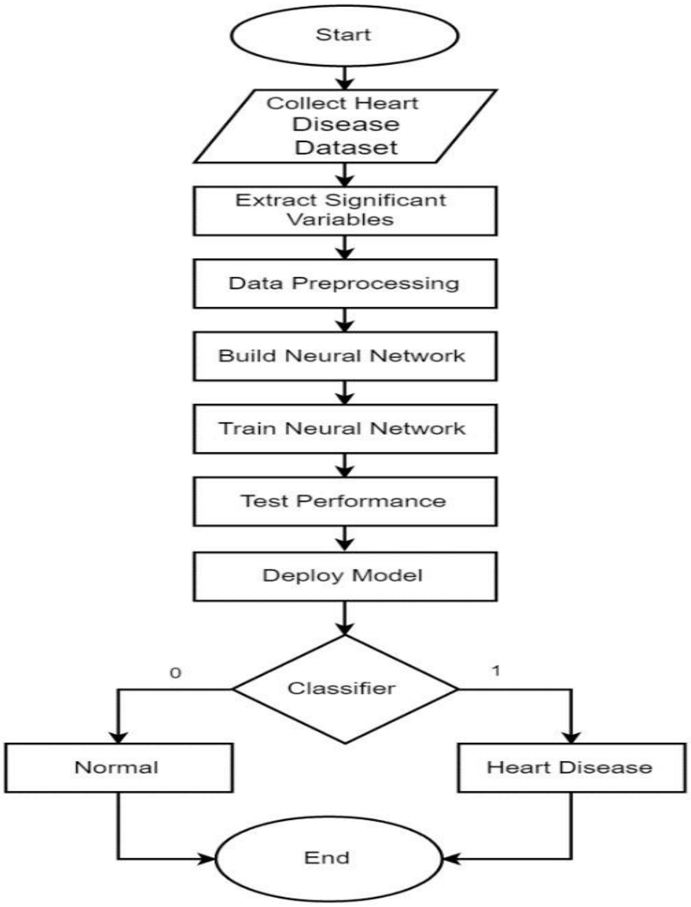
2.4.1.1 Flowchart Of Naïve Bayes Decision Tree Algorithm. [3.1]



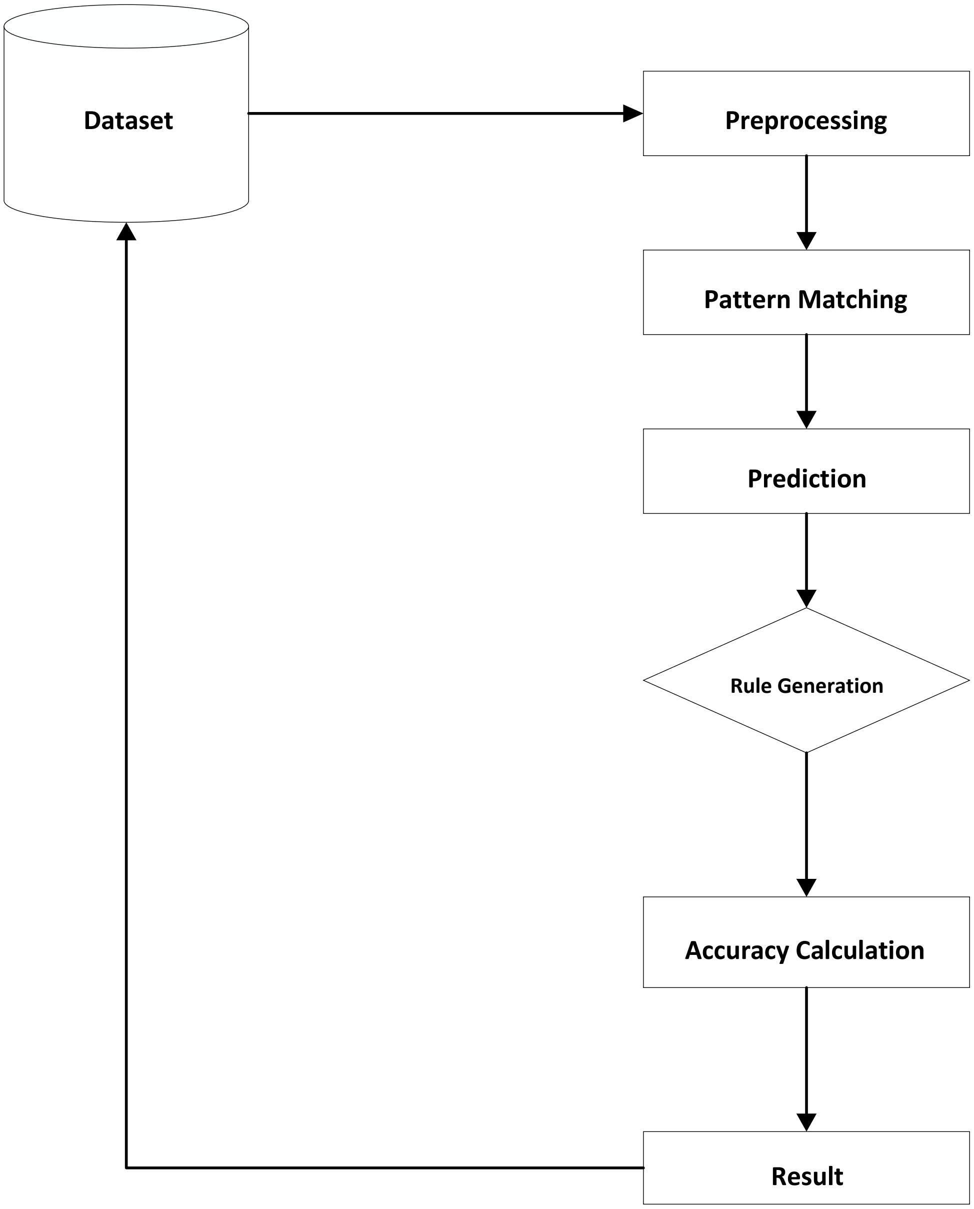
The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

2.4.2 Project Flow Chart.

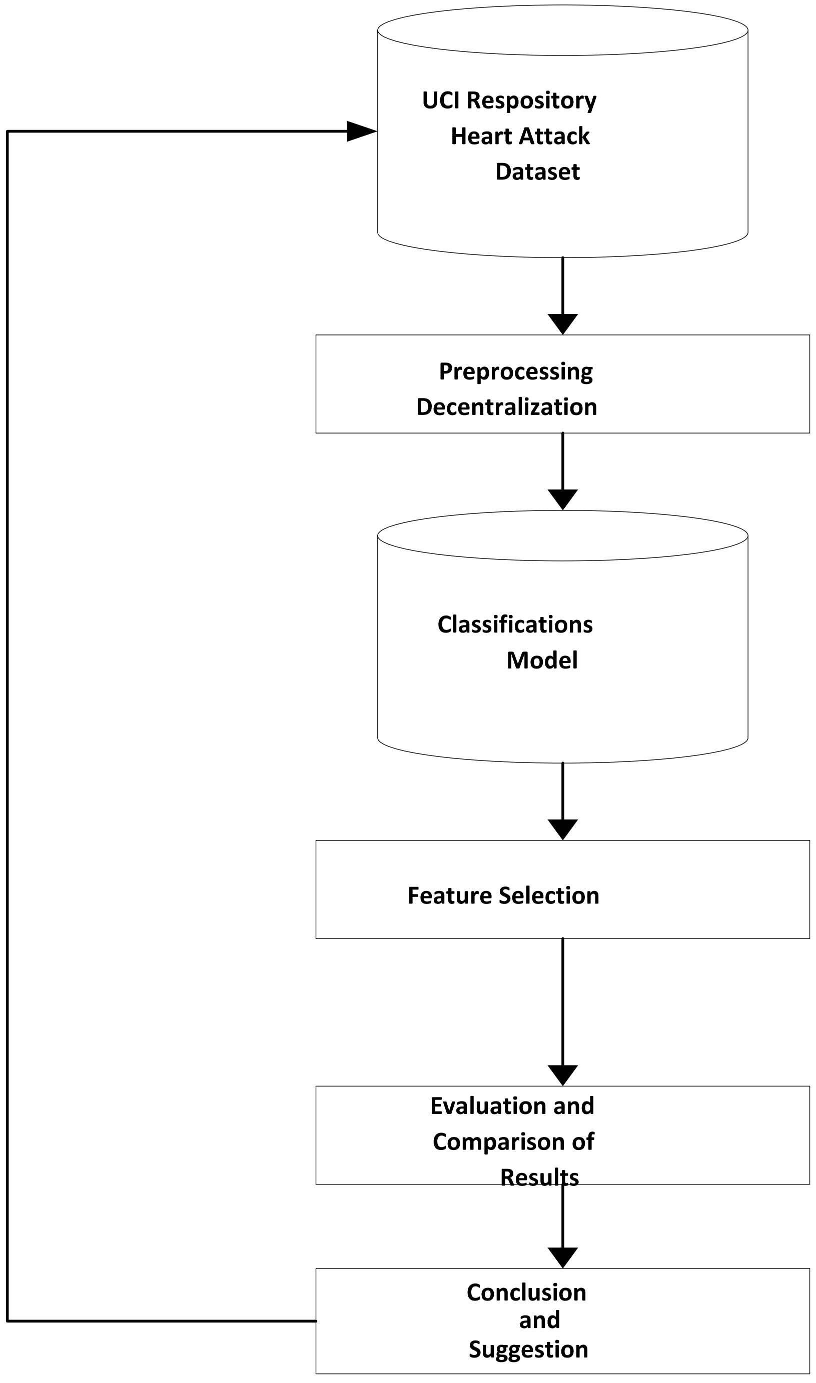
This will be the proposed flow chart that the system will look like



2.4.3 Data Flow Diagram



2.4.4 Proposed Model[4]



Chapter 3: Research Methodology

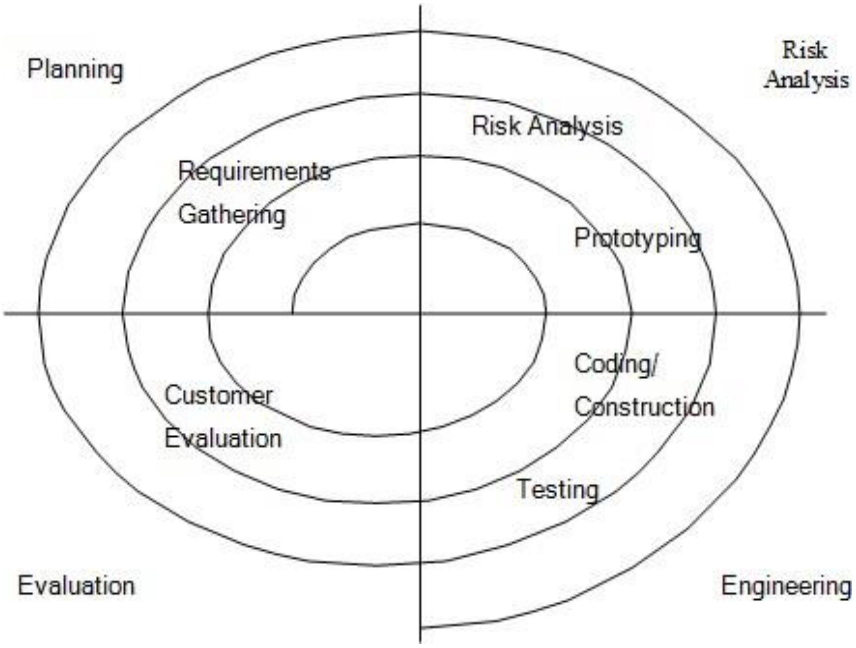
3.1 Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results**[5]**.

3.1.1 System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development. Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users**[6]**.



[7]

There are four phases that involve in the spiral model:

1) Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

2) Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

3) Engineering phase

At this phase, a software are created and testing are done at the end this phase.

4) Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

3.1.2 Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

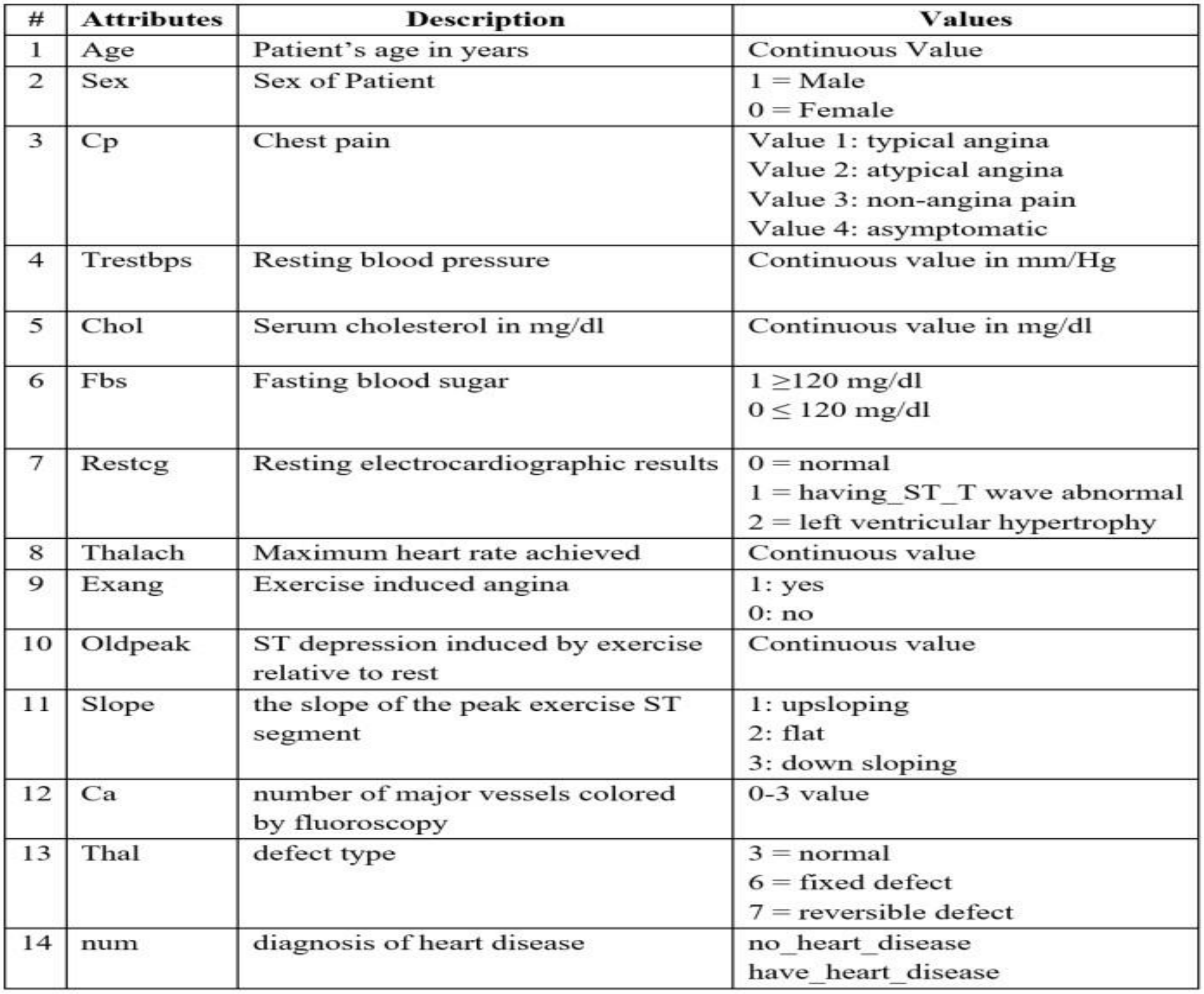
The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes**[8]**

His database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.
In particular, the Cleveland database is the only one that has been used by ML researchers to this date.
The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

3. 2 Algorithms and Classifiers Used for Experiments[9].



#	Attributes	Description	Values
1	Age	Patient's age in years	Continuous Value
2	Sex	Sex of Patient	1 = Male 0 = Female
3	Cp	Chest pain	Value 1: typical angina Value 2: atypical angina Value 3: non-angina pain Value 4: asymptomatic
4	Trestbps	Resting blood pressure	Continuous value in mm/Hg
5	Chol	Serum cholesterol in mg/dl	Continuous value in mg/dl
6	Fbs	Fasting blood sugar	1 ≥120 mg/dl 0 ≤ 120 mg/dl
7	Restcg	Resting electrocardiographic results	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value
9	Exang	Exercise induced angina	1: yes 0: no
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	Slope	the slope of the peak exercise ST segment	1: upsloping 2: flat 3: down sloping
12	Ca	number of major vessels colored by fluoroscopy	0-3 value
13	Thal	defect type	3 = normal 6 = fixed defect 7 = reversible defect
14	num	diagnosis of heart disease	no_heart_disease have_heart_disease

3.2.1 Naïve Bayesian

It is a probabilistic classifier based on Bayes' theorem specified by the prior probabilities of its root nodes. The Bayes theorem is given in Equation 1 and normalization constant is given in Equation 2. It proves to be an optimal algorithm in terms of minimization of generalized error. It can handle statisticalbased machine learning for feature vectors and assign the label for feature vector based on maximal probable among available classes {XX1, X2..., XM}. It means that feature "y" belongs to Xiclass, when posterior probability is maximum ie Max. The Bayesian classification problem may be formulated by aposterior probabilities that assign the class label ωi to sample X such that is maximal. The Bayesian classification problem may be formulated by a-posterior probabilities that assign the class label ωi to sample X such that is maximal.

Likelihood Prior

$$P(X_i \mid \underline{y}) = \frac{p(\underline{y} \mid X_i)P(X_i)}{p(\underline{y})} \quad (1)$$

Normalization Constant

$$p(\underline{y}) = \sum_{i=1}^{2} p(\underline{y} \mid X_i)P(X_i) \quad (2)$$ **[10]**

Application of Bayes' rule with the mutual exclusivity in diseases and the conditional independence in findings is known as the Naïve Bayesian Approach. It is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. Naïve Bayesian classifier despite its simplicity, it surprisingly performs well and often outperforms in complex classification. Simple Naïve Bayesian can be implemented by plugging in the following main Bayes' formula:

P (X1, X2…, Xn | Y) = P (X1 | Y) P (X2 | Y) … P (Xn | Y) (3)

The above-mentioned Naïve Bayesian network produces a mathematical model, which is used for modeling the complicated relations of random variables of disease attributes and decision outcome. The algorithm uses the formula to calculate conditional probability with respect to disease condition attributes value and decision attribute value. Based on prior knowledge, the algorithm classifies the decision attribute into labels assigned, and hence the conditional support is computed for each variable attribute.

3.2.2 Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications**[11]**. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

3.2.3 Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers**[12]**. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

3.2.4 Random Forest

RF is a machine learning algorithm based on the ensemble of decision trees. In traditional decision tree methods such as C4.5 and C5.0, all the features are used for generating the decision tree. In contrast, RF builds multiple decision trees and chooses the random subspaces of the features for each of them. Then, the votes of trees are aggregated and the class with the most votes is the prediction result. As an excellent classification model, RF can successfully reduce the overfitting and calculate the nonlinear and interactive effects of variables. Besides, the training of each tree are done separately, so it could be done in parallel, which reduced the training time needed. Finally, combining the prediction result of each tree could reduce the variance and improve the accuracy of the predictions**[13]**.

3.2.5 Extreme learning machine

ELM was first proposed by Huang et al. Similar to a single layer feed-forward neural network(SLFNN), ELM is also a simple neural network with a single hidden layer. However, unlike a traditional SLFNN, the hidden layer weights and bias of ELM are randomized and need not to tune, and the output layer weights of ELM are analytically determined through simple generalized inverse operations

3.2.6 Logistic regression

LR is a generalized linear regression model. Therefore, it is similar with multiple linear regression in many aspects. Usually, LR is used for binary classification problems where the predictive variable y∈[0,1]y∈[0,1], 0 is negative class and 1 is positive class. But it can also be used for multi-classification.
In order to distinguish heart disease patients from healthy people, a hypothesis h(θ)=θTXh(θ)=θTX is proposed. The threshold of classifier output is hθ(x)=0.5hθ(x)=0.5, which is to say, if the value of hypothesis hθ(x)≥0.5hθ(x)≥0.5, it will predict y=1y=1 which means that the person is a heart disease patient, otherwise the person is healthy. Hence, the prediction is done.
The sigmoid function of LR can be written as**[14]**:

hθ(x)=11+e−z,hθ(x)=11+e−z,

where z=θTXz=θTX.
The cost function of LR can be written as:

J(θ)=1m∑i=1mcost(yi,y'i),J(θ)=1m∑i=1mcost(yi,yi'),

where *m* is the number of instances to be predicted, yiyi is the real class label of the *i*th instance, and y'iyi' is the predicted class label of the *i*th instance.

cost(yi,y'i)={0,1,yi=y'iotherwise.

3.2.7 Support vector machine

Invented by Cortes and Vapnik, SVM is a supervised machine learning algorithm which has been widely used for classification problems. The output of SVM is in the form of two classes in a binary classification problem, making it a non-probabilistic binary classifier. SVM tries to find a linear maximum margin hyperplane that separates the instances.

Assume the hyperplane is wTx+b=0wTx+b=0, where *w* is a dimensional coefficient vector, which is normal to the hyperplane of the surface, *b* is offset value from the origin, and *x* is dataset values. Obviously, the hyperplane is determined by *w* and *b***[15]**. The data points nearest to the hyperplane are called support vectors. In the linear case, *w* can be solved by introducing Lagrangian multiplier αiαi. The solution of *w* can be written as:

w=∑i=1mαiyixi,w=∑i=1mαiyixi,

where *m* is the number of support vectors and yiyi are target labels to *x*. The linear discriminant function can be written as:

g(x)=sgn(∑i=1mαiyixTix+b),g(x)=sgn(∑i=1mαiyixiTx+b),

s*gn* is the sign function that calculates the sign of a number, sgn(x)=−1sgn(x)=−1 if x<0x<0, sgn(x)=0sgn(x)=0 if x=0x=0, sgn(x)=1sgn(x)=1 if x>0x>0. The nonlinear separation of data set is performed by using a kernel function. The discriminant function can be written as:

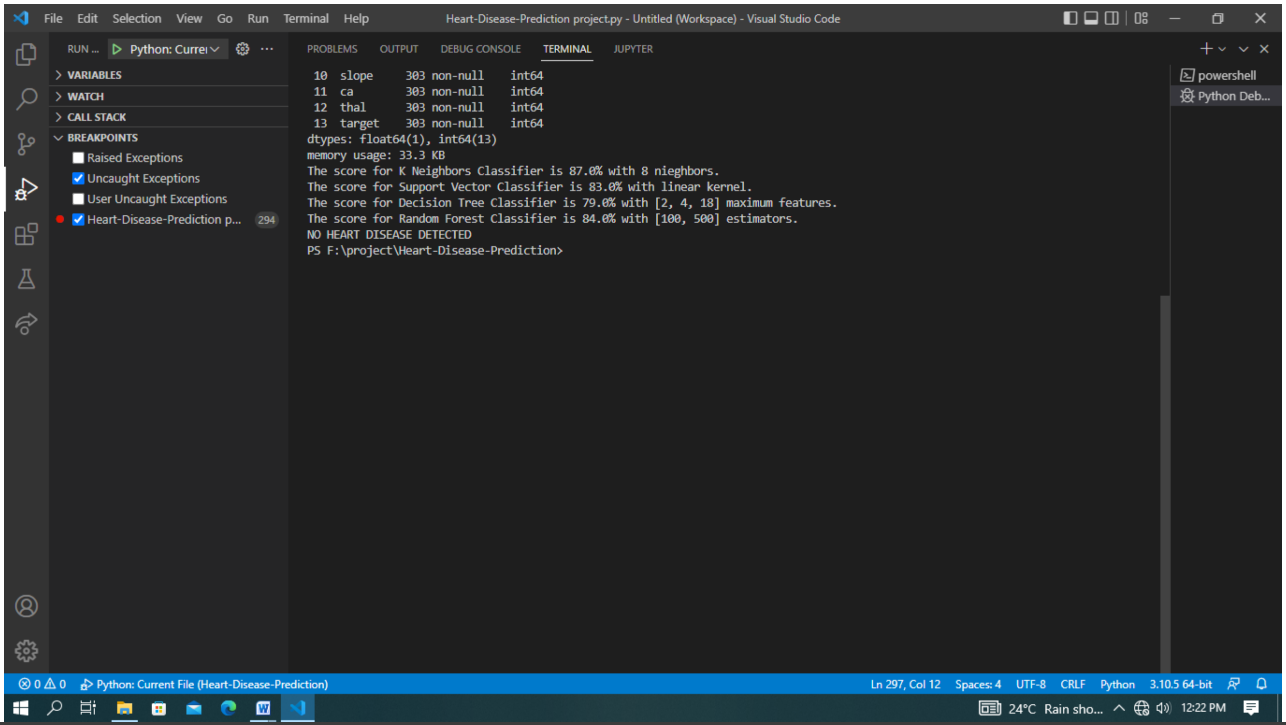
g(x)=sgn(∑i=1mαiyiK(xi,x)+b),g(x)=sgn(∑i=1mαiyiK(xi,x)+b),

where K(xi,x)K(xi,x) is the kernel function**[16]**.

3.2.8 PART

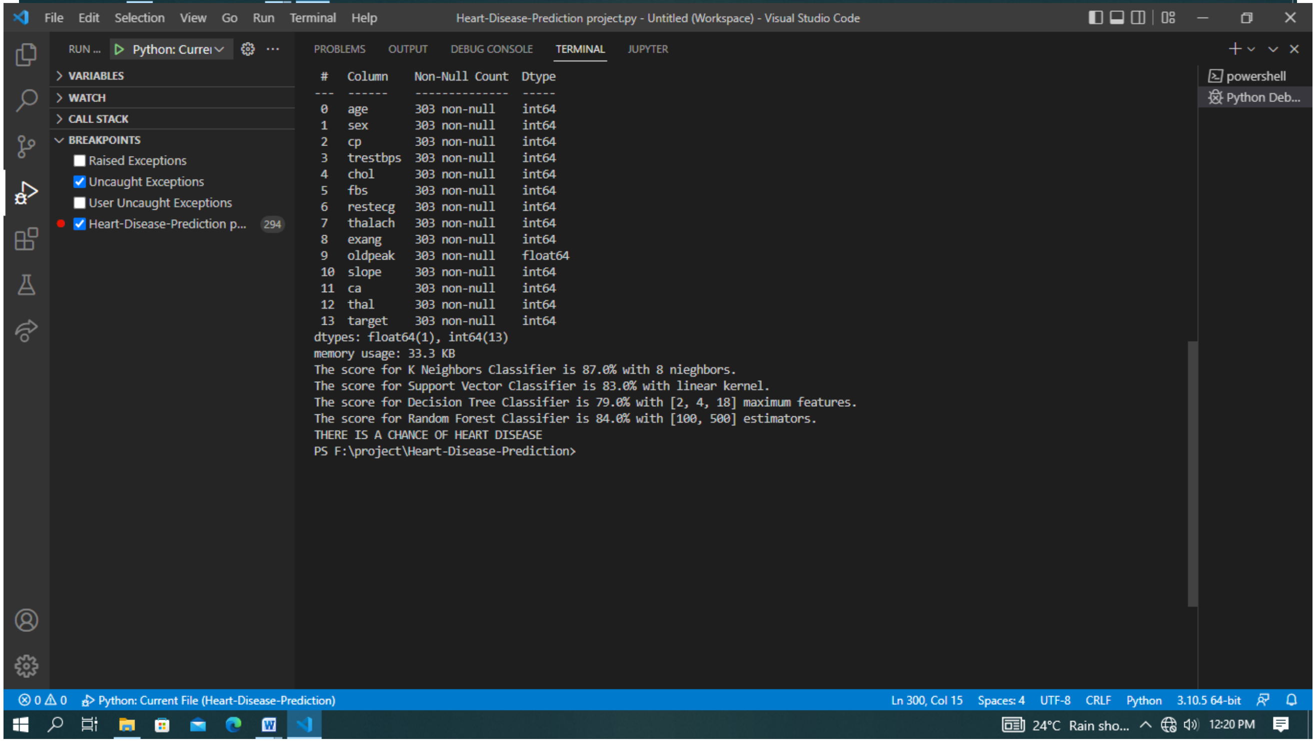
PART is the acronym for Projective Adaptive Resonance Theory. PART is a rule-based classification algorithm. It is a neural network developed by Cao and Wu. It is an advanced version of the C4.5 and RIPPER algorithms. The PART algorithm is suitable for high dimensional datasets. The key feature of the PART network lies is the presence of a hidden layer of neurons, which calculate the variations between the output and input neurons, and work on reducing the similarity differences**[17]**.

3.2.9 Output



File Edit Selection View Go Run Terminal Help
Heart-Disease-Prediction project.py - Untitled (Workspace) - Visual Studio Code
RUN ... Python: Curre
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
VARIABLES
WATCH
CALL STACK
BREAKPOINTS
Raised Exceptions
Uncaught Exceptions
User Uncaught Exceptions
Heart-Disease-Prediction p... 294
powershell
Python Deb...
⊗ 0 ⚠ 0 Python: Current File (Heart-Disease-Prediction)
Ln 297, Col 12 Spaces: 4 UTF-8 CRLF Python 3.10.5 64-bit
24°C Rain sho... 12:22 PM

```
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
The score for K Neighbors Classifier is 87.0% with 8 nieghbors.
The score for Support Vector Classifier is 83.0% with linear kernel.
The score for Decision Tree Classifier is 79.0% with [2, 4, 18] maximum features.
The score for Random Forest Classifier is 84.0% with [100, 500] estimators.
NO HEART DISEASE DETECTED
PS F:\project\Heart-Disease-Prediction>
```



File Edit Selection View Go Run Terminal Help
Heart-Disease-Prediction project.py - Untitled (Workspace) - Visual Studio Code
RUN ... Python: Curre
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
VARIABLES
WATCH
CALL STACK
BREAKPOINTS
Raised Exceptions
Uncaught Exceptions
User Uncaught Exceptions
Heart-Disease-Prediction p... 294
powershell
Python Deb...
⊗ 0 ⚠ 0 Python: Current File (Heart-Disease-Prediction)
Ln 300, Col 15 Spaces: 4 UTF-8 CRLF Python 3.10.5 64-bit
24°C Rain sho... 12:20 PM

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
The score for K Neighbors Classifier is 87.0% with 8 nieghbors.
The score for Support Vector Classifier is 83.0% with linear kernel.
The score for Decision Tree Classifier is 79.0% with [2, 4, 18] maximum features.
The score for Random Forest Classifier is 84.0% with [100, 500] estimators.
THERE IS A CHANCE OF HEART DISEASE
PS F:\project\Heart-Disease-Prediction>
```

3.2.10 Testing and evaluating algorithms

Testing an algorithm

One way to test short programs is to do what is known as a dry run using paper. A dry run involves creating a trace table, containing all the variables a program contains. Whenever the value of a variable changes, the change is indicated in the trace table.

Trace tables help a programmer to determine the point in a program or algorithm where an error has occurred.

Consider this simple program:

```
1 total is integer
2 number is integer
3 set total = 0
4 for count = 1 to 3
5 input “Enter number”, number
6 total = total + number
7 next count
8 output total
```

Each instruction has been given a line number (1-8). The program has three variables - total, count and number. These variables will be put into a trace table.

Instruction	total	count	number

Next, the program is tested using test data. If the numbers 5, 7 and 9 are input, the resulting total should be 21.

The instruction number is entered in the table. If a variable changes with that instruction, the new variable value is written in the appropriate box, as follows:

Instruction	total	count	number
1			
2			
3	0		

Instruction	total	count	number
4		1	
5			5
6	5		
7			
4		2	
5			7
6	12		
7			
4		3	
5			9
6	21		
7			
8	21		

At each step, the programmer is able to see if, and how, a variable is affected.

Trace tables are extremely useful because they enable a programmer to compare what the value of each variable should be against what a program actually produces. Where the two differ is the point in the program where a logic error has occurred.

Evaluating an algorithm

When developing an algorithm it is useful to formalise exactly what that algorithm is designed to do. This makes it possible to check whether it meets all of those requirements. In addition, there might be several possible algorithms for the same problem. A good algorithm will not only do what it is supposed to do but will also do it efficiently. Inefficient algorithms may slow down the speed of the program execution. For example, a bubble sort is less efficient than a merge sort.

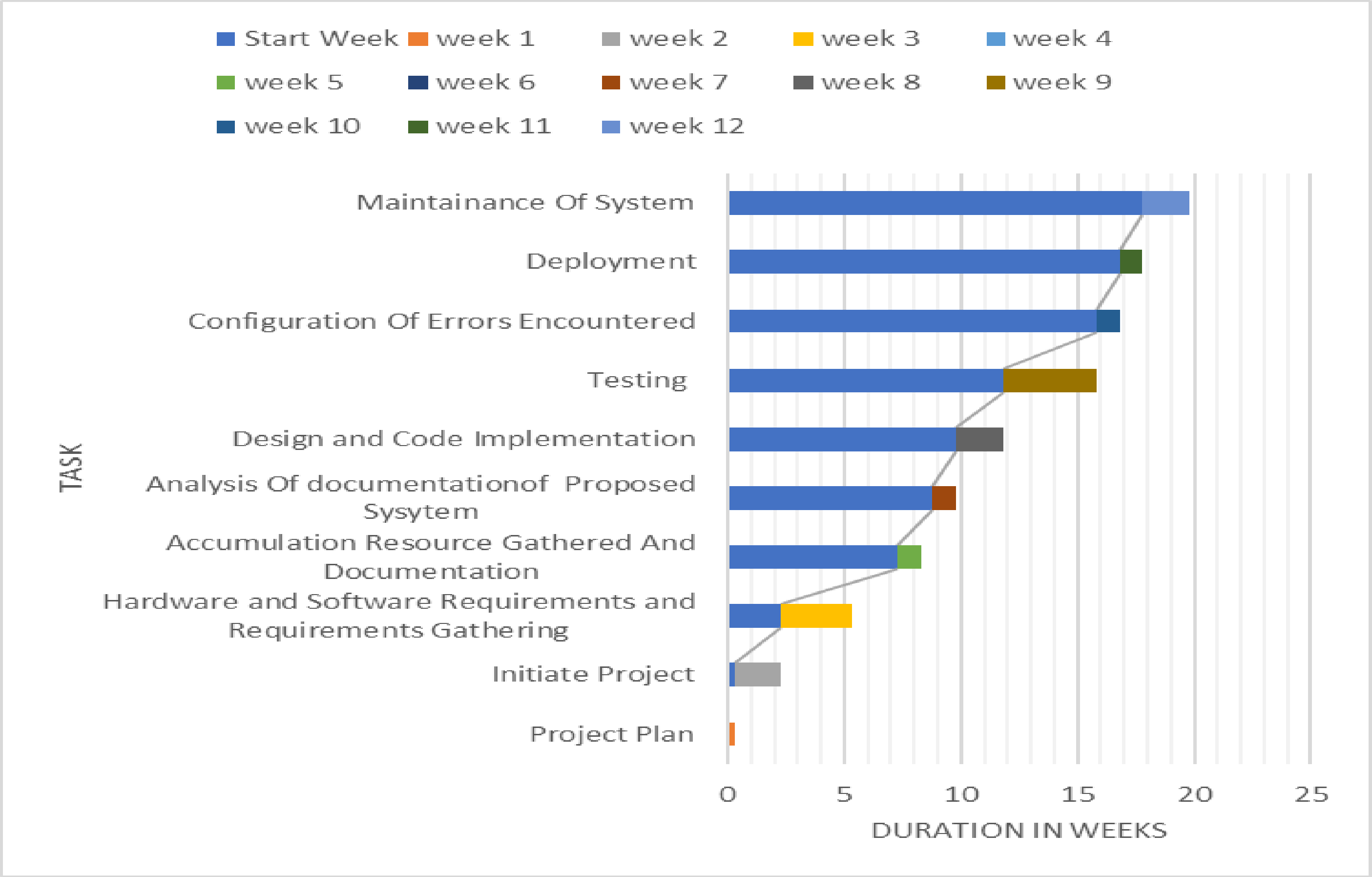
Evaluation is the process that allows us to make sure the solution does the job it has been designed to do efficiently and to think about how it could be improved.

Ways that algorithms may be faulty

Algorithms may fail because:

- the original problem may not be fully understood
- the algorithm is incomplete
- the algorithm is inefficient and may be too complicated or too long
- the algorithm does not meet the original design criteria, so it is not fit for purpose

3.4 Work Plan



3.6 Conclusion And Future Work.

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.

References.

1. A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–281. https://doi.org/10.15680/IJIRSET.2016.0505545

2. Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. *International Journal of Pure and Applied Mathematics*, *118*(Special Issue 8), 165–173. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

3. Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

4. Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. *Oriental Journal of Computer Science and Technology*, *10*(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

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

6. Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In *Procedia Computer Science* (Vol. 85, pp. 962–969).

7. https://www.kaggle.com/ronitf/heart-disease-uci

8. Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. *Ijcsc 0973-7391*, *7*, 1–9. Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. *Heart Disease*, *3*(6), 2385–2392.

9. Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In *2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016* (pp. 1–5). https://doi.org/10.1109/CEEICT.2016.7873142

10. Artificial, I. T. O. (1995). Chapter 2. *Neuron*, 36–62. https://doi.org/10.1109/ETD.1995.403491 Bozzo, R., Conca, A., & Marangon, F. (2014). 11.

11. Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. *Transportation Research Procedia*, *3*(July), 518–527. https://doi.org/10.1016/j.trpro.2014.10.033

12. Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. *ICT Innovations 2012 Web Proceedings*, 577–582.

13. David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques, *6956*(October), 1817–1823. https://doi.org/10.21917/ijsc.2018.0253

14. Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Backpropagation neural network versus logistic regression in heart disease classification. *Advances in Intelligent Systems and Computing*, *702*, 133–144. https://doi.org/10.1007/978-981-13-0680-8_13

15. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. *Indian Journal of Science and Technology*, *9*(39), 1291–1293. https://doi.org/10.17485/ijst/2016/v9i39/102078

16. Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. *International Journal of Advanced Research in Computer Engineering & Technology*, *5*(7), 2278–1323.

17. Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. *Journal of Physics: Conference Series*, *1000*(1).