A CAL PROJECT REPORT

ON

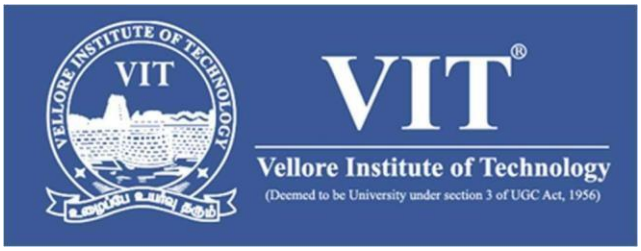
A COMPARATIVE STUDY OF DATA MINING ALGORITHMS ON HEART DISEASES PREDICTION

FOR THE COURSE

DATA MINING TECHNIQUES–SWE 2009(A1+TA1)

BY

VALLIAMMAI.M 16MIS0488



FALL SEMESTER 2019-20

ACKNOWLEDGEMENT

The project report entitled **A COMPARATIVE STUDY OF DATA MINING ALGORITHMS ON HEART DISEASES PREDICTION** submitted to school of information technology and engineering, VIT university by Valliammai.M registration number 16MIS0488 respectively under guidance of Prof. Prabhavathy P.

.

.

.

.

.

.

.

.

.

.

.

prof Prabhavathy P

Associate Professor

School of Information Technology and Engineering

TABLE OF CONTENTS

ABSTRACT

In recent decades, heart disease has been identified as the leading cause of death across the world. According to World Health Organization (WHO), the early and timely diagnosis of heart disease plays a remarkable role in preventing its progress and reducing related treatment costs. Although cardiovascular diseases have been identified as the leading cause of death in the world in the past decade, they have been introduced as the most preventable and controllable diseases. Considering the ever-increasing growth of heart disease-induced fatalities, researchers have adopted different data mining techniques to diagnose it. Data mining scholars have long studied the application of tools and equipment in improving the process of data analysis in large and complex datasets. Adopting data mining techniques in the medicine field is of high importance in diagnosing, predicting and deeply understanding of healthcare data .These applications include treatment centers analysis aimed at improving treatment policies and prevention of any mistake in hospitals, early diagnosis of diseases, prevention of diseases and hospital death reduction.

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

According to the World Health Organization heart disease is the first leading cause of death in high and low income countries and occur almost equally in men and women. By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases. Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths. For CVDs specifically, in 2005, the age standardized mortality rate for developing nations like India, China, and Brazil was between 300-450 per 100,000, whereas it was around 100-200 per 100,000 for developed countries like USA and Japan . According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data . From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization and healthcare. Medical Data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately

and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. In this project we have analyzed the several data mining techniques proposed in recent years for the diagnosis of heart disease. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, Random forests, Naïve Bayes, Decision tree, Linear Regression, Genetic algorithm which are showing accuracy at different levels. We have concentrated on four algorithms mainly: KNN, Linear Regression, Random Forests & Support Vector Machine(SVM).

1.2. OBJECTIVE OF THE WORK

The main objective of this project is to compare a data mining modeling techniques and find the best accuracy among the data modelling techniques ,namely, KNN, SUPPORT VECTOR MACHINE ,LOGISTIC REGRESSION and RANDOM FOREST this project helps in discovering and extracting hidden knowledge associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, ialso helps to reduce treatment costs.

1.3. SCOPE OF THE WORK

Data mining techniques are used for variety of applications. In health care industry, data mining plays a significant role for predicting diseases. For identifying a disease, number of tests should be essential from the patient. But using data mining technique the number of test should be condensed. This reduced test plays an important role in time and accuracy. This project analyzes a comparative analysis of how data mining techniques are carried out to prediction the best accuracy of algorithms used for heart disease prediction .

CHAPTER 2

LITERATURE REVIEW

Machine Learning has been documented as a classification mechanism for medical and otherapplications. Below are some of the relevant work which has already made its mark in this field of study:

In [1] In this paper data mining techniques such as Naïve Bayes, Hunts Salgorithm & bagging are used. Experiments are conducted using Weka tool and the results are compared with bagging

and without bagging using 10-fold cross validation. Finally Bagging algorithm is stated to be the successful data mining technique used in heart diseases detection.

In [2] a weighted fuzzy rule-based clinical decision support system (CDSS) is presented for the diagnosis of heart disease, automatically obtaining knowledge from the patient's clinical data. The proposed clinical decision support system for the risk prediction of heart patients consists of two phases: (1) automated approach for the generation of weighted fuzzy rules and (2) developing a fuzzy rule-based decision support system. In the first phase, we have used the mining technique, attribute selection and attribute weightage method to obtain the weighted fuzzy rules.Then, the fuzzy system is constructed in accordance with the weighted fuzzy rules and chosen attributes.Finally, the experimentation is carried out on the proposed system using the datasets obtained from the UCI repository and the performance of the system is compared with the neural network- based system utilizing accuracy, sensitivity and specificity.

In [3] In this paper three classification tree techniques are compared Tree based decision stump technique, LMT, Random forest technique.It is observed that decision stump classification tree technique turned out to be the best classifier.

In [4] In this paper heart diseases detection using data mining & artificial neural networks is done. The pre-processed heart disease data warehouse was clustered with k- means clustering obtain data most applicable to heart attack.The frequent items are mined successfully with the aid of MAFIA Algorithm.

In [5] This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. Different supervised machine learning algorithms i.e. Naïve Bayes, Neural Network, along with weighted association Apriori algorithm, Decision algorithm have been used for analyzing the dataset and the conclusion is Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

In [6] Naïve Bayes, neural networks, decision tree, logistic regression,association rule is the data mining algorithms applied to the data set. After implementing the algorithms into the data set a pattern for each data set is created. And it is proved association rule provide accurate results as compared to the remaining techniques.

In [7] Classification is a process used to find a model that describes and differentiate data classes or concepts for the purpose of using the model to predict the class of objects whose class label is unknown. Some of the tools used for classification are: decision tree,ID3,C4.5,C5.0,J48.The conclusion from the above paper is that decision tree is best suited for detecting heart disease. It is well known for its simplicity and accuracy.

In [8] The various methodology section discusses the voting technique and KNN. The techniques are: Voting, K-nearest neighbour, 10 fold cross validation,heart disease data.This paper proves that K-nearest algorithm has the highest accuracy with 97.4%.

In [9] The objective of this paper is to provide a study of different data mining techniques that can be employed in automated disease prediction sytem. The paper concludes that Naïve Bayes has the highest accuracy of all.

In [10] In this different data mining techniques have been used such as Neural Networks,Decision trees, Naïve Bayes. Associative classification is a recent & rewarding technique which intergrates association rule mining and classification a model for prediction and achieves maximum accuracy. The overall objective of this paper is to study the various data mining techniques and to compare the best method for prediction.

In [11] Different Apriori strong rules were contrasted with aim to foresee coronary illness. A one of a kind model comprising of one filter and evaluation techniques are developed. Three strong rules and distinctive assessment techniques, are applied to find the superior software this paper concludes that using Apriori algorithm with strongs rules good results are acquired as compared with neural networks.

In [12] the authors have a medical diagnosis system for predicting the risk of cardiovascular disease. multilayered back propagation neural networks are particularly suited to complex classification problems. the weights of the neural network are determined using genetic algorithm

because it finds acceptably good set of weights in less number of iterations. genetic- neural network is used for training the system. the classification accuracy obtained using this approach is 90.17%. and also system this sytem will thereby also provide suggestions with the diet chart for the patients on the basis of the various parameters compared.

In [13] this paper emphasizes on the study of prediction ofheart diseases using data mining and machine learning algorithms and their tools 8 algorithms different algorithms are used which includes decision tree, j48 algorithm, logistic model tree algorithm, random forest algorithm, naïve bayes, knn, support vector machine, nearest neighbour to predict the heart diseases. and accuracy on using this algorithms are predicted.

In [14] This paper the emphasizes the various data mining techniques available to predict heart disease and to compare them to find the best method of predictions. The authors have focused on classification method and prediction method of data mining using Naive Bayes and Improved K-means algorithm. The accuracy of the algorithm used in each technique can be enhanced by hybridizing or combining algorithm to single algorithm.(Naive Bayes and Improved K-means algorithm).

In [15] studyof heart disease prediction system (EHDPS) using neural network for predicting the risk level of heart disease. The EHDPS predicts the likelihood of patients getting heart disease. The authors have implemented the multilayer perceptron neural network with backpropagation as the training algorithm. From ANN, an MLPNN together with BP algorithm is used to develop the system.and the final experimental results show that the system predicts heart disease with ∼100% accuracy by using neural networks.

In [16] The author designed a heart disease prediction system using various data mining techniques and to perform the analysis of the results obtained for all implemented techniques. The performance analysis is done by using Naïve Bayes and Genetic Algorithm and proved that out of all algorithms these two provides correct results.

In [17] The authors designed a perceptive model for heart illnesses acknowledgment using data mining strategies that are fit for enhancing the constancy of heart infections conclusion. They divide the data into Training and Testing Data Sets and employ Naïve Bayes technique to obtain relatively higher prediction accuracy .they focused on accuring highly accurate prediction of Heart Disease.

In [18] The paper describes about a prototype using data mining techniques mainly Naïve Bayes and WAC (Weighted Associated Classifier). Authors presented a efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart disease. The system indicates whether patient had a risk of heart disease or not.

In [19] The authors have carried out a analysis to develop a prototype Health Care Prediction System using, Naive Bayes. The System will discover and extract hidden data related to diseases (heart attack, cancer and diabetes) from a historical heart disease database. It will answer complicated queries for diagnosing sickness and so assist care practitioners to form intelligent clinical selections which ancient call support systems cannot. By providing effective treatments, it conjointly helps to reduce treatment prices.

In [20] In this paper the authors have presented an Efficient Heart Disease Prediction System using data mining. The main contribution of this study is to help a non-specialized doctors to make correct decision about the heart disease risk level. The rules generated by the proposed system are prioritized as Original Rules, Pruned Rules, Rules without duplicates, Classified Rules, Sorted Rules and Polish This system can help medical practitioner in efficient decision making based on the given parameter. They had train and tested the system using 10 fold method and found the accuracy of 86.3 % in testing phase and 87.3 % in training phase.

2.2. PROBLEM DEFINITION AND APPROACH

Problem Definition

The diagnostic of heart disease remains more or less the most difficult and tedious task in the medical field and it various factors and symptoms of prediction which is involved in several layered issue that could engender the negative presumptions and unpredictable effects. Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?”, “Identify the female patients who are single, above 30 years old, and who have been treated for cancer.” However, they cannot answer complex queries like “Identify the important preoperative predictors that increase the length of hospital stay”, “Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?”, and “Given patient records, predict the probability of patients getting a heart disease.” Clinical decisions are often made based on doctors’ 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. Wu, et al proposed 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 [23]. This suggestion is promising as data modelling 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.

CHAPTER 3

EXPERIMENTAL DETAILS

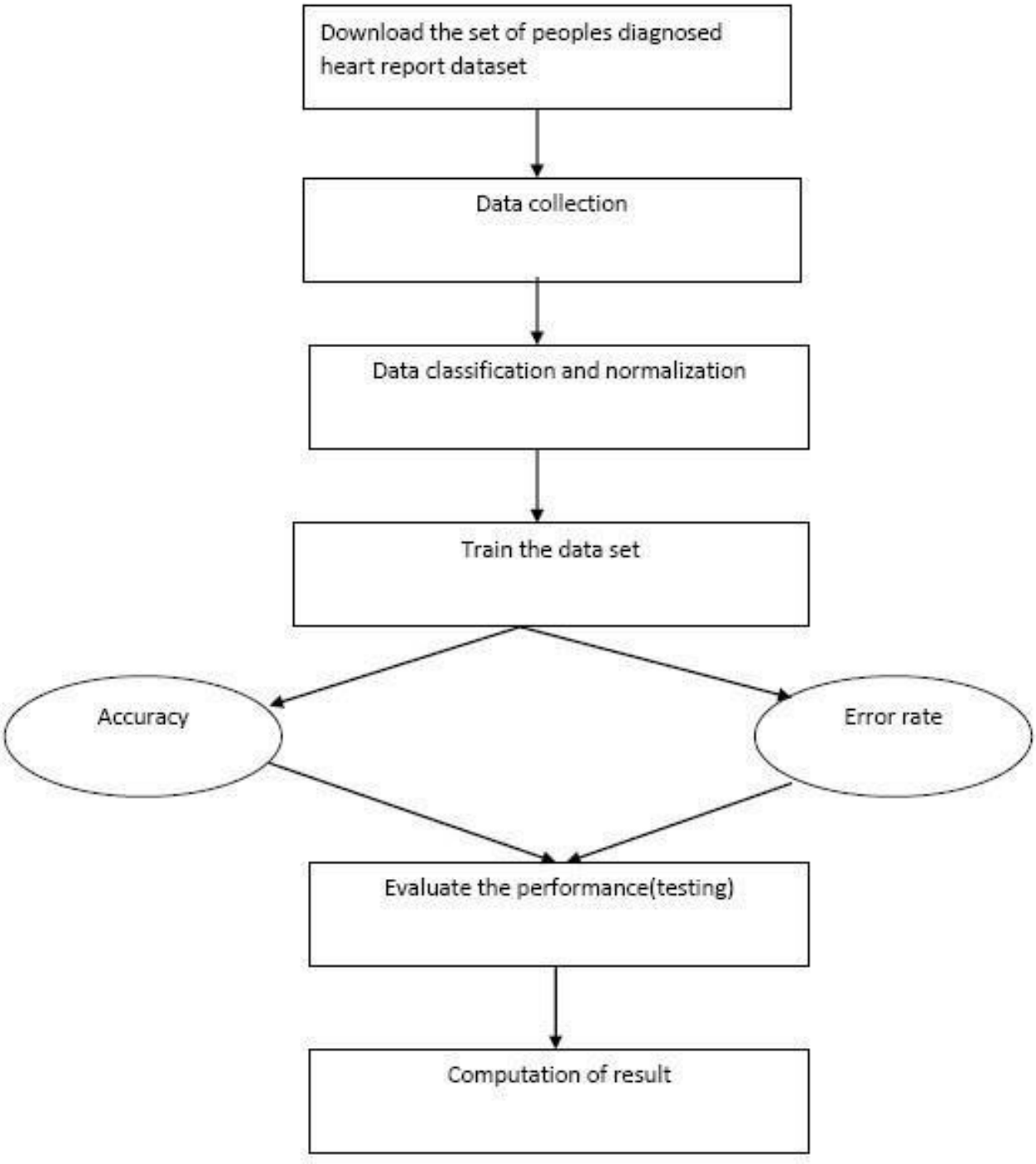
3.1. MACHINE LEARNING METHODS

Healthcare environment generally includes rich data about patients and their treatment, which are stored in health management systems. Such data is valuable especially if we cultivate the existing information into useful practices. Data mining techniques can help in extracting valuable knowledge from hidden relationships and trends among data. In fact, data mining techniques have been extensively used in healthcare research as a stage of knowledge discovery process , which offers promising ways to uncover hidden patterns within large amounts of health data . It has been applied to a diversity of healthcare domains to improve and accelerate decision making . The data mining techniques are divided into four types: Classification, Clustering, Regression and Association rule mining.

Classification methods are the most widely used algorithms in Healthcare sector as it helps predicting the status of patient by classifying patient records and find the class that matches the new patient record . Basically, classification is known as supervised learning techniques which requires the data to be initially classified into initial classes or labels . Then these data are entered into a classification algorithm in order to be learned as shown in. Particularly, the relationship between attributes needs to be discovered by the algorithm to predict the outcome. In this phase the classification algorithms builds the classifier from the training set made up of dataset tuples and their associated class labels. Every tuple that constitutes the training set is referred to as a category or class. When a new case is arrived the developed classification algorithm is used to classify it into one of the predefined classes as shown in Figure . The term which specifies how “good the algorithm is” is called prediction accuracy. For instance, the training set in the medical database would have much relevant patient information recorded already, where the prediction outcome is whether or not the patient had a heart disease.

here are many classification algorithm that can be used for healthcare datasets. Among them, we have chosen K-nearest algorithm ,support vector machine ,Random forest, and Support vector Machine(SVM).

3.2. DESIGN FRAMEWORK



3.3. DATA SET, DATA SOURCE, CHARACTERIZATION , PREPROCESSING

DATA SET:

Four datasets are joined and preprocessed .the dataset includes cleveland, hungarian, long-beach-va, and switzerland databases. Each database has the same number of features, which is 14, but different numbers of records.so they are join .the dataset 299 instances . Table 1 shows the 14 attributes/features as they exist in the dataset alongside the description of each attribute

DATASET DESCRIPTION:
TABLE 1:

Field	**Description**	**Range and Values**
Age	Age of the patient	0-100 in years
Sex	Gender of the patient	0-1 (1:Male 0:Female)
Chest Pain	Type of chest pain	1-4 (1: Typical Angina, 2: Atypical Angina, 3: Non-anginal, 4: Asymptotic)
Resting Blood Pressure	Blood pressure during rest	mm Hg
Cholesterol	Serum Cholesterol	mg / dl
Fasting Blood Sugar	Blood sugar content before food intake if >120 mg/dl	0-1 (0: False, 1: True)
ECG	Resting Electrocardiographic results	0-1 (0: Normal, 1: Having ST-T wave)
Max Heart Rate	Maximum heart beat rate.	Beats/min
Exercise Induced Angina	Has pain been induced by exercise	0-1 (0: No, 1: Yes)
Old Peak	ST depression induced by exercise relative to rest	0-4
Slope of Peak Exercise	Slope of the peak exercise ST segment	1-3 (1: Up sloping, 2: Flat, 3: Down sloping)
Ca	Number of vessels colored by fluoroscopy	0-3
Thal		3- normal 6- Fixed Defect 7- Reversible Defect
Num	Diagnostics of Heart Disease	0-1 (0: <50% Narrowing 1: >50% Narrowing)

DATA PREPROCESSING:

The performance and accuracy of the predictive model is not only affected by the algorithms used, but also by the quality of the dataset and the preprocessing techniques. Preprocessing refers to the steps applied to the dataset before applying the machine learning algorithms to the dataset. The preprocessing stage is very important because it prepares the dataset and puts it in a form that the algorithm understands. Datasets can have errors, missing data, redundancies, noise, and many other problems which cause the data to be unsuitable to be used by the machine learning algorithm directly. Another factor is the size of the dataset. Some datasets have many attributes that make it harder for the algorithm to analyze it, discover patterns, or make accurate predictions. Such problems can be solved by analyzing the dataset and using the suitable data preprocessing techniques.

Data preprocessing steps includes: data cleaning, data transformation, missing values imputation, data normalization, feature selection, and other steps depending on the nature of the dataset[24].

STATISTICAL SUMMARY OF ALL THE ATTRIBUTES:

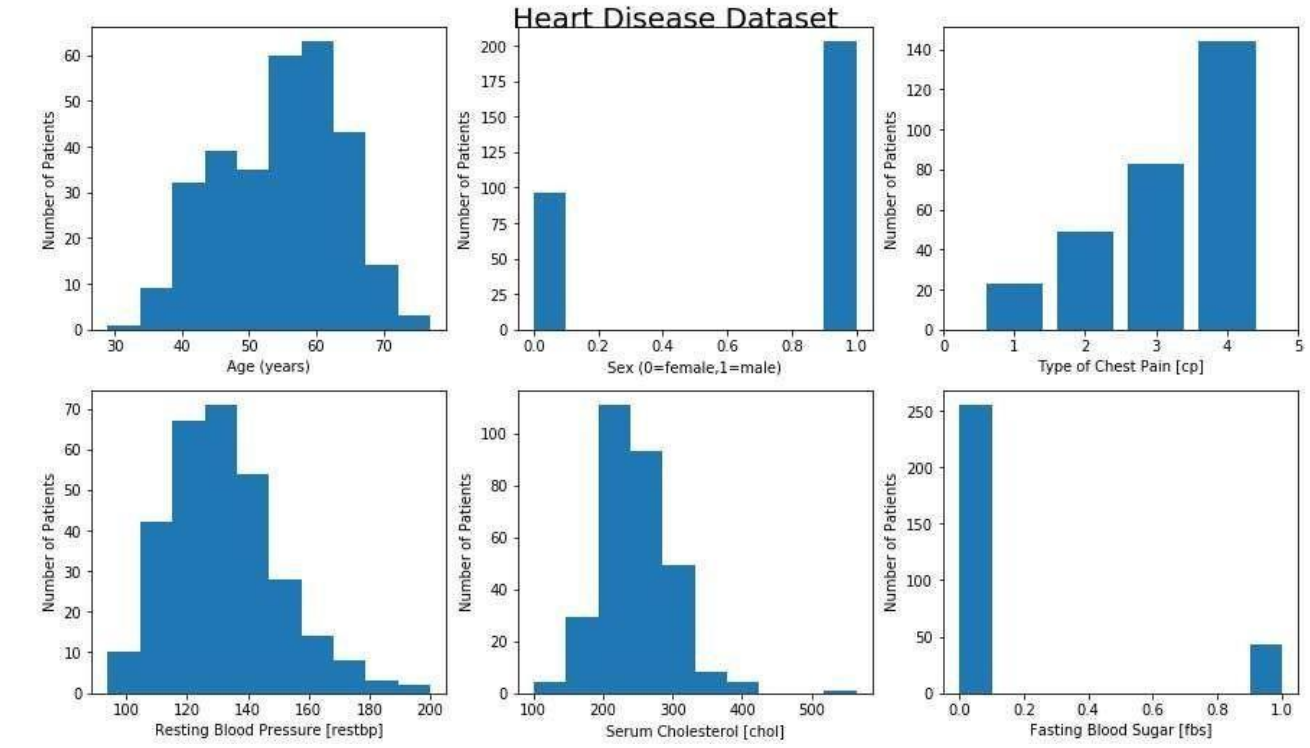
	age	**sex**	**cp**	**restbp**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**num**
Count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000
Mean	54.521739	0.67893	3.163880	131.715719	246.785953	0.143813	0.989967	149.327759	0.331104	1.051858	1.605351	0.672241	4.745819	0.946488
Std	9.030264	0.46767	0.964069	17.747751	52.532582	0.351488	0.994903	23.121062	0.471399	1.161276 9	0.616962	0.937438	1.940977	1.230409
Min	29.000000	0.000000	1.000000	94.000000	100.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	132.500000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	242.000000	0.000000	1.000000	152.000000	0.000000	0.800000	2.000000	0.000000	3.000000	0.000000

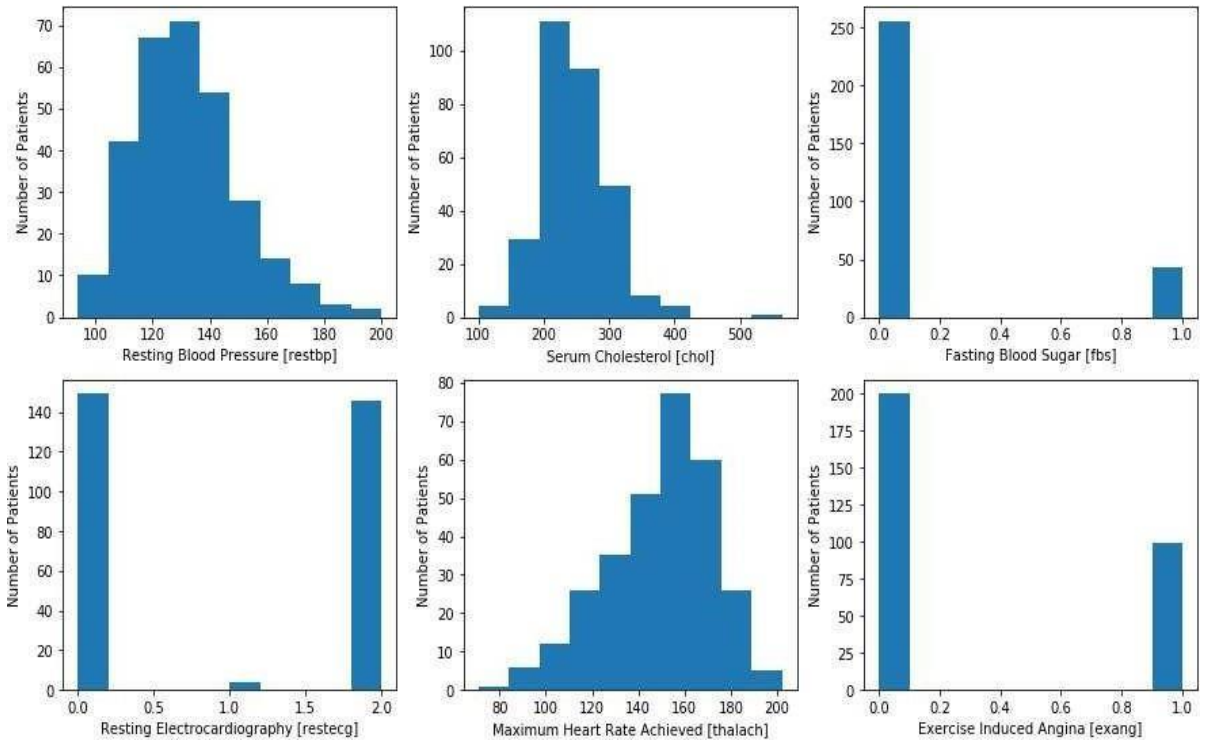
	age	sex	cp	restbp	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
75%	61.000000	1.000000	4.000000	140.000000	275.500000	0.000000	2.000000	165.500000	1.000000	1.600000	2.000000	1.000000	7.000000	2.000000
Max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000	4.000000

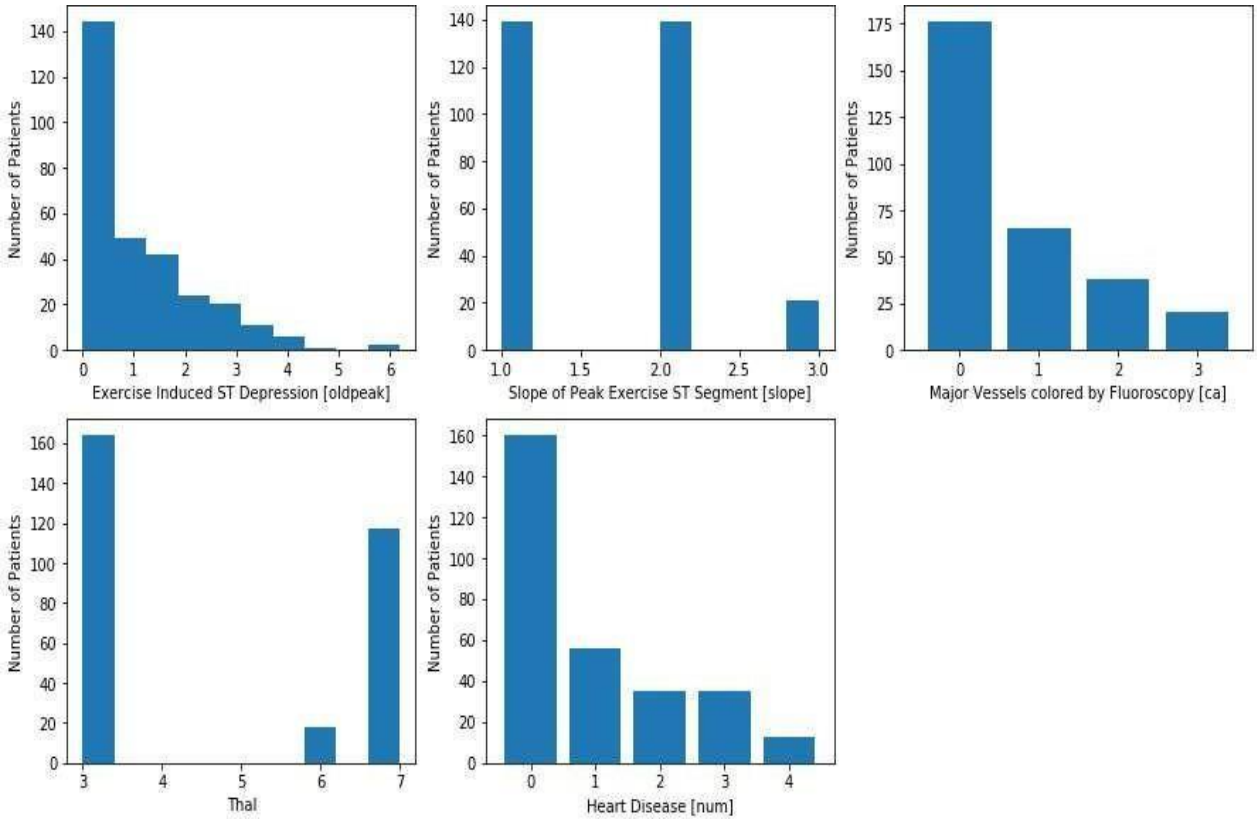
GENERATING HISTOGRAMS

Histograms are generated for viewing distribution of values in dataset. With simple histogram of the data, the distribution of different attributes can be easily observed .And it is extremely easy for us to see which attributes are categorical values and which arenot.

For example : distribution of ages and fbs (fasting blood sugar).the age distribution is closely resembling of Gaussian distribution while fbs is a categorical value.


Heart Disease Dataset
Number of Patients
Age (years)
Sex (0=female,1=male)
Type of Chest Pain [cp]
Resting Blood Pressure [restbp]
Serum Cholesterol [chol]
Fasting Blood Sugar [fbs]


Number of Patients
Resting Blood Pressure [restbp]
Serum Cholesterol [chol]
Fasting Blood Sugar [fbs]
Resting Electrocardiography [restecg]
Maximum Heart Rate Achieved [thalach]
Exercise Induced Angina [exang]


Number of Patients
Exercise Induced ST Depression [oldpeak]
Number of Patients
Slope of Peak Exercise ST Segment [slope]
Number of Patients
Major Vessels colored by Fluoroscopy [ca]
Number of Patients
Thal
Number of Patients
Heart Disease [num]

Converting categorial values into discrete values:

Number of patients in dataframe: 299, with disease: 139, without disease: 160

	age	sex	restbp	chol	fbs	thalach	exang	oldpeak	ca	hd	cp_1	\
0	63.0	1.0	145.0	233.0	1.0	150.0	0.0	2.3	0.0	0.0	1	
1	67.0	1.0	160.0	286.0	0.0	108.0	1.0	1.5	3.0	1.0	0	
2	67.0	1.0	120.0	229.0	0.0	129.0	1.0	2.6	2.0	1.0	0	
3	37.0	1.0	130.0	250.0	0.0	187.0	0.0	3.5	0.0	0.0	0	
4	41.0	0.0	130.0	204.0	0.0	172.0	0.0	1.4	0.0	0.0	0	

	cp_2	cp_3	recg_1	recg_2	slope_1	slope_3	thal_6	thal_7
0	0	0	0	1	0	1	1	0
1	0	0	0	1	0	0	0	0
2	0	0	0	1	0	0	0	1
3	0	1	0	0	0	1	0	0
4	1	0	0	1	1	0	0	0

	age	sex	restbp	chol	fbs	thalach	\
count	299.000000	299.00000	299.000000	299.000000	299.000000	299.000000	
mean	54.521739	0.67893	131.715719	246.785953	0.143813	149.327759	
std	9.030264	0.46767	17.747751	52.532582	0.351488	23.121062	
min	29.000000	0.00000	94.000000	100.000000	0.000000	71.000000	
25%	48.000000	0.00000	120.000000	211.000000	0.000000	132.500000	
50%	56.000000	1.00000	130.000000	242.000000	0.000000	152.000000	
75%	61.000000	1.00000	140.000000	275.500000	0.000000	165.500000	
max	77.000000	1.00000	200.000000	564.000000	1.000000	202.000000	

	exang	oldpeak	ca	hd	cp_1	cp_2	\
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	
mean	0.331104	1.058528	0.672241	0.464883	0.076923	0.163880	
std	0.471399	1.162769	0.937438	0.499601	0.266916	0.370787	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.800000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	1.600000	1.000000	1.000000	0.000000	0.000000	
max	1.000000	6.200000	3.000000	1.000000	1.000000	1.000000	

	cp_3	recg_1	recg_2	slope_1	slope_3	thal_6	\
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	
mean	0.277592	0.013378	0.488294	0.464883	0.070234	0.060201	
std	0.448562	0.115079	0.500701	0.499601	0.255970	0.238257	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

	thal_7
count	299.000000
mean	0.391304
std	0.488860
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

3.4. PROCESSING TECHNIQUES

Classification machine learning techniques

Classification, which is a type of supervised ML techniques perform predictions for future cases based on a previous dataset.

(i) **K-Nearest Neighbor**

KNN characterization is a standout amongst the most basic and straightforward arrangement strategies and ought to be one of the main decisions for an order study when there is almost no earlier learning about the dissemination of the information. KNN order was produced from the need to perform discriminated examination when dependable parametric assessments of likelihood densities are obscure or hard to decide. K-Nearest Neighbor is also known as lazy learning classifier.

An object is classified in a class to which its k-nearest neighbors belong. In the kNN algorithm, the classification of a new test feature vector is determined by the classes of its k-nearest neighbors. Here, the kNN algorithm was implemented using Euclidean distance metrics to locate the nearest neighbor. The Euclidean distance metrics d(x,y) between two points x and y is calculated using the equation below. Where N is the number of features such that x = {x1, x2, x3,...,xn} and y = {y1, y2, y3,...,yn}

$$d(x,y) = \Sigma_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. SVM has become more popular tool for machine learning tasks. It is supervised learning model that is applied mainly for classification, but it can also work for regression problems. The basic SVM works as binary classifier where the training data is divided into two classes. For multiclass problem, the basic SVM algorithm is executed repeatedly on the training data .

The SVM algorithm mapped feature vector into a higher dimensional vector space, where a maximum margin hyper-plane is established in this space. The distance from the hyper-plane to the nearest data point on each side is maximized. Maximizing the margin and thereby producing the largest possible distance between the separating hyper-plane and the instances on each side of it has been proven to reduce an upper bound on the expected generalization error. In this algorithm,

each data item as a point in n-dimensional space is plotted with the value of each feature being the value of a particular coordinate. Then, classification is done by finding the hyper plane that differentiate the two classes very well. SVM is very effective in high dimensional spaces. Also, it is effective in cases where number of dimensions is greater than the number of samples. Moreover, it uses a subset of training points in the decision function as a result it is also memory efficient. It has also some versatility like different Kernel functions can be specified for the decision function. There are also some common kernels provided, but it is also possible to specify custom kernels

3 . LINEAR REGRESSION:

Linear regression is a very simple approach for supervised learning. Though it may seem somewhat dull compared to some of the more modern algorithms, linear regression is still a useful and widely used statistical learning method. Linear regression is used to predict a quantitative response Y from the predictor variable X.Linear Regression is made with an assumption that there's a linear relationship between X and Y.

Mathematically, we can write a linear relationship as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Where:

- *y* is the response
- *β* values are called the **model coefficients**. These values are "learned" during the model fitting/training step.
- *β0* is the intercept
- *β1* is the coefficient for *X1* (the first feature)
- *βn* is the coefficient for *Xn* (the nth feature)

Linear Regression is a very powerful statistical technique and can be used to generate insights on consumer behaviour, understanding business and factors influencing profitability. Linear regressions can be used in business to evaluate trends and make estimates or forecasts. For example, if a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.

Linear regression can also be used to analyze the marketing effectiveness, pricing and promotions on sales of a product. For instance, if company XYZ, wants to know if the funds that they have invested in marketing a particular brand has given them substantial return on investment, they can use linear regression. The beauty of linear regression is that it enables us to capture the isolated impacts of each of the marketing campaigns along with controlling the factors that could influence the sales. In real life scenarios there are multiple advertising campaigns that run during the same time period. Supposing two campaigns are run on TV and Radio in parallel, a linear regression can capture the isolated as well as the combined impact of running this ads together.

4 RANDOM FOREST

Random Forest is a supervised learning algorithm. Like you can already see from it's name, it creates a forest and makes it somehow random.Random forests build multiple decision trees and merge them to get a more stable and accurate prediction.One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Important Hyper parameters:

The Hyper parameters in random forest are either used to increase the predictive power of the model or to make the model faster. I will here talk about the hyper parameters of sk learns built-in random forest function.

1. Increasing the Predictive Power

Firstly, there is the **„n_estimators"** hyper parameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking averages of predictions. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation. The last important hyper-parameter we will talk about in terms of speed, is **„min_sample_leaf "**. This determines, like its name already says, the minimum number of leafs that are required to split an internal node.

2. Increasing the Models Speed

The hyper parameter tells the engine how many processors it is allowed to use. If it has a value of 1, it can only use one processor. A value of "-1" means that there is no limit**„ random_state"** makes

the model's output replicable. The model will always produce the same results when it has a definite value of random_state and if it has been given the same hyperparameters and the same training data.

CHAPTER 4

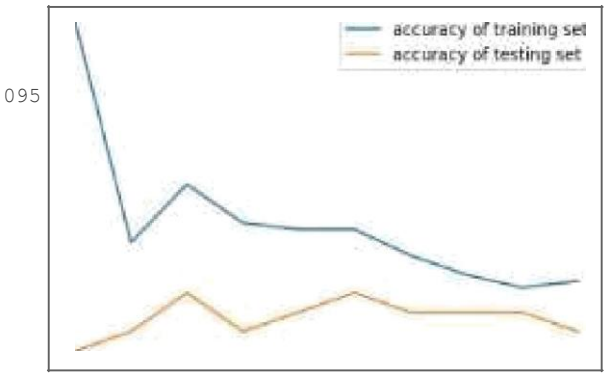
RESULTS AND DISCUSSION:

In [17]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline

x_train,x_test,y_train,y_test = train_test_split(Xall,yall,random_state=42)

knn= KNeighborsClassifier(n_neighbors=4)
knn.fit(x_train,y_train)
```

Out[17]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=4, p=2,
                     weights='uniform')
```

```
n[1B]:x=knn.predi'ct(x_traio}
      y=knn.predict{x_test}


      for i in range(124):
          if x{i)  1:


              k+=l

      for i in ranget75):
          if y[i] .l:


              k+=l
      print(1en(x))
      print(1en(y)}
      print('number patients affected(malignamt)',j)
      print('number patients not affected(bqnign)',k)

      print('accuracy o* Ann n   5, cn the t^aininq set: (:.3f)'.format(knn:score(a train,y_traio))}
      print('accuracy of #nn n   3, on the tePt set:(:.3f}'.fommt((knn.score1x_test,y_test})))


      75
      number patients affected(malignant) 116
      n'mber patients not affectedlbenignl 183
      accuracy of knn n - 9, on the trainingq sqt: 0.862
      accuracy of knn n - 9, on the test set: 0.7o7
```

```
In  -15..
          .':. .        .. ..  . .
         . raining_ac::pracy=
         test_accurac   [


         'foP'n''ñei$   rs''''         'oxs'''set:mgs:
             'clF*KNeij   rsClié ifier(h_neighbors=n néi'    Drs)
             clf'..fiCfx_   aih y_ti&iñ1''"
             :traini   aqcugacy.append clf=scqre(x_€rain,y_train)
             test_accuracy.append(clf.score(x_test,y_test))

         plt.plñt(néij   rs_settiñ  , traiñi@g_ac   %cy, lébel="atcuracy of'tiaini   se 'j:
          plD.ploD(nerghbors_settiñgs, est_accuxacv,iabel- zczuracv of  estiñq'set )
         pit..label{'gc=  âcy'l
         plt.xlabel('number of neighbours')


iut?i5}: -oâatplDtlib=lege:xl.Imgend.at 0a649b0ñ"470>
```



niimLri nf i ei §hhoui s

```
In {55). frosn sklee*o.neighbors ioport BEeighbozsClassifier
         fiom sklearn.ɑ⁻⁻el_selection import train_test_split'
         isport mqtplotlib.pyplot aua plt
         %matqilotlib idline

         x_train,a_test y_trein,y_test = train_test_split(Xsll,yall,random_state=4x)

         kon= KNeighborsCla9sifier(n_neighbors=6)
         kmn.Jit?x_tiain,y_€rain)

         x=tnn.predict(x_t*ein)
         y=tnn.predict(x_lest)
         S=S

         #for i in range(len(mlp.predict(x_train)):
         for i io rshge(224T.
             if x{i] l':




         for i io range(751:
             iE y{i] l:



         prinbClenl*l)
         print(len(x_test))
         print('nueber patients affered(malignant)',j)
         print.('number palienbs no* gffectedCbenign,',k)

         prinbC'accuracy of knn q - 6, on lhe lzaining see. (..2f}'.fomRac(Dnn.scorq(x_braip,y_t*ainl))
         prin1C'accuracy of knn ñ - G, on lhe 1es1'seb: {..3f}'.format(Ckrm.sco*elx_tes*,y_test))]l



         nuRber oaliencs affected(malignanbt U7
         number patients not affectedCbenign) l7z
         aicuracy of knn n - G, on the trainiq'se€: O.8B7
         accuracy of knn n - 6, on lhe tesl seb. 0.815
```

```
In {<0}. from skleaxn.lireaz —osel 9—rt LogisticRegression
         lrclf=[ogisiicReqressionCxancMsn_stat  J sol:n=r=”lbfgs”)
        ’.irclf.Ii€(x traio,y tzain)
         x=lrclI.predict(x train)
         v=lrclW.predict(x_test)

         k=0
         #for i in range(len(mlp.predict(x_train)):

             iE.x1i]  1.

             elee.


         for i in range(75):
             if y[i]==1:


         priat(lerCx))
         primu‹ler.(x_test))
         print('nuAber’ pa€ients affeJted(naiignzrt)’,j}
         printi'numMer pabieants nccaffezsed›ter.ign}',k)

         priat('accuracv on €he training set. {..3f)'.formet(lrcl€.score'x_txain,y_tréin)))
         print('accuracy on the test set. {..ofj’.format{(lrclf.score{a_test,y_test))))


         nuRber patieñte affected(malignant) 153
         nrnrbex patience not affe:tedCbeniqr) Z76
         accuracy on che training set- 0.857
         aoczzz act ozt size cesc set . 0 . E 67
```

```
Io i58D. from sklearn.svm impowt PVC
         svc = E2C(kernel='zbf', randceo_state='@, gaJoea='auro'}
         évc.fit{x_txain„y_train)
         x=svc.predict(x_train)
         y=9vc.pzedict{R_tePt)


         #for i in range(len(mlp.predict(x_train))):
         for i in range(Z2?l .
             if.x[i]  1.



         for i io rangef7S).
             iE y(i]  1.


         print(lenix)}
         piintilen(x_test))
         print('nuAbez patients affe:ted(malignzrtl',j)
         print('number patiente nct affe:ted(ker.ign} '.t)

         print('accuracy on the training set: {:.3f)'.format(rvc.score(x traio,y train)I)
         piiat('accuracy on the test set. {..'of)'..format{(s:zc.score(r_test,y_test}))j



         nusber oatients affected(maiigoant] 154
         nu&her patience not affected(benign) 175
         accuracy on the training set. 1.8E*
         accuracy on the test set. 0.E13
```

```
In [59]: from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier(n_estimators=250, criterion='entropy',random_state=0 )
rfc.fit(x_train,y_train)
x=rfc.predict(x_train)
y=rfc.predict(x_test)
j=0
k=0
#for i in range(len(mlp.predict(x_train))):
for i in range(224):
    if x[i]==1:
        j+=1
    else:
        k+=1

for i in range(75):
    if y[i]==1:
        j+=1
    else:
        k+=1
print(len(x))
print(len(x_test))
print('number patients affected(malignant)',j)
print('number patients not affected(benign)',k)

print('accuracy on the training set: {:.3f}'.format(rfc.score(x_train,y_train)))
print('accuracy on the test set: {:.3f}'.format((rfc.score(x_test,y_test))))
```

```
224
75
number patients affected(malignant) 136
number patients not affected(benign) 163
accuracy on the training set: 1.000
accuracy on the test set: 0.853
```

The table show the final accuracy after optimizing the algorithm by changing the parameter of the function.

ALGORITHMS	TRAINING SET	TESTING SET
KNN-k nearest neighbor algorithm	0.857	0.813
SVM-support vector machine	0.862	0.813
Logistic regression	0.857	0.867
Random Forest	1.00	0.853

From the accuracy table we found that multi-layer Random Forest Classifier provides best results. It has detected 136 patients as affected out of 299 patients and actual patient affected is 139 patients this algorithm detects most of the patient.

CHAPTER 5

SUMMARY AND CONCLUSIONS

An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. . we used the jupyter notebook for implementing the ML algorithms like K-Nearest Neighbors (KNN), Support Vector Machine (SVM) , logistic regression and random forest . From the implementation of these algorithm provide different set of accuracies for testing and training cancer dataset. We have observed that training dataset Random forest classifier provides the best result and accuracy It predicts almost all the affected patient. Logistic regression also provide better accuracy but it did not meet the maximum accuracy we have successfully implemented the rest of the algorithms and cform those algorithm Random Forest classifier detect more Heart disease patients. It has detected 136 patients as affected out of 299 patients and actual patient affected is 138 patients this algorithm detects most of the patient. Thus, we have analyzed different algorithms and found that certain type of algorithm is suitable for certain type of scenario.

REFERENCES

[1] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol*, *2*, 56-66.

[2] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27-40.

[3] Vijiyarani, S., & Sudha, S. (2013). An efficient classification tree technique for heart disease prediction. In *International Conference on Research Trends in Computer Technologies (ICRTCT-2013) Proceedings published in International Journal of Computer Applications (IJCA)(0975–8887)* (Vol. 201).

[4] Patil, S. B., & Kumaraswamy, Y. S. (2009). Extraction of significant patterns from heart disease warehouses for heart attack prediction. *IJCSNS*, *9*(2), 228-235.

[5] Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, 53-59.

[6] Asmi, S. P., & Samuel, S. J. (2015). An analysis and accuracy prediction of heart disease with association rule and other data mining techniques. *Journal of Theoretical and Applied Information Technology*, *79*(2), 254.

[7] Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

[8] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, *2*(3), 220-223.

[9] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using differentdata mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

[10] Purushottama,c, Kanak Saxenab, Richa Sharmac Bhopal,2016)Efficient Heart Disease Prediction System ScienceDirect*Procedia Computer Science*, *85*, 962-969.

[11] Wadhawan, R. (2018). Prediction of coronary heart disease using Apriori algorithm with data mining classification. *International Journal of Research in Science and Technology*, *3*(1), 1-15.

[12] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, *13*(T-NANO 2014 Abstracts), 121.

[13] Kumar, M. N., Koushik, K. V. S., & Deepak, K. (2018). Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools.

[14] Shirwalkar, N., Gursalkar, S., Tak, T., & Kalshetti, A. (2018). Human heart disease prediction system using data mining techniques. *International Journal of Innovations & Advancement in Computer Science*, *7*(3), 357-360.

[15] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, *13*(T-NANO 2014 Abstracts), 121.

[16] singh,Sonika Jindal(2018).Heart Disease Prediction System using Hybrid Technique of Data. Mining Algorithms, International Journal of Advance Research, Ideas and Innovations in Technology.

[17] Navdeep Singh,Sonika Jindal,Shaheed Bhagat (2018),Heart disease prediction using Classification and feature selection techniques, International Journal of Advance Research, Ideas and Innovations in Technology

[18] Ajad Patel, Sonali Gandhi, Swetha Shetty,Prof. Bhanu Tekwani(2017)Heart Disease Prediction Using Data Mining, , International Journal of Advance Research, Ideas and Innovations in Technology

[19] Singh, G., Bagwe, K., Shanbhag, S., Singh, S., & Devi, S. (2017). Heart disease prediction using Naïve Bayes. *International Research Journal of Engineering and Technology (IRJET) e-ISSN*, 2395-0056.

[20] Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962-969.