ANALYSIS OF DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION USING MACHINE LEARNING

BY

ATIYA SHOBNOM
ID: 213-25-060

This report presented in partial fulfillment the requirements of the Degree of Master of Science in Computer Science and Engineering.

Supervised By

Shah Md.Tanvir Siddiquee
Assistant Professor,
Department of CSE
Daffodil International University

Co-Supervised By

Abdus Sattar
Assistant Professor & Coordinator-MSCSE
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2022

APPROVAL

This Thesis titled **"Analysis of Data Mining Techniques for Heart Disease Prediction Using Machine Learning"**, submitted by **Atiya Shobnom**, ID No: **213-25-060** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 19-01-2021.

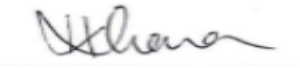
We have examined this report and recommended its acceptance:

BOARD OF EXAMINERS

Chairman

Dr. S M Aminul Haque, PhD
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Ms. Most. Hasna Hena
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Mr. Md. Abbas Ali Khan
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

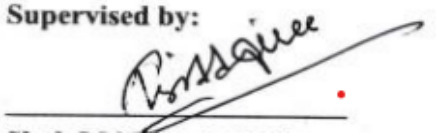
External Examiner

Dr. Mohammad Shorif Uddin, PhD
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

The project is done under the supervision of **Shah Md.Tanvir Siddiquee, Assistant Professor,** Department of Computer Science and Engineering, Daffodil International University.We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Shah Md.Tanvir Siddiquee
Assistant Professor,
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Abdus Sattar
Assistant Professor & Coordinator-MSCSE
Department of CSE
Daffodil International University

Submitted by:

Atiya Shobnom
ID: -213-25-060
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and thankfulness to almighty God for His godly blessing makes us possible to complete the final year thesis successfully.

We really thankful and wish our profound our debt to Shah Md.Tanvir Siddiquee, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our director in the field of "Analysis of Data Mining Technique for Heart Disease Prediction Using Machine Learning" to carry out this design. His endless forbearance, scholarly guidance, continual goad, constant and energetic supervision, constructive review, precious advice, reading multitudinous inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to teachers and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate Daffodil International University, who took part in this bandy while completing the course work.

Ultimately, we must admit with due respect the constant support and cases of our parents.

The Author

Abstract

A Heart Complaint or Disease (HC) is a fatal experience that stops blood flow from supplying oxygen to the heart muscle, that's when a heart attack occurs .There are no noticeable symptoms before a heart attack occurs, which is why it is so important to detect the early symptoms correctly. Through this thesis work an attempt has been made to predict heart attack by reviewing patient history and initial symptoms based on various findings. Here the important information is collected from two popular & reputed medical college name as Ziaur Rahman Medical & TMSS Medical College for searching based on different attributes divided into 1212 rows and 14 different column. Also various classifier algorithm has been use to predict the causes of observed symptoms. Significant risk factors for heart disease were identified using the first Hunt method of the dataset to be explored. LR, SVM, DT, RF,KNN Algorithms are prognosticate IHC using the most significant 13 threat factors. The vaticination accuracy of heart disease is shown from 81.89% to 96.30% with various classifier algorithms. Where the accuracy efficiency of Random Forest is 96.30%.

Keywords: Classifier,Heart Complaint, Accuracy, KNN(K-Nearest Neighbor Algorithm), LR(Logistic Regression), SVM(Support Vector Machine), DT(Decision Tree),RF(Random Forest).

Table of Contents

CHAPTER 3

3.THEORETICAL FRAMEWORK

CHAPTER 4

4. DESIGN AND IMPLEMENTATION

CHAPTER 5

5. RESULTS AND DISCUSSIONS

CHAPTER 6

6. LIMITATIONS AND FUTURE WORKS

CHAPTER 7

7. Conclusion

List of Figure

List of table

CHAPTER 1
INTRODUCTION

In this chapter we will introduce our thesis background, objectives of the study, motivation, expected outcome and assumption. In section 1.1 we will discuss the background, in section 1.2 we will discuss objectives of the study, in section 1.3 in section we will discuss motivation, in 1.4 we will discuss expected outcome, in section 1.5 we will discuss assumption.

1.1 Background

A Heart Complaint (HC) is a fatal experience that stops blood flow from supplying oxygen to the heart muscle, that cuts off the oxygen supply to the heart muscle from the blood, resulting in a heart attack. When the oxygenated blood is reduce to the heart muscles then it is block by fat, cholesterol and other substances. Heart attack is a common causes of death worldwide. Coronary arteries transport blood but no blood pressure is felt by the heart muscle. Most heart complaints are due to atherosclerosis. Usually, present when the through-way lumen appears normal by angiography. It is damages the heart muscles by reducing the supply of blood.

Coronary heart disease mortality data distributed in Bangladesh in 2020 as indicated by the most recent WHO or 15.16% of total deaths. According to the American Heart Association (AHA), A person in the United States suffers a heart attack every 40 seconds, which is really scary . Bangladesh is #118 in the world at 94.27 age-adjusted mortality per population [1].Currently, Coronary Arteries (CA) in Bangladesh is one of the most problems in the healthcare sector as 1.06 lakh deaths in 2019 were due to ischemic heart complaints, lakhs due to stroke and hypertensive heart complaints. The common reason for this condition is the eating habits of Bangladeshis. They prefer to eat highly oleaginous, rich food and lack exercise habits although many of them are smokers [2]. Unfortunately, we don't get any warning before a heart attack because this process (atherosclerosis) has no serious symptoms.

According to a new report, about 51 of U.S. Adults are affected by certain types of cardiovascular disease (CD) and diabolic blood pressure, also known as hypertension, which can lead to heart failure and stroke [3].

1.2 Objectives

Objectives of thesis to find out following significant threat factors for heart complaint in Bangladeshi cases and predict HC using different bracketing methods. Overall, the objects of this thesis can be described as follows

1. To describe the important threat factors of heart complaint.
2. To identify correlation between heart disease and threat factor.
3. To prognosticate Ischemic Heart Complaint (HC) using substantially significant threat factors by applying different bracket algorithms.

1.3 Motivation

People all around the world many people are affected many kinds of heart problem.
About 150 millions peoples are died because of many kinds of heart problem.

1.4 Expected Outcome

In today's world there are many advance technologies for detecting heart problem. But our expectation is to create a system by using some machine learning algorithms so that a person can know in advance (with higher accuracy) whether he/she has heart disease or not we are using Patient Health Questioner (PHQ) for detecting he/she is has heart disease or not. First, we collect people data from field and they we use this dataset in our proposed model. When a user gives his/her data for knowing he/she has heart disease or not our proposed model analysis on it and then it answers that he/She has heart disease or not.

1.5 Assumption

The following parameters are assumed for the effective run of the proposed system:

- ✓ User input data of required questionnaire.
- ✓ Dataset follows the format (.csv).
- ✓ The rows correspond to people's data (Except Age and Gender) data and the columns correspond to the sample/experiment/patient.
- ✓ It is assumed that training & testing dataset is free from missing value.

CHAPTER 2
LITERATURE REVIEW

From our previous chapter we have already learn about the introduction of our thesis. In this chapter we will discuss about related study of our thesis. Various machine learning algorithms have resulted in many applications in heart problem detection. In many research projects, several classifiers and data mining models have been used to target appropriate classifications such as heart disease detection, autism detection, etc.In section 2.1 we will discuss heart problem detection's many approaches. In section 2.2 we will discuss some works that are related to some classifiers and data mining models.

2.1 Heart Problem Detection

A number of researchers have applied various machine learning methods to detect swish results and identify the causes of new problems that cause various non-communicable conditions such as heart attacks. A machine learning algorithm was proposed to predict diabetic disease using Naive Bayes and Decision Tree algorithms. An exploratory study on the correlation of association rules and their operation was carried out on coronary heart disease data.They performed the study on two various datasets, American drug and Western drug order 3 videlicet, and the study generated 6 rules for American drug while generating 7 rules for Western drug order. Three algorithms namely RF,NB and SVM used voting style to predict complaints and the precision for Random-Forest (RF) was 87.50, NB (Naive Bayes) and SVM achieved the same precision of 85.11. Shaheed Ziaur Rahman Medical College dataset contains 303 cases with 74 different characteristics but only 14 of them like age, sex, blood sugar serum cholesterol, type of casket pain, maximum heart rate, resting blood pressure, exercise etc. Confirmed angina, etc.Taken as prescribed by other publishers. The precision of this system ranges from 86.66 to 95.55 minimum for hidden estates of different sizes. SVM and gain rate methods were used for point selection and RF and NB were used to predict HD and the dataset used for this search was "age", "trestbps", "chol", "cp", "fbs" , "restecg", "thalach","exang", "oldpeak", "thal", "num", "ca", "pitch". A trial was used for approximately 1212 cases with 800 tested negatives and 412 tested cons, and the missing values replacement function in the WEKA tool was used to handle missing data.

2.1.1 Using Patient Health Questionnaire

PHQ 9 is a tool through which every area of mental health is assessed in the screening process. B. Yalamanchili et al. [4] they are using Patient Health Questionnaire (PHQ-8) system for detect depression. Firstly, they are collecting some depressed and non-depressed people's data they are analysis on it after analysis on it they build a model for detecting depression.

2.2Machine Learning Algorithm

Machine learning has been around since 1959. Machine learning algorithms are a part of AI. It learns itself based on machine learning algorithms. A machine is trained using machine learning algorithms to automatically run itself as a human being and in the future can itself make a decision operation from fast data or experience.

2.2.1 Random Forest

Random-Forest a popular classification algorithm which consists of many decision trees. Let's say there are n numbers of data in a database and n numbers of single tree are formed with the data from there. The random forest is basically constructed from the majority of the output from each single decision tree.

2.2.2 Naive Bayes

A Naive Bayes classifier is a classifier that predict the probability of an earlier event relative to the probability of a later event.

$$p(c|x) = \frac{p(x|c).p(c)}{p(x)})$$

Above,

- p(c|x) is the posterior probability of class given predictor.
- p(c) is the prior probability of class.
- p(x|c) probability of predictor given class.
- p(x) is the prior probability of predictor.

2.2.3 Logistic Regression

The three types of learning systems in machine learning are supervised learning, unsupervised learning and reinforcement learning. Logistic regression is a supervised learning-based classification algorithm. Two types of regression are logistic regression and linear equation. Discrete values are obtained as output through logistic regression. There are three types of logistic regression namely Binary logistic regression, Multinational logistic regression, Ordinal logistic regression. Binary logistic regression will predict the output in 0/1 or true/false form. If the predicted output follows a promo it is called ordinal logistic regression. Logistic regression is similar to the S shape, The logistic function can be drawn using the sigmoid function as:$f(x) = \frac{1}{1+e^{-x}}$

2.2.4 Decision Tree

Decision Tree is a method of data mining technique which helps us to take decision easily. It has a specific structure through which the data is divided into small sub parts. One part of which is called root node and other part is called leaf node. Decision tree can operate of numerical and categorical data.

CHAPTER 3

THEORETICAL FRAMEWORK

From our previous chapter we have learnt about the related literature of our thesis. In this chapter we will study the theoretical framework of our thesis. In section 3.1 we will discuss the definition of terms; in section 3.2 we will discuss types of heart problem . There are many kinds of heart problem we will discuss it in this section; in section 3.3 we will discuss about causes of heart disease ; in section 3.4 we will discuss about complication of heart diseases; in section 3.5 we will discuss about patient health questioner; in section 3.6 we will discuss about machine learning. In section 3.7 we will discuss about Naive Bayes; in section 3.8 we will discuss about Logistic Regression. In section no. 3.9 we will discuss to Decision Tree and section no. 3.10 we will discuss about Random Forest; in section 3.11 we will discuss about Performance Evaluation.

3.1 Definition of terms

Heart disease is a kind of disease which affect the blood vessels or heart. Risk of heart disease to grow by high blood pressure, high cholesterol, smoking, diabetics, unhealthy diet and obesity [5].

3.2 Types of Heart Complaints

The different types of heart complaint that are given below:

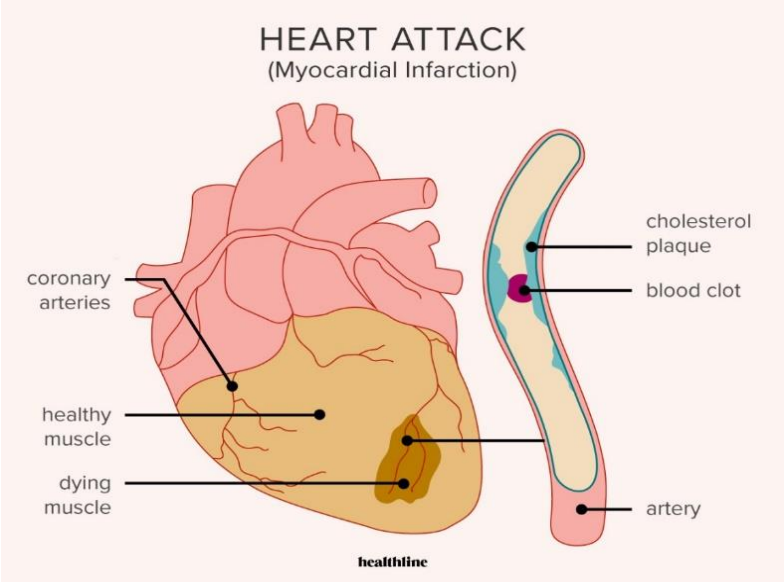


Fig 3.2: Types of Heart Attack

3.3 Causes of Heart Complains

Many causes are responsible for heart complaint like:

- Cholesterol level .
- BP high.
- Diabetes.
- Blood pressure high during pregnancy.
- Drug abuse.
- Smoking.
- Stress.
- Congenital heart defect etc.

3.4 Complications of Heart Disease

Heart Failure: Heart failure means disruption of the normal functioning of the heart. Heart is a pumping muscular organ that circulates blood [5]. The heart has two main functions. Heart takes the contaminated blood from the body and pumps that blood to different parts of the body. When these two functions are hampered, we call the situation Heart failure.

Heart Attack: An adult's heart beats seventy-two times per minute. Through this beat, the heart pumps blood throughout the body. This blood circulation takes place from its two blood vessels called coronary arteries. If cholesterol or fat accumulates in those blood vessels, blood cannot flow normally, then the muscle of the heart suffers from anemia, blood clots and a heart attack occurs.

Stroke: A stroke occurs when blood flow to the brain stops. For some reason, the blood vessel in the brain ruptures and the excess blood clots, then the blood flow to the brain is blocked and this condition is called stroke. Strokes occur for the same reasons that cause heart attacks.

3.5 primary Symptoms of Heart Attack:

- Severe pain with pressure or twisting in the chest This pain spreads to other parts of the body Chest pain comes in both arms Pain is felt more in the left arm.
- Jaw, neck-back, and stomach pain is felt.
- Head feels lightheaded and sweats
- Can't breathe fully
- Nausea and cough
- Feeling restless or panicked, body sweats profusely.

3.6 Patient Health Questioner

Important problems are identified based on information from patient questions.Firstly the system asks some question like as a doctor. After completing all question, the patient will give some answer based on his/her answer the system detect that this testing person is affected in heart problem or not.

3.7 Machine Learning

Machine learning has been around since 1959. Machine learning algorithms are a part of AI. It learns itself based on machine learning algorithms [5]. A machine is trained using machine learning algorithms to automatically run itself as a human being and in the future can itself make a decision operation from fast data or experience.

There are three main branches of machine learning algorithms -

1. Ability to process decisions.
2. Error function used for comparing accuracy.
3. Ability to perform model optimization.

Feature of Machine Learning:

- ✓ Machine learning algorithms can easily test, training and finally predicts through pattern recognition.
- ✓ Machine learning algorithms do not require human intervention
- ✓ Managing multivariate dataset.
- ✓ The application of machine learning algorithms are wide ranging.

Importance of Machine Learning: It is a branch of AI whose ideas are being used in all fields such as finance, healthcare, marketing, social sites, cyber security and many other fields [5]. Currently big data analysis ,data extraction and other tasks can be done with machine learning.

3.7.1 How does Machine Learning Works

There are three main branches of machine learning algorithms works -

1. Ability to process decisions.
2. Error function used for comparing accuracy.
3. Ability to perform model optimization.

- **Ability to process decisions:** Machines can learn automatically after training and testing datasets during decision processing.
- **Error function used for comparing accuracy:** Perdition can be done by comparing the accuracy using the error function.
- **Ability to perform model optimization:** Machine learning algorithm repeats the optimization process and evolution hence gives better performance.

3.7.2 Supervised Machine Learning

One of the most important types of machine learning is the supervised machine learning algorithm. A machine is told what to do with all the features in a data set, the machine is given a prediction based on those features and the machine can make a decision accordingly.

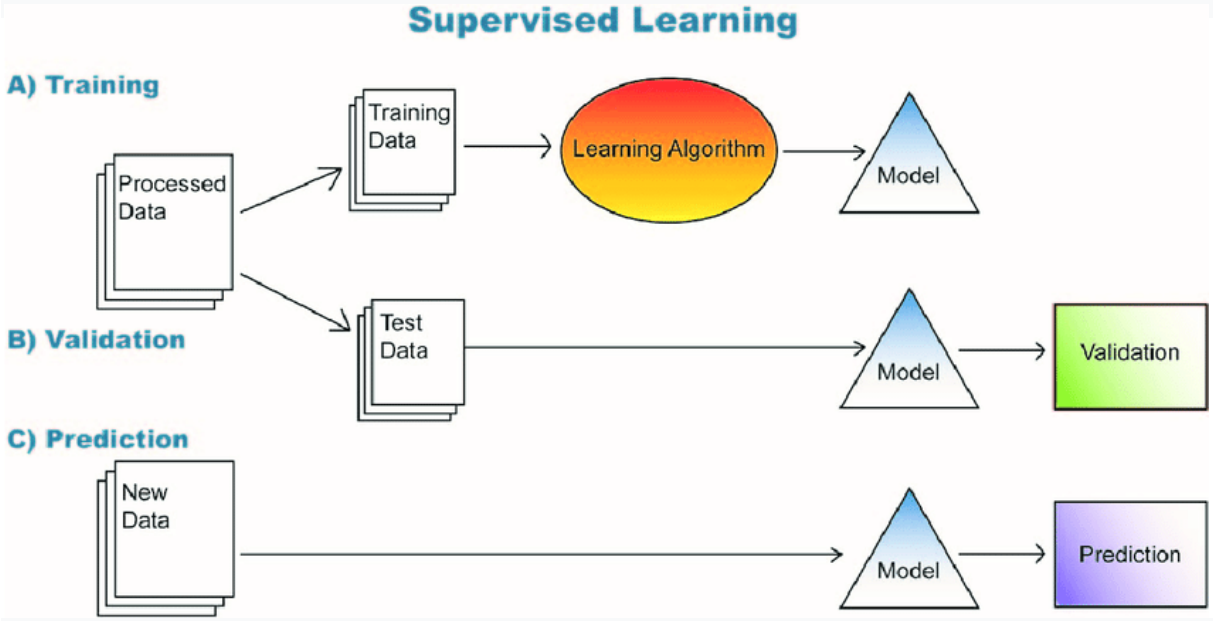


Fig 3.7.2 : Supervised Machine Learning

3.7.3 Unsupervised Machine Learning

Unsupervised learning algorithm is another type of machine learning algorithm. In this algorithm, decisions are not made based on the characteristics of dataset, but only the data. Unsupervised Machine Learning has no idea except data and algorithm.

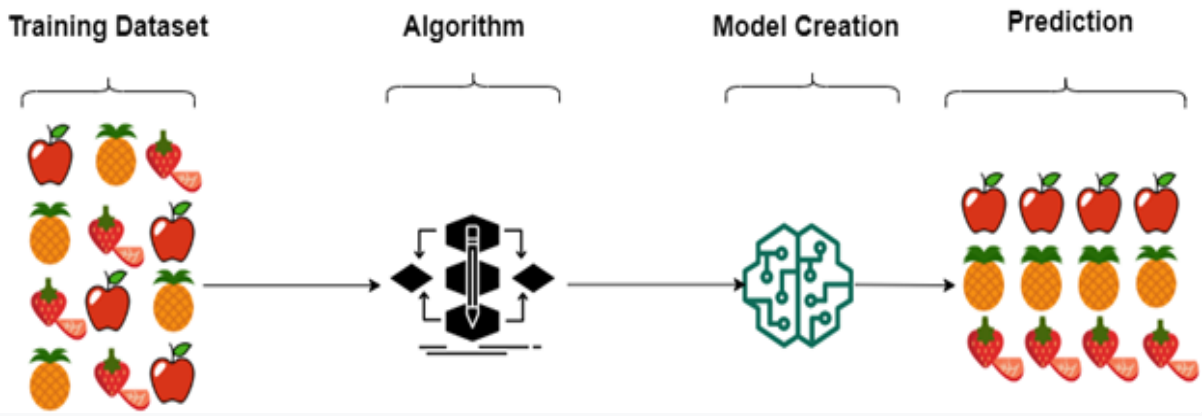


Fig 3.7.3: Unsupervised Machine Learning Example

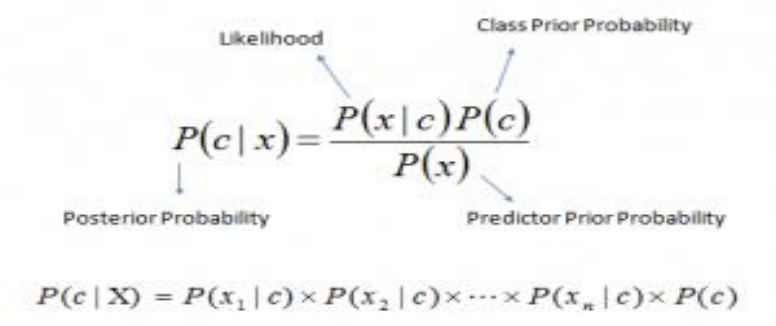
3.8 Naive Bayes

A Naive Bayes classifier is a classifier that predict the probability of an earlier event relative to the probability of a later event.

$$p(c|x) = \frac{p(x|c).p(c)}{p(x)})$$

Above,

- p(c|x) is the posterior probability of class given predictor.
- p(c) is the prior probability of class.
- p(x|c) probability of predictor given class.
- p(x) is the prior probability of predictor.



Circuit Learn has three types of libraries for the Naive Bayes classifier algorithm.

- Gaussian NB
- Multinomial NB
- Bernoulli NB

- **Gaussian NB:** A Gaussian NB is used for classification and assumption whose features follow a normal distribution.
- **Multinomial**: It is used for multinomial distribution of data. The primary task of Naive Bayes classifier algorithm is to classify any particular documents into specific branches.
- **Bernoulli:** Bernoulli classifier algorithm takes independent Boolean variables. It works like polynomial classifier. It is a model suitable for classification.

3.9 Logistic Regression

The three types of learning systems in machine learning are supervised learning, unsupervised learning and reinforcement learning. Logistic regression is a supervised learning-based classification algorithm. [6] Two types of regression are logistic regression and linear equation. Discrete values are obtained as output through logistic regression. There are three types of logistic regression namely Binary logistic regression, Multinational logistic regression, Ordinal logistic regression. Binary logistic regression will predict the output in 0/1 or true/false form. If the predicted output follows a promo it is called ordinal logistic regression. Logistic regression is similar to the S shape, The logistic function can be drawn using the sigmoid function as: $f(x) = \frac{1}{1+e^{-x}}$

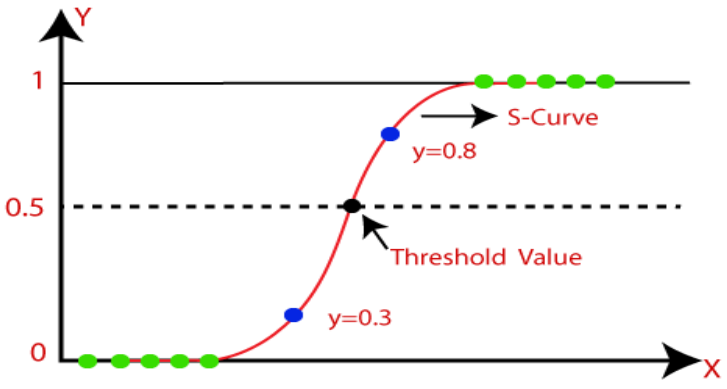


Fig 3.8: Logistic Regression Curve

Equation of Logistic Regression:

We know, the straight line equation is:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots\ldots\ldots\ldots + b_nx_n$$

- ✓ Here, y can be between 0 and 1 only, so it divided the above equation by (1-y):

$$\frac{y}{1-y}; 0 for\ y = 0, and\ infinity\ for\ y = 1$$

- ✓ so, the range is [infinity] to +[infinity],and the equation will be:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots\ldots\ldots\ldots + b_nx_n$$

3.10 Decision Tree

Decision Tree is a method of data mining technique which helps us to take decision easily. It has a specific structure through which the data is divided into small sub parts. One part of which is called root node and other part is called leaf node. Decision tree can operate of numerical and categorical data.

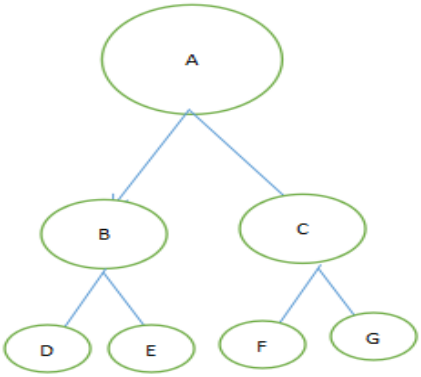


Fig 3.9: Decision tree

Decision tree algorithm look like a large configuration but quite simple technique:

This algorithm works based on few parameters. They are refers to the training tuples. Attribute list is refers to the tuple sets. A heuristic process is chosen by Attribute Selection methods.

Advantages of Decision tree:

- ✓ A decision tree does not require data scaling.
- ✓ Missing values in the data also do not significantly affect the process of generating the choice tree
- ✓ A decision tree model is automated and easy to explain to technical teams and stakeholders.
- ✓ Compared to other algorithms, decision trees require less effort in data preparation during pre-processing.

3.11 Random Forest

Random-Forest is a popular classification technique which consists of many decision trees. Let's say there are n numbers of data in a database and n numbers of single tree are formed with the data from there. The random forest is basically constructed from the majority of the output from each single decision tree.

For example, after training the data by forming N number of single trees from a website, the majority of decision trees can be used to generate random forest decisions. While training the data, randomly generated trees will be value-added for each spouse in the testing period. Single Decision Free has high probability of error while Rainforest has low probability of error because of variance in Rainforest algorithm. Firstly, by creating several learner's models from the Random Forest algorithm based on two factors:

1. Raw Sampling with Replacement.
2. Feature Sampling With Replacement.

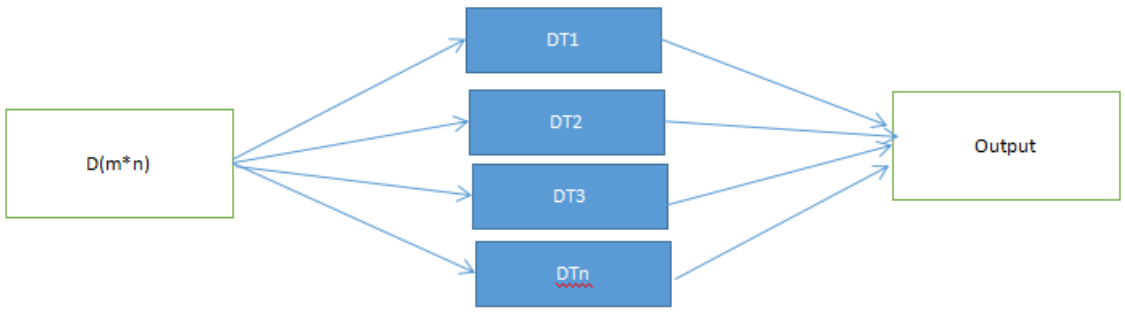


Fig 3.10: Random Forest

3.12 Support Vector Machine

Support vector machine learning algorithm is an important classifier. Support Vector Machine is abbreviated as SVM. It is a supervised machine learning algorithm used for regression or classification problems. The goal of SVM is to create a best line of classification in n-dimensional space. That the new data can be placed in the right place and the decision boundary is called the hyper-plane. A Support Vector Machine or SVM helps in creating hyper-plane.

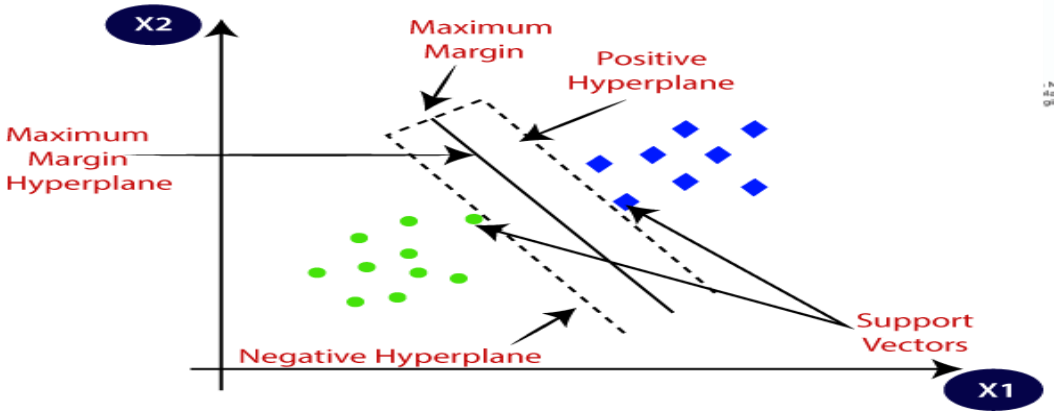


Fig 3.12: Support Vector Machine

3.13 Performance Evaluation

In our proposed models we have done k-fold cross-validation and calculate confusion matrix for performance evaluation. So firstly, we will know about k-fold cross-validation and then we calculate confusion matrix.

K-Fold Cross Validation:

Cross validation is a model of machine learning algorithms that can evaluate a small number of iterations in a resembling process [6]. Here k is an argument expressing the number of groups of data to divide by. That is why it is called K- Fold Cross Validation. Let's say that k = 10 means that the specific value of k will be used in reference to the model. That is, there will be 10 fold cross validation.

Rules for performing k-fold cross validation in a general manner:

- ✓ The data included in the data set must be randomized.
- ✓ The dataset should be partitioned into k^{th} partitions.

- To be done separately for each group-
  - ✓ Each part of the data set should be treated as a test dataset.
  - ✓ An additional group should be assumed as the training dataset.
  - ✓ Training datasets are taken into a model and evaluated.
  - ✓ The evaluated results should be saved and the model discarded.

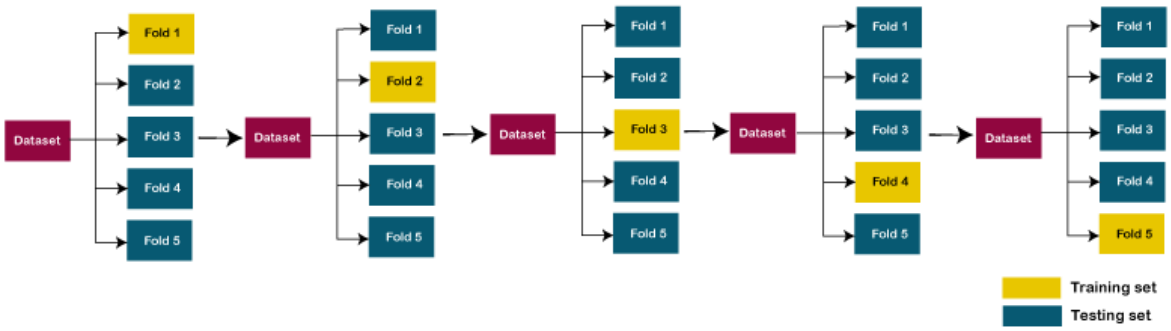


Fig 3.13: K- Fold Cross Validation.

Confusion Matrix:

Confusion matrix is the number of correct and incorrect predictions by dividing the data into categories and expressed briefly. The model shows the ways in which confusion occurs when making predictions and looks for mistakes made by the classifier. This is the key to the confusion matrix.

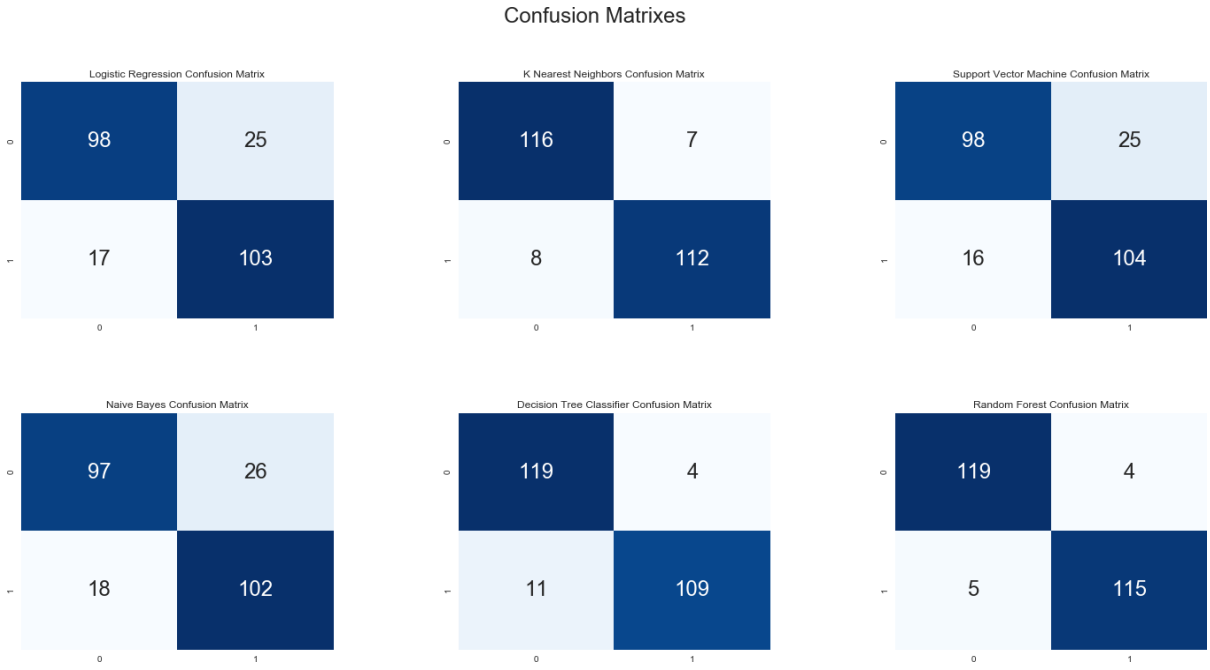


Fig 3.13.1 Confusion Matrix

How to calculate Confusion Matrix:

The process is given below:

- ✓ A valid dataset is required for good results.
- ✓ Have to prepare for every rows in the dataset.
- ✓ Predictions and Expected Outcomes must be calculated.
- ✓ Count the number of predictions for each class,
- ✓ We should organize the categories for each incorrect prediction.
- ✓ These numbers will be in matrix form.

Confusion matrix looks likes that table:

n = Total Prediction	No	Yes
No	True Negative	False Positive
Yes	False Negative	True Positive

Table 01: Confusion Matrix

We can calculate using this formula is given below:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Recall: It is defined as, how our model predicted correctly?

$$Recall = \frac{TP}{TP + FP}$$

F-measure: $\mathrm{F} - measure = \frac{2+recall+precision}{recall+precision}$

CHAPTER 4

DESIGN AND IMPLEMENTATION

From our previous chapter we learned about the theoretical framework of our thesis. In this chapter we discuss the design and implementation of our proposed model. In Section 4.1 we discuss datasets; In Section 4.2 we discuss how to implement our system; In Section 4.3 we will discuss the technical requirements.

4.1 Dataset

Data collected from TMSS Medical Council and Shaheed Ziaur Rahman Medical College, Bangladesh. About 1212 case data are collected with 14 features and added in appendix and contract documents are added in appendix section. Four orders of data were collected. Some individual, some case histories, some demographic and some characteristic data.

This disquisition work database is taken from TMSS Medical College and Shaheed Ziaur Rahman Medical College Repositories. It includes 13 features. The heart complaint dataset included in this disquisition work contains a total of 270 cases with no standard.

The dataset is commonly employed for typical angina, atypical angina and non-anginal pain and a variety of asymptomatic cardiac conditions. This disquisition task is aimed at predicting external heart complaints by type of complaint.

Specialty is a numeric data type that represents the age of the case and ranges from 29 to 65 times. ‘‘cp” is a specialty to determine the type of pain, represented by a range of 1 to 4. “trestbps” is a resting blood pressure between 92 and 100; FBS overestimates the blood sugar position that represents either 1 or 0 boolean values as true or false. “Restecg” is the resting electrocardiographic findings presented in three cases from 0 to 2. “exang” is exercise convinced angina which is a Boolean value. Complaint is the target class in the dataset that indicates the presence of a heart complaint by yes or no Also, all the features and their values are presented in Table 1:

Attribute and Description of Dataset

S. No.	Attribute Name	Type	Description	Range
1	Age	Numeric	Age in years	29-65
2	Sex	Nominal	Sex in number	Male = 0, Female = 1
3	Cp	Nominal	Chest pain type	typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic = 4
4	trestbpd	Numeric	Resting blood pressure	92-200
5	serumCho	Numeric	serum cholesterol in mg/dl	126-564
6	fbs	Nominal	Fasting blood sugar level	Yes =1, No = 0
7	restecg	Nominal	Resting electrocardiographic results	Normal = 0, having ST-T wave abnormality=1, showing probable or definite left ventricular hypertrophy = 2
8	thalach	Numeric	Maximum heart rate achieved	82-185
9	exang	Nominal	Exercise induced angina	Yes = 1, No = 0
10	oldpeak	Numeric	ST depression induced by exercise	71-202
11	peakSlope	Numeric	the slope of the peak exercise ST segment	1-3
12	numVesse ls	Numeric	number of major vessels (0-3) coloured by fluoroscopy	0-3
13	thal	Nominal	The defect type of the heart	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Disease	Nominal	Identification of a heart attack.	Yes=2, No=1

Table 02: Attribute and Description of Dataset

Data Pre-processing:

Data processing is the act of converting data from a given form into a more usable and desirable form [7] .We can pre-process dataset using those method:

- ✓ Data Cleaning/Cleansing. Cleaning "dirty" data.
- ✓ Data Integration.
- ✓ Data Transformation.
- ✓ Data Reduction.

Parameter Tuning:

Parameter tuning needs when there is inconsistency between test data and train data accuracy [33,34]. That means when overfitting/ underfitting occur that time we need to parameter tuning. But in this study, there is no overfitting or underfitting problem so that we no need to parameter tuning.

4.2 System

System means how we are implementing our proposed system. Our system has two section one is implementing and other is user using. In design section we have to describe about how a user can use our system. Using machine learning algorithms, We are developing a model to identify whether a person is suffering from heart problems or not. In our paper we have done label encoding and use this for our three proposed system to detect heart problem using Patient Health Questioner. For evaluation we divided 75% dataset for training and 25% for testing data.

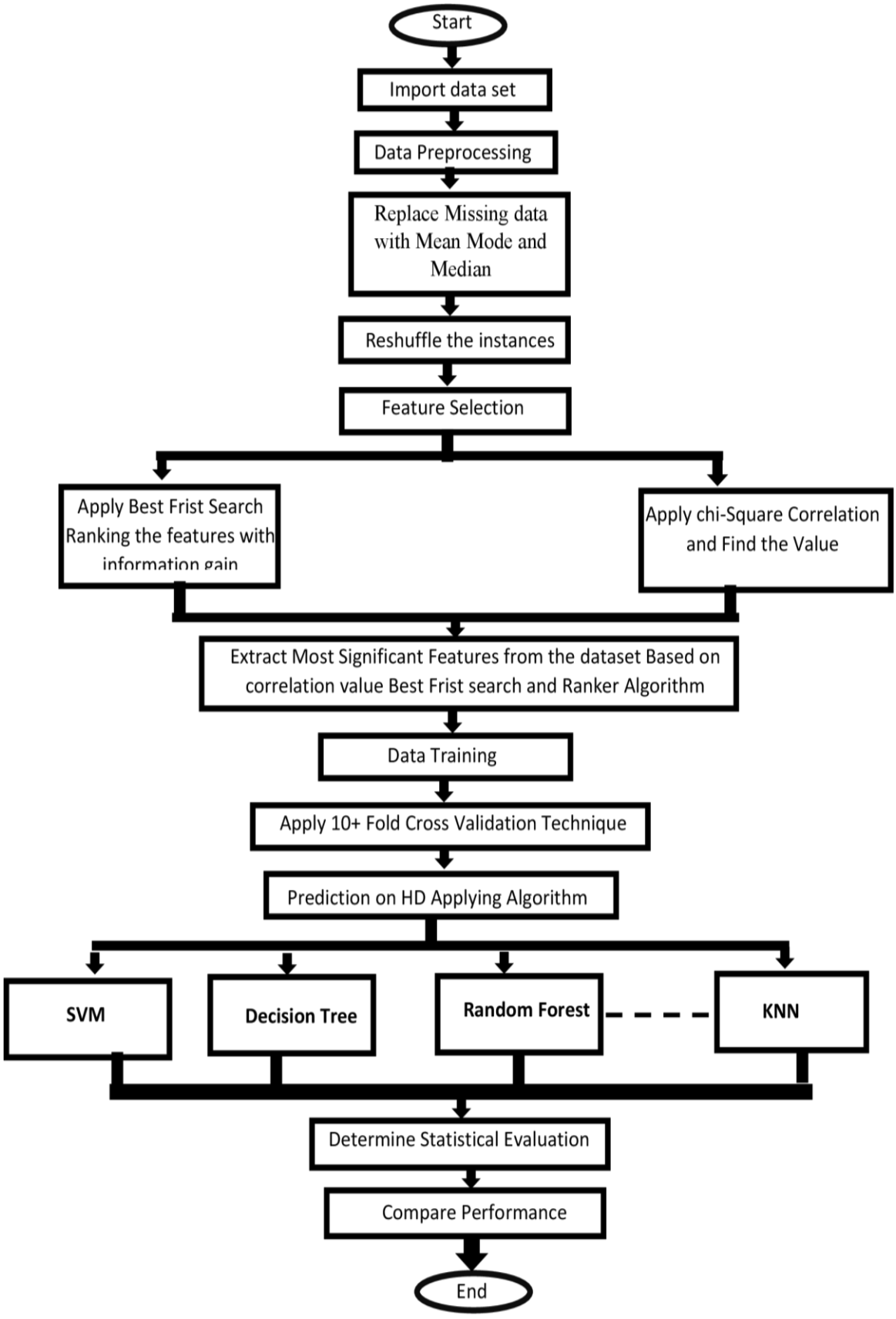


Fig 4.2: Working Flow of the System.

Our main work (model building) is divided into those sections . They are:

- ✓ Step-1: Load the dataset
- ✓ Step-2: Normalized the dataset (Label encoding)
- ✓ Step-3: Split the dataset to 75% of training and 25% of testing randomly.
- ✓ Step-4: Build the model using the training data and train the model.
- ✓ Step-5: Evaluate the model using test data.
- ✓ Step-6: Tuning the parameter & optimize model error.
- ✓ Step-7: Perform k-fold cross validation & calculate Confusion matrix.
- ✓ Step-8: Change the value of k & try to get low error.
- ✓ Step-9: Finally gets the calculated result.

4.2.1 Implementation Steps Using Naive Bayes

Now, we have already a good idea how we have implemented Naive Bayes algorithm in our proposed model. The implementation steps are given bellow:

Step-1: At first load the dataset. In this step we should be careful about the format of the input dataset because our models work with .csv (Comma-Separated Value) format so that dataset must be in .csv format. We have to be careful about load dataset because to build a model dataset is the main factor.

Step-2 : Pre-process the data to convert the labels into a numerical form so that they can be converted into a machine-readable form.

We are collecting data in categorical form so we need to convert it into numerical form. For this reason, we must pre-process our dataset's data.

Step-3: Generate the Naïve Bayes model by using Bayes Theorem. We have already told this in third chapter into Naïve Bayes section. Analysis the summary result to check our input variables are significant to our response variable or not.

Step-4: Split the dataset to 75% of training and 25% of testing randomly.

.

Step-5: Generate the model using training data & trained the model.

Step-6: Evaluate the model using test data. Check the prediction of our test data in order to check the patient is depressed or not. If he/she is depressed then system analysis that which type depression he/she is suffer.

Step-7: Perform K-Fold Cross Validation.

Step-8: Change the values of K and try it gets low error.

Step-9: Finally gets the calculated result which is acceptable.

4.2.2 Implementation Steps Using Logistic Regression

Now, we have already a good idea how we have implemented Logistic Regression algorithm in our proposed model. The implementation steps are given bellow:

Step-1: At first load the dataset. In this step we should be careful about the format of the input dataset because our models work with .csv (Comma-Separated Value) format so that dataset must be in .csv format. We have to be careful about load dataset because to build a model dataset is the main factor.

Step-2 : Pre-process the data to convert the labels into a numerical form so that they can be converted into a machine-readable form.

We are collecting data in categorical form so we need to convert it into numerical form. For this reason, we must pre-process our dataset's data.

Step-3: Generate the Naïve Bayes model by using Bayes Theorem. We have already told this in third chapter into Naïve Bayes section. Analysis the summary result to check our input variables are significant to our response variable or not.

Step-4: Split the dataset to 75% of training and 25% of testing randomly.

.

Step-5: Generate the model using training data & trained the model.

Step-6: Evaluate the model using test data. Check the prediction of our test data in order to check the patient is depressed or not. If he/she is depressed then system analysis that which type depression he/she is suffer.

Step-7: Perform K-Fold Cross Validation.

Step-8: Change the values of K and try it gets low error.

Step-9: Finally gets the calculated result which is acceptable.

4.2.3 Implementation Steps Using Decision Tree

Now, we have already a good idea how we have implemented Decision tree algorithm in our proposed model. The implementation steps are given bellow:

Step-1: At first load the dataset. In this step we should be careful about the format of the input dataset because our models work with .csv (Comma-Separated Value) format so that dataset must be in .csv format. We have to be careful about load dataset because to build a model dataset is the main factor.

Step-2 : Pre-process the data to convert the labels into a numerical form so that they can be converted into a machine-readable form.

We are collecting data in categorical form so we need to convert it into numerical form. For this reason, we must pre-process our dataset's data.

Step-3: Generate the Naïve Bayes model by using Bayes Theorem. We have already told this in third chapter into Naïve Bayes section. Analysis the summary result to check our input variables are significant to our response variable or not.

Step-4: Split the dataset to 75% of training and 25% of testing randomly.

.

Step-5: Generate the model using training data & trained the model.

Step-6: Evaluate the model using test data. Check the prediction of our test data in order to check the patient is depressed or not. If he/she is depressed then system analysis that which type depression he/she is suffer.

Step-7: Perform K-Fold Cross Validation.

Step-8: Change the values of K and try it gets low error.

Step-9: Finally gets the calculated result which is acceptable.

.

4.2.4 Implementation Steps Using Random Forest

Now, we have already a good idea how we have implemented Random Forest algorithm in our proposed model. The implementation steps are given bellow:

Step-1: At first load the dataset. In this step we should be careful about the format of the input dataset because our models work with .csv (Comma-Separated Value) format so that dataset must be in .csv format. We have to be careful about load dataset because to build a model dataset is the main factor.

Step-2 : Pre-process the data to convert the labels into a numerical form so that they can be converted into a machine-readable form.

We are collecting data in categorical form so we need to convert it into numerical form. For this reason, we must pre-process our dataset's data.

Step-3: Generate the Naïve Bayes model by using Bayes Theorem. We have already told this in third chapter into Naïve Bayes section. Analysis the summary result to check our input variables are significant to our response variable or not.

Step-4: Split the dataset to 75% of training and 25% of testing randomly.

.

Step-5: Generate the model using training data & trained the model.

Step-6: Evaluate the model using test data. Check the prediction of our test data in order to check the patient is depressed or not. If he/she is depressed then system analysis that which type depression he/she is suffer.

Step-7: Perform K-Fold Cross Validation.

Step-8: Change the values of K and try it gets low error.

Step-9: Finally gets the calculated result which is acceptable.

4.3 Technical Requirements

Technical requirements, otherwise known as technical specifications or specifications, refer to solutions used by professionals to solve technical problems and software-related problems [8].

4.3.1 Python

Python is a programing language that is used to develop different programming tasks. Generally, python is used to develop software, data analysis, automated tasks etc. In 2022, conducted a survey from RedMonk firm and they ware declared python was the second common favorite programming language to the among developers [9]. We are creating a heart complaints detection system. So we are using python for implementing of our proposed model. For developing our system we have to must using machine learning algorithms and machine learning algorithms can only be using python language. Using python language we can impingement different kinds of machine learning algorithms and we for get good accuracy we have to must using python language. We can developing machine learning algorithms using java but implementing machine learning algorithms using java language using java is very complex. So if we are using python for developing our proposed model it is very easy and safe.

CHAPTER 5

RESULT AND DISCUSSION

From our previous chapter we have learned about the design and implementation of our proposed models. In this chapter we discuss the results of our model.

5.1 Results and Analysis

In our former chapters we have mooted about different algorithms, former factory in this field and the dataset we used for our trials. All those were the foundation for this chapter. In this chapter we mooted about results that we plant after administering the algorithms and analyzed-them.

5.2 Delicacy of Models with All Features

We are implementing many machine learning algorithms to detecting heart complaints. We are implementing different types of machine learning algorithms and we go a result. Our applied machine learning algorithms are Support Vector Machine, Decision Tree, Random Forest, Gaussian Naive Bayes, Logistic regression [10]. This is the model accuracy table of our proposed model.

Classifier Algorithm	Accuracy(%)
SVM	84.77
Random Forest	96.30
Decision Tree	93.83
Naive Bayes	81.89
Logistic Regression	83.12

Table 03: Accuracy of our Algorithm.

5.3 Point Engineering

We are implementing some feature engineering to increasing accuracy on our datset.
We are doing using feature engineering for some reasons that's are

1. Choosing lower features helps us to train hastily.
2. By picking up the most important features, we can use relations between them as new features. Sometimes
this gives surprising improvement.
3. Some features are linearly related to others. This might put a strain on the model.
4. Point Selection means to handpick only the important features in- order to meliorate the delicacy of the algorithm. 6. It reduces training time and reduces over befitting.

5.4 Point Significance

In dataset there may be some attributes which do not effect the prophecy that important In some cases, numerous attributes may drop the delicacy position of a model. So, it's important to work with the correct attributes. So far we have worked with all the features of the dataset and listed the delicacy of different models.

Support Vector Machine	85.11%	sex,age,chol,cp,exang,fbs,ca,oldpeak,slope
Random Forest	87.50%	sex, cp,slope,ca,thal,thalach,fbs
Naive Bayse	84.89%	sex,ca,thalach,exang, oldpeak
Logistic Regression	82.72%	sex, age, chol, cp, ca, oldpeak, thal, restecg, slope
Decision Tree	83.83%	sex, fbs, cp, oldpeak, ca, restecg, thal

Table 04: Result of performance measure.

5.5 Results

In this part we are represent the achievement of our output based on 9 following attributes such as "sex", " fbs", "cp", "restecg", "oldpeak", "exang", "thal", "pitch" and "ca". Among all the algorithms top 3 are produce the best accuracy and they are Random Forest, Naive Bayes and Support Vector Machine (SVM).

5.6 Results of performance measure

There are 5 data mining techniques is used to perform on 1212 instance of features with 13. All the accuracy were gather for future works. In this figure shows the accuracy of following algorithms in different levels. We can see the most accuracy is 96.30% in Random Forest.

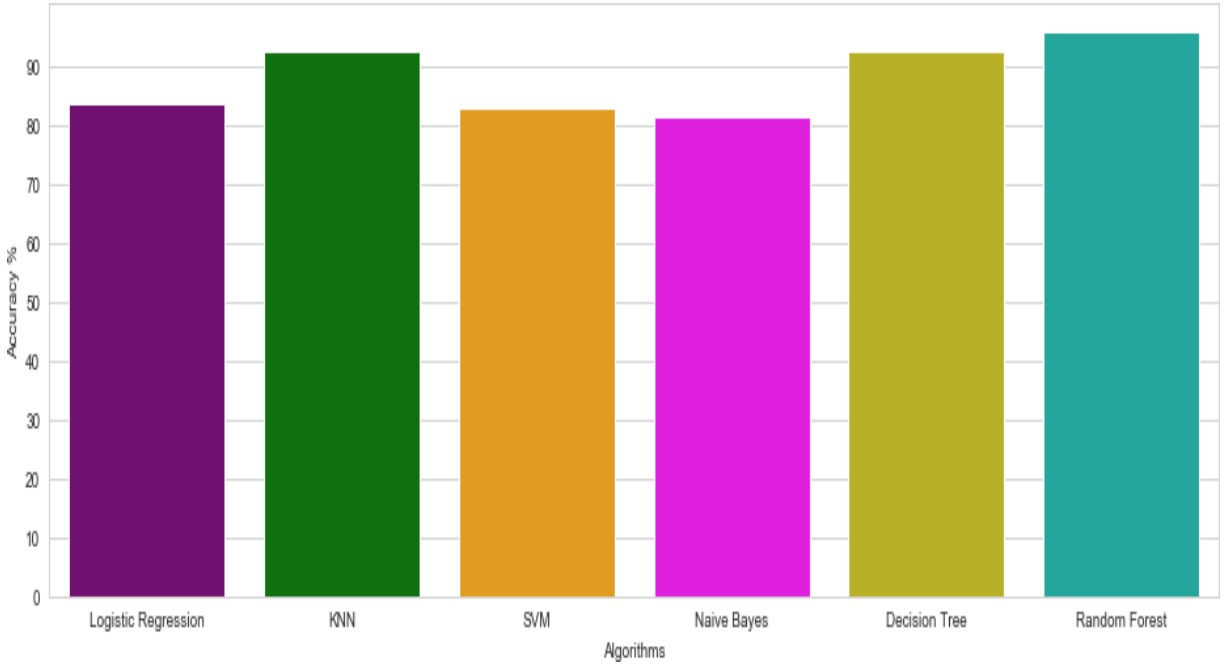


Fig: 08 Accuracy of All Algorithm

- The following table shows average accuracy in the each data mining technique:

Data Mining Technique	Average Accuracy Level
Support Vector Machine	78.15%
Random Forest	78.20%
Naive Bayes	78.20%
Logistic Regression	78.05%
Decision Tree	65.50%

Table 05: The average accuracy achieved by each data mining technique

5.7 Analysis of Features and Data Mining Fashion

In this section we are talking about the significant on the results attained from the test. By these results, the significant features and data mining techniques that have significant impacts in models are linked to prognosticate heart disease [11].

Features Occurrence	commerce	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Occurrence in the high accuracy	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in the highest F-measure	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in the highest precision	0	6	4	2	1	2	2	2	4	2	4	5	4
Total Number of Occurrence	4	20	1	4	5	12	10	8	1	14	12	19	1

Table 06: Comparison Among Attributes

- The following table shows the comparison between TMSS Medical College and Shaheed Ziaur Rahman Medical College:

Comparison Category	TMSS Medical college Dataset					Shaheed Ziaur Rahman Medical College Dataset	
No. of Attributes	13					13	
Attributes	age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal					age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal	
Class Attribute	num					num	
Different values for "num"	0,1,2,3,4					1,2	
	0	1	2	3	4	1	2
Distributions of "num"	164	55	36	35	13	150	120
Records with Missing Values	6					0	
Total number of instance	303					270	

Table 07: comparison between TMSS Medical College and Shaheed Ziaur Rahman Medical College

5.8 Selection of Data Mining Ways

Data mining techniques are essential for finalizing our model with features [12]. This model achieves the highest average accuracy using three main data mining techniques. These three techniques are Random Forest, Support Vector Machine (SVM), Naive Bayes. The Random Forest, Naive Bayes and SVM are named to develop the heart disease prophecy models.

5.9 Evaluation

In this study, we are trial to another dataset from Shaheed Ziaur Rahman Medical College and TMSS Medical College by using UCI machine respository.The datasets are quite similar. Table 6 shows the comparison between Shaheed Ziaur Rahman Medical College and TMSS Medical College dataset. Both of them are 13 features and 1 predicted feature of heart disease. All the features are same but values are not same. No missing values in the dataset. In this why, the dataset of Shaheed Ziaur Rahman Medical College was link to the original dataset for the evaluation.

5.10 Evaluation Results

we have achieve the accuracy from the data using techniques using the following attributes and the accuracy of top 3 techniques are as follow:

- ✓ Accuracy of the Random forest from the dataset with 13 features is =(133+130)/270=0.97
- ✓ Accuracy of the Support vector Machine from the dataset with 13 features is =(97+130)/270=0.84
- ✓ Accuracy of the Naive Bayes from the given dataset with13features is =(89+130)/270=0.81

Table 7: Shows the comparison of the accuracy with 13 features of attribute. The most accuracy for 13 attributes is 96.30% accuracy by the Random forest .

5.11 Evaluation of Classification Algorithm

Confusion matrix looks likes that table:

n = Total Prediction	No	Yes
No	True Negative	False Positive
Yes	False Negative	True Positive

Table 01: Confusion Matrix

We can calculate using this formula is given below:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Recall: It is defined as, how our model predicted correctly?

$$Recall = \frac{TP}{TP + FP}$$

F-measure: $\mathrm{F} - measure = \frac{2+recall+precision}{recall+precision}$

• For each algorithm, test choice cross-validation was employed. Testing, cross-validation repeats the training and testing process with arbitrary wood samples many times. Its value is 10-foldcross-confirmation.

5.12 Discussion

The mentioned results indicate that the accuracy of the three data mining techniques based on certain features is sufficient. Hard risk partitioning using data mining techniques influenced population data. These are the significant features to use data mining techniques Random Forest achieved 96.30 accuracy. Hence the random forest switch performing fashion. Based on the prediction of evaluation results, the Swish prediction model was proposed 9 significant traits (Commerce, CP, fbs, Restecg, xang, Oldpeak, Pitch, CA and thal) and discussion using random forest crossbred fashion.

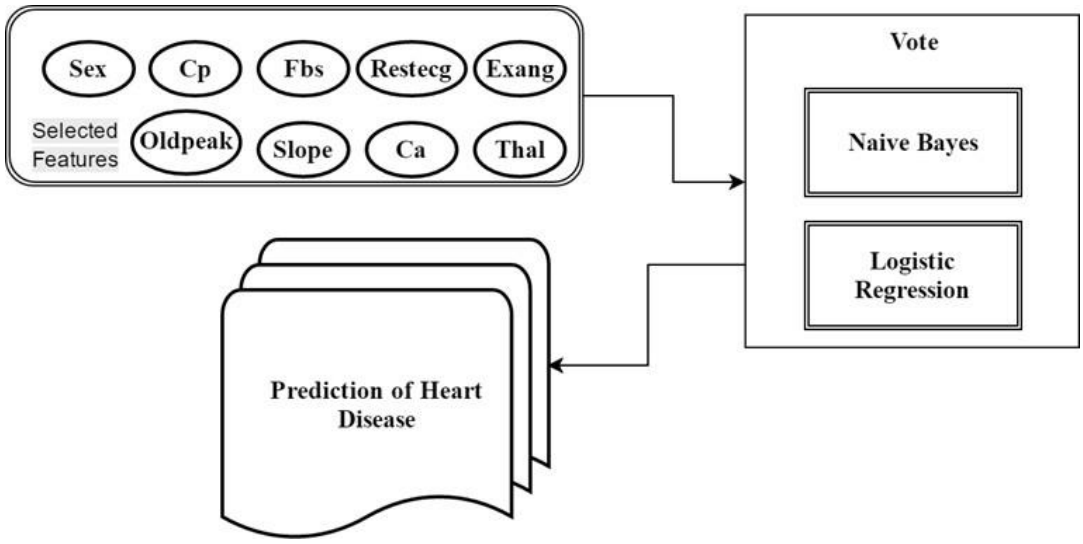


Fig : Prediction Model

5.13 Comparison of Performance

Classifier	New Accuracy(%)	Previous Accuracy(%)	Accuracy Increase(%)
SVM	84.77	78.15	6.62
Naive Bayse	81.89	78.20	3.69
Random Forest	96.30	78.20	18.10

Table 08: Accuracy Comparison

CHAPTER 6

LIMITATIONS AND FUTURE WORKS

In the previous chapter we will discuss our proposed model's result. In this chapter we will discuss the limitations and future works that are included in our thesis. In section 6.1 we will discuss limitations; in section 6.2 we will discuss future works.

6.1 Limitations

For doing this thesis firstly we are collecting dataset from Ziaur Rahman Medical & TMSS Medical College. The dataset has only 1212 cases and every case has 14 features. We are using patient Health Questioner system to collect data from patient [13]. When we are successfully collecting dataset. Then we are implementing this dataset with our proposed model. But there have some limitations. Firstly we are only working on Patient Health Questioner system we can not implementing any detection method due to short of time. And we can not virtualization of our dataset.

6.2 Future Work

We are implementing our dataset for detecting heart complaints using many machine learning algorithms and using Patient Health Questioner method. We got a final accuracy depend on our dataset. But our job is not done yet. We want to do more due to short time we can not implementing any other method so we are want to implementing more method for detection heart complaints.

CHAPTER 7
CONCLUSION

7.1 Conclusion

There are no noticeable symptoms before a heart attack occurs [13], which is why it is so important to detect the early symptoms correctly. Through this thesis work an attempt has been made to predict heart attack by reviewing patient history and initial symptoms based on various findings. Here the important information is collected from two popular & reputed medical college name as Ziaur Rahman Medical & TMSS Medical College for searching based on different attributes divided into 1212 rows and 14 different column. Also various classifier algorithm has been use to predict the causes of observed symptoms [13]. Significant risk factors for heart disease were identified using the first Hunt method of the dataset to be explored. LR, SVM, DT, RF,KNN Algorithms are prognosticate IHC using the most significant 13 threat factors [15]. The vaticination accuracy of heart disease is shown from 81.89% to 96.30% with various classifier algorithms. Where the accuracy efficiency of Random Forest is 96.30%.

References

[1] A. Hiam, "Coronary Heart Disease in Bangladesh," *World Life Expectancy,Available https//www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease.*, p. online, 2020.

[2] P. Palma, "Heart condition now a pest in Bangladesh," The Daily Star, 2019.

[3] "American Heart Association statistical report tracks global numbers of first time," (Online), 2019.

[4] www.heart.org, "Prostrating a Fitness Plateau," Online), 2019.

[5] Kinge andS. Gaikwad, ""Survey on data mining ways for complaint vaticination",," *International Research Journal of Engineering and Technology (IRJET),*, vol. vol. 05, no. no. 01,, pp. pp. 630-636, , 2018..

[6] "“ Understand Your Pitfalls to Help a Heart Attack",," *www.heart.org, 2016. (Online). Available https//www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack.*, (Penetrated 08-Sep-2018)..

[7] B. S. a. J. Shailaja, "" Machine Literacy in Healthcare A Review"," in *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA),*, in Coimbatore, India,, 2018 .

[8] J. S. a. S. A. Iyer, "" Opinion of Diabetes Using Bracket Mining Ways"," *International Journal of Data Mining & Knowledge Management Process*, Vols. vol. 5,, no. no. 1, pp. pp. 01-14, , 2015..

[9] J. S. a. S. Iyer, "" Opinion of Diabetes Using Bracket Mining Ways"," *International Journal of Data Mining & Knowledge Management Process*, Vols. ,vol. 5,, no. no. 1,, pp. pp. 01-14, 2015..

[10] I. P. a. S. K. Rajmohan, "" Vaticination and Opinion of Cardio Vascular Disease-- A Critical Check",," *2014 World Congress on Computing and Communication Technologies, 2014.* , vol. Available10.1109/ wccct.2014.74, (Penetrated 20 April 2018)..

[11] RedMonk, "https://redmonk.com/sogrady/2022/03/28/language-rankings-1-22/," 2022.

[12] S. V. S. P. A. a. G. Pouriyeh, ""A comprehensive disquisition and comparison of Machine Learning Ways in the sphere of heart complaint"," *in 2017 IEEE Symposium on Computers and Dispatches (ISCC), Heraklion, Greece,*, 2019.

[13] M. I. G. S. C. M. a. M. M. Raihan, ""A Comprehensive Analysis on Threat Vaticination of Acute Coronary Syndrome Using Machine Learning Approaches",," in *in 2018 21st International Conference of Computer and Information Technol*, 2018.

[14] B. Rathnayakc andG. Ganegoda, ""Heart conditions vaticination with Data Mining and Neural Network Ways",," *in 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, India,*, pp. pp. 1-6., 2018.

[15] M. Marcus, ""Some jobs feel unsafe when it comes to heart health",," CBS NEWS,, (Penetrated 9-Sep-2018)..

[16] C. Moran, ""Farmers 7 times more likely to die from heart complaint",," *Available https//www.agriland.ie/farming-news/farmers-7-times-more-likely-to-die-from-cardiovascular-complaint/.*, (Penetrated 11-Sep-2018)..

[17] "“ Understand Your Pitfalls to Help a Heart Attack",," *www.heart.org, 2016. (Online). Available https//www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack. ,* (Penetrated 08-Sep-2018)..

[18] D. Kinge andS. Gaikwad, ""Survey on data mining ways for complaint vaticination",," *International Research Journal of Engineering and Technology (IRJET),*, vol. vol. 05, no. o. 01, pp. pp. 630-636,, 2018..