PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

ABINAYA.B	**813816205003**
AKSHAYA PRIYA.R.K	**813816205008**
CHANDRA LEKA.P	**813816205014**
HARNI.R	**813816205018**

in partial fulfillment for the award of the degree

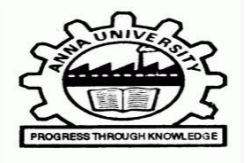
of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

SARANATHAN COLLEGE OF ENGINEERING, TIRUCHIRAPALLI



ANNA UNIVERSITY : CHENNAI 600 025

SEPTEMBER 2020

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report "**PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING"** is the bonafide work of

ABINAYA.B	**813816205003**
AKSHAYA PRIYA.R.K	**813816205008**
CHANDRA LEKA. P	**813816205014**
HARNI.R	**813816205018**

Who carried out the project work under my supervision.

Dr.R.SUMATHI M.E., Ph.D.,
HEAD OF THE DEPARTMENT
Professor
Department Of Information Technology
Saranathan College of Engineering
Panjappur
Trichy - 620012

MR.P.ANAND,M.E.,
SUPERVISOR
Assistant Professor
Department Of Information Technology
Saranathan College of Engineering
Panjappur
Trichy - 620012

VIVA – VOICE EXAMINATION

PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

ABINAYA.B	**813816205003**
AKSHAYA PRIYA.R.K	**813816205008**
CHANDRA LEKA.P	**813816205014**
HARNI.R	**813816205018**

The viva – voice Examination of this project work done as a part of B.Tech Information Technology was held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We sincerely thank **Shri. S. RAVINDRAN, Secretary, Saranathan College of Engineering,** for giving us a platform to realize our project.

We express our sincere thanks to **Dr. D. VALAVAN Ph.D., Principal, Saranathan College of Engineering,** for giving us an opportunity and immense support for the successful completion of the project.

We are obliged **to Dr. R. SUMATHI, Professor, M.Tech., Ph.D., Professor & Head of Department, Information Technology, Saranathan College of Engineering,** for her valuable suggestion and encouragement to our project. We express our heartfelt thanks to our project coordinator and our project guide **Mr .P. ANAND, M.E., Assistant Professor, Department of Information Technology, Saranathan College of Engineering,** for us with his valuable ideas.

We would like to thank all our department faculty members and technical assistant for their support and help rendered by them in completion of this project .We are also thankful to the department for providing Wi-Fi connection throughout our project period.

We would like to thank our parents for their constant encouragement and support without which this project would not be possible. Last but not the least we would like to thank our friends who have been instrumental in providing idea and material for the construction of our project.

Above all, we thank the God almighty for his bountiful blessings.

ABSTRACT

This Project describes a method to detect possible heart disease using Machine learning algorithm. In recent years, leading cause of death for both men and women is cardiovascular disease. proactive predication of risk of heart diseases will mitigate the situation to a great extent. This can be achieved by automating the prediction of heart diseases by saving time and effort. The recent development in medical field has important role in predicting heart diseases. Machine learning is an application of Artificial Intelligent that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The main contribution of this project is to identify whether a patient has heart disease or not with help of Random Forest Algorithm(RFA) is a combination of many binary decision trees and it is an supervised learning model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

WHO	World Health Organization
HDC	Health Discovery Corporation
ECG	Electrocardiogram
RFA	Random Forest Algorithm
SVM	Support Vector Machine
KDD	knowledge Discovery in Database
EMR	Electronic Medical Records

CHAPTER 1

INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

The World Health Organization (WHO) classifies cardiovascular diseases as the number one cause of death globally. In total, 17.9 million people died from cardiovascular diseases in 2016, representing 31% of all global deaths. Cardiovascular diseases are disorders of the heart and blood vessels. Four out of five cardiovascular diseases deaths are due to heart attacks and strokes. Individuals at risk of cardiovascular diseases may demonstrate raised blood pressure, be overweight or obese.

Among the adult population, cardiovascular diseases are the main health problem in general. It mainly affects the heart and the arteries of the brain, heart and legs. Therefore, the lack of blood supply not only damages the heart, but also the legs and brain, which can lead to health disorders prompting a risk of heart attacks, thrombosis or rupturing of blood vessels, among others. The main risk factors were defined in the Framingham Heart study published in 1952 and are listed as follows

- Age
- Gender
- Body Mass Index
- Smoking Condition
- Homocysteine
- Reactive C-Protein
- Fibrinogen
- Previous familiar cases
- Diet

- Cholesterol HDL Triglycerides Lipoprotein
- Sedentary Condition
- Glucose Tolerance and Metabolic System
- High blood pressure

The WHO defines unhealthy diet, physical inactivity, tobacco use and excessive use of alcohol as the most important behavioral risk factors for heart disease. These “intermediate risk factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack and other complications Some of this information can be provided immediately, while in the other cases tests need to be done. These can include blood tests or an electrocardiogram. An electrocardiogram is a diagnostic tool that is routinely used to measure and record different electrical potentials of the heart. Willem Einthoven developed the ECG method in the early 1900s, and while it is a relatively simple test to perform, the interpretation of ECG tracing requires a significant amount of training

The P wave of the ECG looks at the atria. The QRS complex looks at the ventricles and the T wave evaluates the recovery stage of the ventricles while they are refilling with blood. The ST slope and ST depression, induced by exercise, is part of the database which is used for the method in this paper.

Generally, many health care organizations are facing a major challenge to offer high quality provisions, like diagnosing patients correctly and administrating treatment at reasonable costs. Machine learning techniques have been widely used to mine information from medical databases. In Machine Learning, classification (e.g.: is this specific patient sick or healthy) is a supervised form of learning that can be used to design models describing important data classes. Using those machine learning techniques can support researchers or physicians

in making medical decisions and they can answer important and related questions concerning health care.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy

Table 1.1 Attribute Description

1.2 SYSTEM DESCRIPTION

Method to detect possible heart disease using the Random Forests algorithm. Cardiovascular diseases are the number 1 cause of death globally - an estimated 17.9 million people died from it in 2016. This machine learning work contributes to healthcare and can detect heart disease on the basis of clinical data and test data from different patients. The result and contribution of this paper is to identify whether a patient has heart disease or not, based on the information of clinical data and test results and so support doctors in making decisions about patient treatments.

1.3Problem Statement

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

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz &

Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still "information rich" but "knowledge poor". There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

1.4 Existing System

The Existing System uses SVM techniques to predict the heart disease in machine learning. this method is to develop an efficacious treatment using data mining techniques that can help remedial situations. Further data mining classification algorithms like decision trees, neural networks, Bayesian classifiers, Support vector machines, Association Rule, K- nearest neighbour classification are used to diagnosis the heart diseases.

Disadvantage of Existing System

- The main disadvantage of the SVM algorithm is that it has several key parameters that need to be set correctly to achieve the best classification results for any given problem.
- Parameters that may results is an excellent classification accuracy for problem A, may result in a poor classification accuracy for problem B**.**

1.5 Proposed System

This machine learning work contributes to healthcare and can detect heart disease on the basis of clinical data and test data from different patients. this method is to identify whether a patient has heart disease or not, based on the information of clinical data and test results and so support doctors in making

decisions about patient treatments. The result and contribution of this paper is to identify whether a patient has heart disease or not, based on the information of clinical data and test results and so support doctors in making decisions about patient treatments. Using Random Forest Algorithm result can be accurately analyzed.

Feature Extraction

In data mining, the past is explained and future is predicted by means of data analysis. This field is a combination of statistics, machine learning, artificial intelligence and database technology. There are plenty of applications in data mining and the most important one is disease prediction. Data Mining is a process of extracting patterns and knowledge from huge amount of data. This is called knowledge mining or extraction or Knowledge Discovery from Data (KDD).

Advantages of Proposed System

- It has an effective method for estimating missing data and maintaining accuracy when large proportion of the data are missing.
- It gives the maximum accuracy 95% .its higher than compare SVM method**.**

CHAPTER-2
LITERATURE SURVEY

CHAPTER-2

LITERATURE SURVEY

2.1 Title: Heart Disease Diagnosis and Prediction Using Machine Learning And Data Mining techniques

Authors: Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee.

The heart is one of the main organs of the human body. It pumps blood trough the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. A popular saying goes that we are living in an "information age". Terabytes of data are produced every day. Data mining is the process which turns a collection of data into knowledge. The health care industry generates a huge amount of data daily. However, most of it is not effectively used. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes are not much prevalent. These works show that rather than applying a single mining technique on a data set, results are far better if a collection of mining techniques are used. Java is chosen in most of the research work for practical execution of the project. The aim of this paper is to summarize some of

the current research on predicting heart diseases using data mining techniques, analyse the various combinations of mining algorithms used and conclude which technique(s) are effective and efficient. Also, some future directions on prediction systems have been addressed.

Year: July 2017

Algorithm: classification techniques.

Advantages: combination of mining techniques and accurate implementation of those techniques on the data set yields a fast and effective implementation of a system for heart disease management

Disadvantages: It has been observed that a properly cleaned and pruned dataset provides much better accuracy than an unclean one with missing values

2.2 Title: Medical Data Mining Method to Predict Risk Factors of Heart Attack and Raise Early Warning to Patients

Authors: M. Ilayaraja and Dr. T. Meyyappann

The healthcare sector faces strong pressures to reduce costs while increasing quality of services delivered. The healthcare domains have a lot of challenges and difficulties in diagnosing diseases. Data mining techniques are used to determine buried information that is useful to healthcare practitioners in effective decision making. In this paper, authors developed a new method to discover the frequent item sets to predict most risk factors for heart attack by analyzing the existing Electronic Medical Records. The new method proposed in this research work will enable medical practitioners to give early warning for patients who are likely to be affected by heart disorders. Coronary Heart

Disease affects people when the plaque builds up inside the coronary arteries which supply heart muscle with oxygen-rich blood. In blood, the plaque is identified which is made up of cholesterol, calcium, fat and other substances. It hardens and narrows the arteries. Finally it reduces blood flow to the heart muscle. And an area of plaque can rupture and leads to blood clot. The flow of oxygen-rich blood to a part of the heart muscle is blocked while the clot becomes large. As a result, the patient is affected by angina or heart attack. Angina is also known as heart attack.

Year: July 2015

Algorithm: Apriori and Cluster Based Association Rule mining Methods.

Advantage: The efficiency of level-wise generation of frequent item sets is improved which helps to reduce the search space.

Disadvantage: Association rule mining requires the minimum support value and minimum confidence value.

2.3 Title: Data Mining Approach to Detect Heart Diseases

Authors: Vikas Chaurasia (Research Scholar, Sai Nath Universityy)

Saurabh pal (Dept.of MCA,VBS Purvanchal Univercity)

The healthcare industry gathers enormous amounts of heart disease data which, unfortunately, are not “mined” to discover hidden information for effective decision making. The reduction of blood and oxygen supply to the heart leads to heart disease. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases

using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. 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 . This research paper aims to analyze the several data mining techniques proposed in recent years for the diagnosis of heart disease.

Each data mining technique serves a different purpose depending on the modeling objective. Naive Bayes is one of the successful data mining techniques used in the diagnosis of heart disease patients . Naive Bayes classifiers have works well in many complex real-world J48 Decision Tree is a popular classifier which is simple and easy to implement. J48 Decision Tree with reduced error. It requires no domain knowledge or parameter setting and can handle high dimensional data. In the diagnosis of heart disease large number of work is carried out, researchers have been investigating the use of data mining techniques to help professionals. Many risk factors associated with heart disease like age, sex, chest pain, blood pressure, cholesterol, blood sugar, family history of heart disease, obesity, and physical inactivity.

Year: November 2013

Algorithm: Naive bayes,J48 decision tree.

Advantage: Naïve bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

Disadvantage: Data scarcity chances of loss of accuracy. zero frequency that is if the category of any categorical variable is not seen in training data set then

model assigns a zero probability to that category and then a prediction cannot be made.

2.4 Title: Heart Disease Prediction Using Data Mining Classification

Author: K.Gomathi1, Dr. Shanmugapriyaa2

Heart disease is the leading cause of death in the U.S. Heart disease can strike suddenly and require you to make decisions quickly. In this paper analyses the heart disease predictions using classification algorithms. They present an analysis of the Heart disease for male patients using data mining techniques. The pre-processed data set consists of 210 records, which have all the available 8 fields from the database. In order to carry out experimentations and implementations weka was used as the data mining tool. They using three data mining techniques to predict heart disease, they are naive Bayes, Artificial neural networks and Decision tree (J48).The accuracy of three data mining techniques is compared. The goal is to have high accuracy. In this paper they discuss some of effective techniques that can be used for heart disease classification and accuracy of classification techniques is evaluated based on selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for medical applications. The performance of Naive Bayes shows high level compare with other classifiers.

Year: Jan 2016

Algorithms: Naive Bayes, Artificial neural networks and Decision tree(J48).

Advantages: In this paper they predict using three data mining techniques are compared in order to calculate accuracy.

Disadvantages: This paper only designed to predict disease for males. They uses limited records to analyse.

2.5 Title: Automated Diagnosis of Heart Disease using Data Mining Techniques

Author: Prof. Priya R. Patil and Prof. S. A. Kinariwala

The accurate diagnosis of a heart diseases, is one of the most important biomedical problems whose administration is imperative. An important task of any diagnostic system is the process of attempting to determine and/or identify a possible disease or disorder and the decision reached by this process. In this paper machine learning algorithms are used. As there are number of data's they need classifier to classify data efficiently .And also for clustering and prediction purpose classification models are used. They are decision trees, neural networks, Bayesian classifiers, Support vector machines, Association Rule, Ensemble techniques are used to diagnosis the heart diseases. This system mainly focuses on the supervised learning technique called the Random forests for classification of data by changing the values of different hyper parameters in Random Forests Classifier to get accurate classification results.

Year: November 2017

Algorithms: Ensemble classifier

Advantages: It doesn't include any tuning parameters.

Disadvantages: Ensemble based classification research has two main criticism- the dearth of publicly available real data to perform the experiments on and published well researched methods and techniques.

CHAPTER 3

SOFTWARE REQUIREMENTS

CHAPTER 3
SOFTWARE REQUIREMENTS

3.1 INTRODUCTION

3.1.1 PURPOSE

This software requirement specification provides a complete description of all the function and specifications of efficient machine learning based detection of heart disease.

3.1.2 SCOPE

The scope of this project is aimed at developing of "Efficient machine learning based on detection of heart disease".

3.2 SYSTEM ENVIRONMENT

Fig 3.1 System Environment

3.3 SPECIFIC REQUIREMENTS

3.3.1 Functional Requirements

Functional requirements are those that refer to functionality of the system (i.e) What service it will provide ,to other information needed to produce the correct system, are detailed separately.

Use Case Model:

1. Preprocessing
2. Cross Validation
3. Random forest training & Testing
4. Predicting

Preprocessing

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

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

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

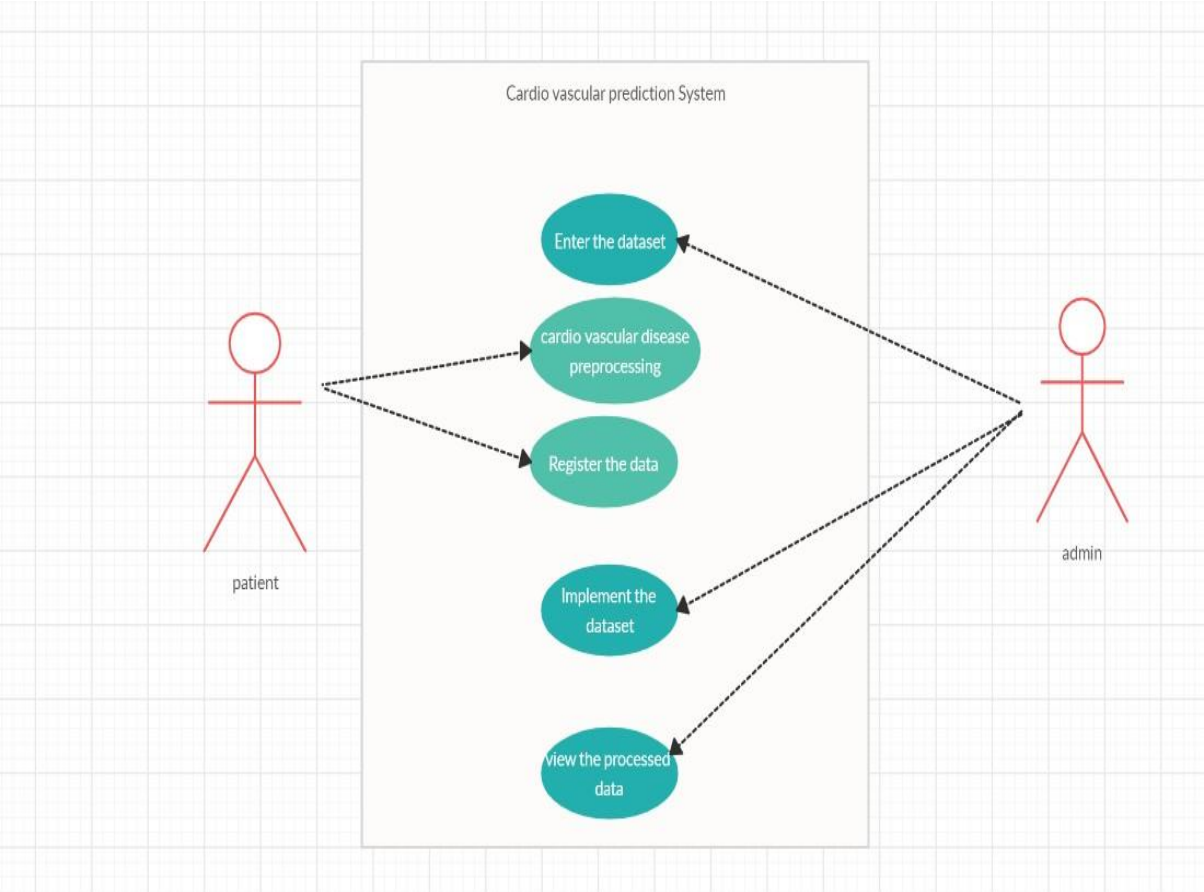


Fig 3.2 Preprocessing

Database upload

For Supervised Machine Learning Algorithms there are multiple techniques. Some examples include: Nearest Neighbor, which classifies a set of test data based on the k Nearest Neighbor algorithm using the training data . Naive Bayes, which is the simplest form of Bayesian network calculates a set of probabilities by counting the frequency and combinations of values in a given data set. Support Vector Machines output an optimal hyper plane which categorizes new examples between labeled training data. The Decision Tree is a tree-based flowchart model, in which each internal node represents a “test” on an attribute. Each branch represents the outcome of the test and the leaves are a

class distribution. The different paths from the root to a leaf represent a classification rule.

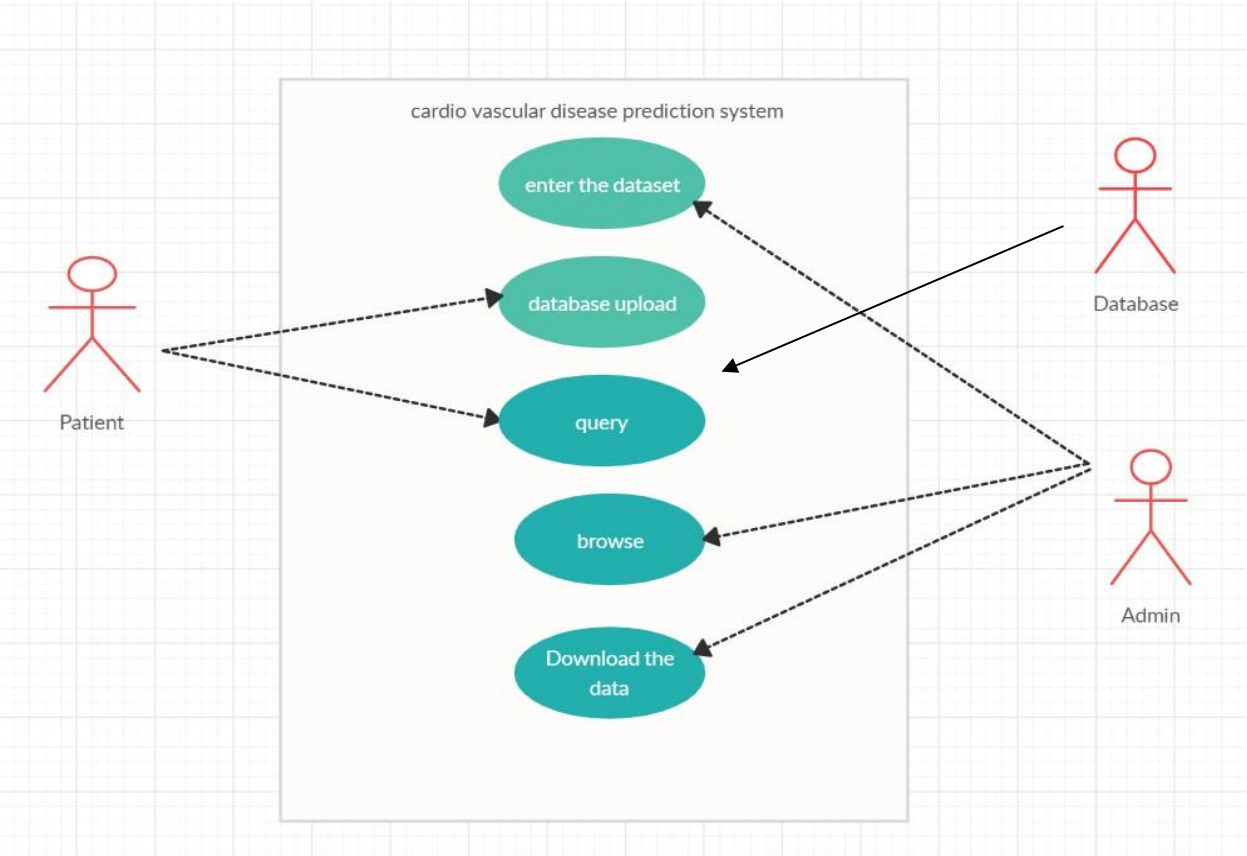


Fig 3.3 Database upload

Cross Validation

In order to achieve a reliable result from the Random Forests, cross-validation was used. Cross validation divides the data set into a specific number of subsets. Each subset is used by repeating both as a training record and as a test record. The error estimates of all rounds are then summarized and averaged [18]. As used in the method by Rieg et al., a 10 times 10-Cross-Validation was applied. As the result, the algorithm revealed which subjects were correctly classified. For this purpose, a confusion matrix was generated, and the cross validation classified as follows [19]:

- True positive: The subject has heart disease and the algorithm has correctly indicated it.

- False negative: The subject has heart disease, but the model has falsely classified him as being without heart disease.

- False positive: The patient does not have heart disease, but the model has classified him as a person with heart disease.

- True negative: The patient does not have heart disease and the algorithm has not classified him as a person with heart disease either

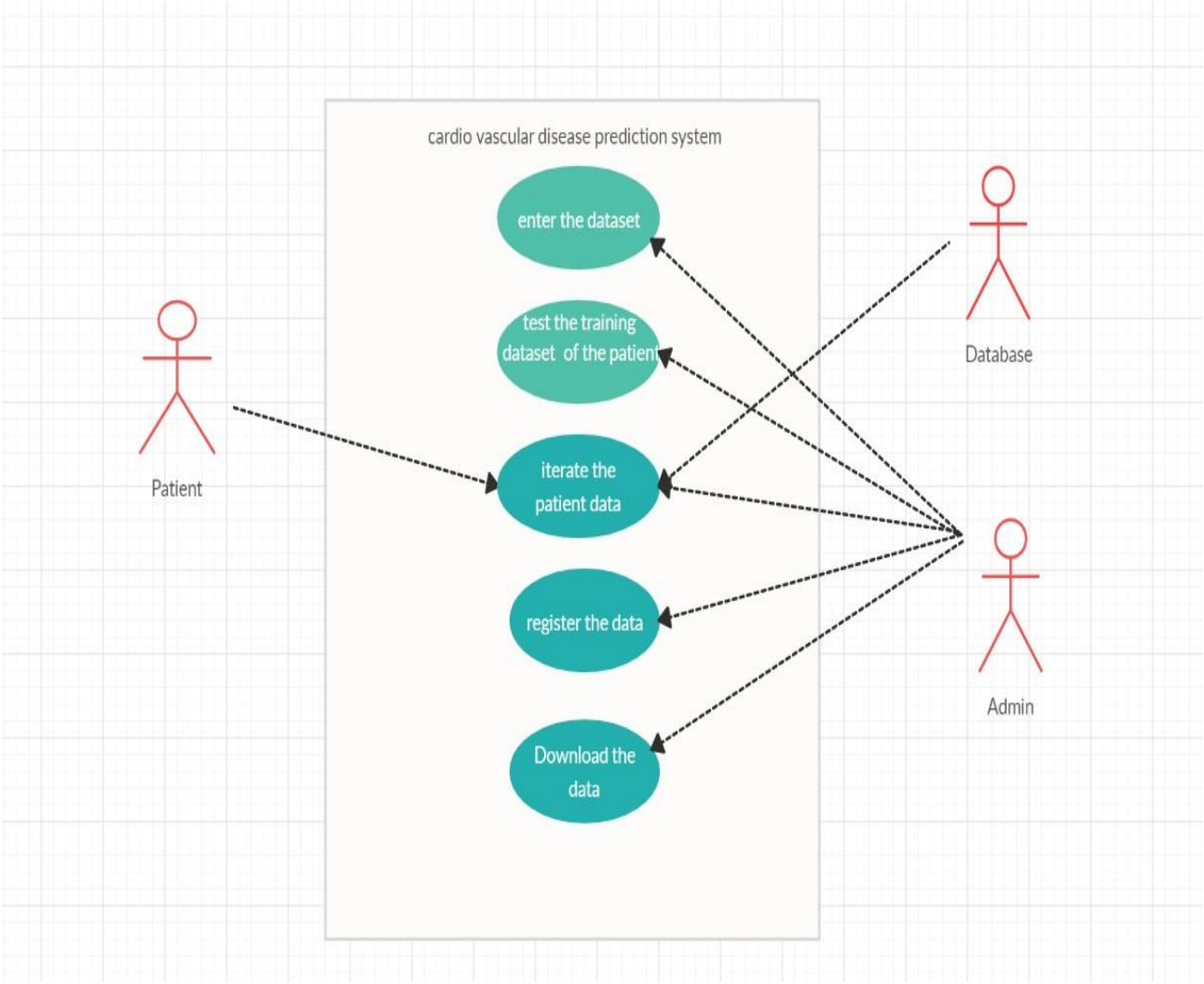


Fig 3.4 Cross validation

Random forest training and testing

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Random forests is great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features

But as stated, a random forest is a collection of decision trees. ... With that said, random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results.

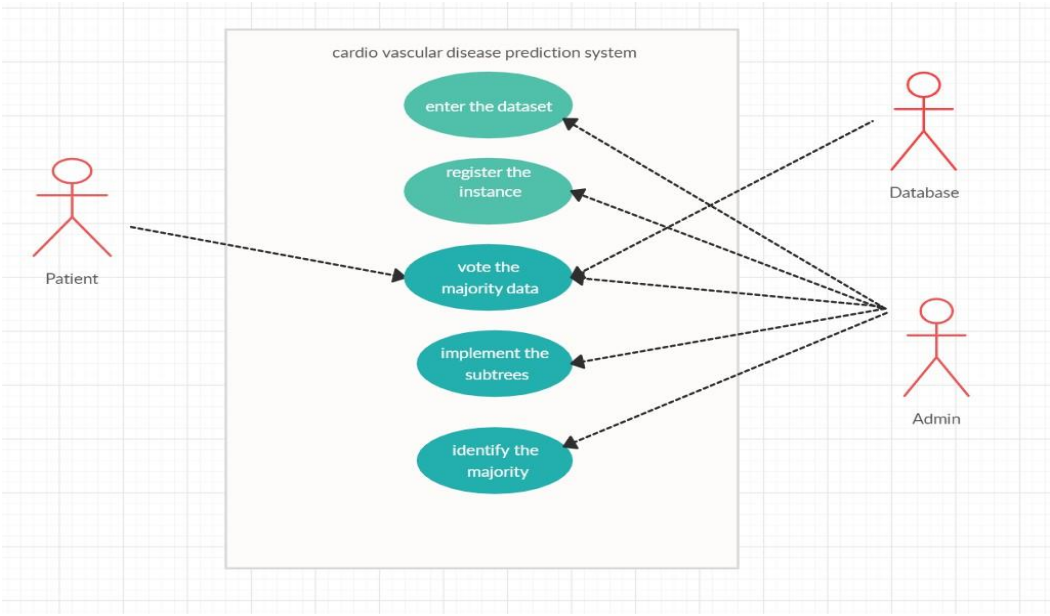


Fig 3.5 Random forest training and testing

The Decision Tree is a tree-based flowchart model, in which each internal node represents a “test” on an attribute. Each branch represents the outcome of the test and the leaves are a class distribution. The different paths from the root to a leaf represent a classification rule.

The machine learning technique used in this paper is the Random Forests. It is used to classify whether a person has a heart disease or not, based on clinical information and test results about a group of patients.

The Random Forests algorithm is a popular and very efficient algorithm, for both classification and regression problems. The principle of Random Forests is to combine many binary decision trees by using several bootstrap samples coming from the learning sample and choosing randomly at each node a subset of variables.

Predict Accuracy

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

Fig 3.6 Prediction

3.3.2 Non-Functional Requirements

Non functional Requirements are the requirements that are non functional in nature. Specifically these are the constraints that the system must work within.

SYSTEM REQUIREMENT SPECIFICATION

Hardware Requirements Processor

: Intel or high

Space on disk : minimum 100mb

For running the application

Device : any device with internet

Minimum space : 20mb

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Software Requirements

Operating System	Any OS with clients to access the internet
Network	Wi-Fi Internet or cellular Network
Visio Studio	Create and design Data Flow and Context Diagram
GitHub	Versioning Control
Google Chrome	Medium to find reference to do system testing, display and run jupyter notebook

Table 3.1 Software Requirements

3.4 SOFTWARE QUALITY ATTRIBUTES

3.4.1 Maintainability

The ability of the system to do the work for which it is intended. The patient list or the UCI data should be maintained.

3.4.2 Security

The patient information is protected. The attribute data are stored in an order. Thousands of data are maintained without collision.

3.4.3 Usability

In order to train the patient, the system provide user friendly.

3.4.4 Reliability

The system will perform the process at the time period without any error.

3.4.5 Availability

The software used to ensure that systems are running and available most of the time. High availability is a high percentage of time that the system is functioning.

CHAPTER 4

SYSTEM DESIGN

CHAPTER 4

SYSTEM DESIGN

4.1 INTRODUCTION

The WHO defines unhealthy diet, physical inactivity, tobacco use and excessive use of alcohol as the most important behavioral risk factors for heart disease. These “intermediate risk factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack and other complications. Some of this information can be provided immediately, while in the other cases tests need to be done. These can include blood tests or an electrocardiogram. An electrocardiogram is a diagnostic tool that is routinely used to measure and record different electrical potentials of the heart. Willem Einthoven developed the ECG method in the early 1900s, and while it is a relatively simple test to perform, the interpretation of ECG tracing requires a significant amount of training The P wave of the ECG looks at the atria. The QRS complex looks at the ventricles and the T wave evaluates the recovery stage of the ventricles while they are refilling with blood. The ST slope and ST depression, induced by exercise, is part of the database which is used for the method in this paper.

Generally, many health care organizations are facing a major challenge to offer high quality provisions, like diagnosing patients correctly and administrating treatment at reasonable costs. Machine learning techniques have been widely used to mine information from medical databases. In Machine Learning, classification (e.g.: is this specific patient sick or healthy) is a supervised form of learning that can be used to design models describing important data classes. Using those machine learning techniques can support researchers or physicians

in making medical decisions and they can answer important and related questions concerning health care.

4.2 ARCHITECTURE

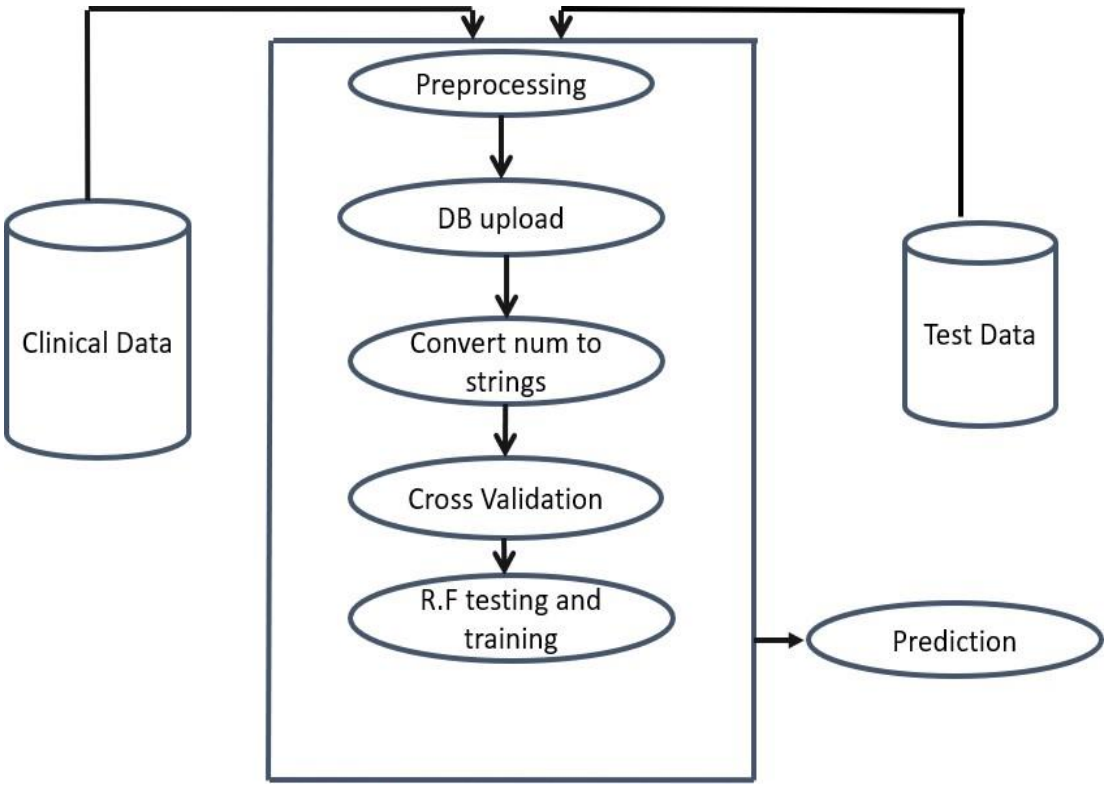


Fig 4.1 System Architecture

4.3 SYSTEM DESIGN DESCRIPTION

In the phase ,The UCI data attributes are collected and their features are extracted. The extracted features are stored as input and passed to the classifier by using the extracted features.

When the patient data are collected from uci data, this data are in exel or cv file format. The random forest algorithm is used to be predict the accuracy.

4.4 DATA FLOW

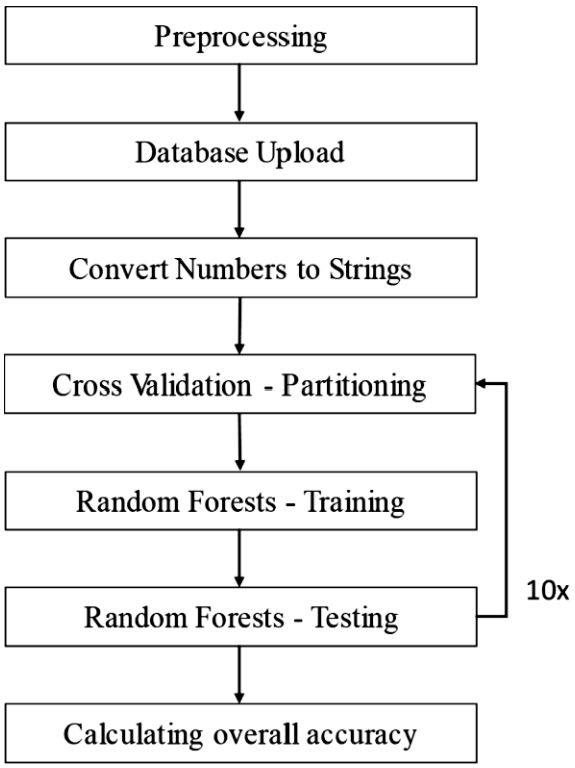


Fig 4.2 Data flow diagram

In the diagram is a way of representing a flow of process. The Preprocessing is the initial process and the prediction is decided in the section . The Uci data are collected and upload to in this section. The patients data are to be compared to the threshold value. Patients details are insert to random forest algorithm for getting more accuracy. Uci data was stored the database and the values are retrieve easily. Build a decision tree and the values are inserted. Then tested the random forest in python. The each and every values are to be tested. At last the accuracy will be given, its more than others algorithm. Finally we calculating overall accuracy.

CHAPTER 5
IMPLEMENTATION

CHAPTER-5

IMPLEMENTATION

5.1 MODULE IMPLEMENTATION

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

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

Modules

- 1.Preprocessing
- 2.Cross Validation
- 3.Random forest training & Testing
- 4.Prediction

5.1.1 Preprocessing

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

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

Code

{

Text/plain:[pandas.core.frame.Dataframe”]

}

<style scoped>\n,

.dataframe tbody tr th {\n,vertical-align: middle;\n”

Info=[“age”,”1”:male, “0”:female,”chest pain type”,1: typical angina, 2:atypical angina, 3:non-anginal pain, 4:asymptomatic”, “resting blood pressure”, “serum cholestral in mg\dl”, “fasting blood sugar>120 mg\dl”, “resting electrocardiographic results(values 0,1,2)”, “maximum heart rate achived”, exercise induced angina”, “oldpeak = st depression induced by excersise relative to rest”, “the slope of the peak exercise ST segment”, “number of majar vessels(0-3) colored by flourosopy”, “thal:3=normal; 6=fixed defect; 7=reversible defect”]\n”,

For I in range(len(info)):\n”,

Print(dataset.coloumns[i]+\”:\\t\\t\\t”+info[i])”

]

}

Database Upload

Nearest Neighbor, which classifies a set of test data based on the k Nearest Neighbor algorithm using the training data ,Naive Bayes, which is the simplest form of Bayesian network calculates a set of probabilities by counting the frequency and combinations of values in a given data set. Support Vector Machines output an optimal hyper plane which categorizes new examples between labeled training data.

Code

Dataset= pd.read_csv(\"heart.csv"\)

dataset.shape

data:{

text/html:[

\n,

<style scoped>\n

.dataframe tbody tr th {\n,

Vertical-align:middle;\n,

}\n,

\n,

<tbody>\n,

</thread>\n,

<tr style=\”text-align:right;\”>\n”,

<th></th>\n

<th>age</th>\n

<th>sex</th>\n

<th>cp</th>\n

<th>trestbps</th>\n

<th>chol</th>\n

<th>fbs</th>\n

<th>restecg</th>\n

<th>thalach</th>\n

<th>exang</th>\n

<th>oldpeak</th>\n

<th>slope</th>\n

<th>ca</th>\n

<th>thal</th>\n

<th>target</th>\n

<\tr>\n,

</theard> </tbody>

5.1.2 Cross Validation

In order to achieve a reliable result from the Random Forests, cross-validation was used. Cross validation divides the data set into a specific number of subsets. Each subset is used by repeating both as a training record and as a test record. The error estimates of all rounds are then summarized and averaged [18]. As used in the method by Rieg et al., a 10 times 10-Cross-Validation was applied. As the result, the algorithm revealed which subjects were correctly classified

Code

Name : “stdout”,

Text: [

Target 1.000000\n,

Exang 0.436757\n,

Cp 0.4333798\n,

Oldpeak 0.430696\n,

Thalach 0.421741\n,

Ca 0.391724\n,

Slope 0.345877\n,

Thal 0.344029\n,

Sex 0.280937\n,

Age 0.225439\n,

Trestbps 0.144931\n,

Restecg 0.137230\n,

Chol 0.85239\n,

Fbs 0.028046\n,

Name : target, dtype:float64\n.

Print(dataset.corr()[\"target\"].abs().sort_values(ascending=False))"

5.1.3 Random forest training and testing

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Random forests is great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features

Code

From sklearn.ensemble import RandomForestClassifier\n,

\n,

Max_accuracy=0\n,

```
    \n,
For x in range(2000):\n,
   rf=RandomForestClassifier(random_state=x)\n,
   rf.fit(x_train , Y_train)\n,
Y_pred_rf = rf.predict(X_test)\n,
Current_accuracy = round(accuracy_score(Y_pred_rf, Y_test)*100,2)\n,
If (current_accuracy >max_accuracy):\n,
Max_accuracy = current_accuracy\n,
Best_x = x\n,
\n,
#print(max_accuracy)\n,
#print(best_x)\n,
rf=RandomForestClassifier(random_state=best_x)\n,
rf.fit(x_train,Y_train)\n,
y_pored_rf =rf.predict(x_test)
```

5.1.4 Predict Accuracy

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble

methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: random forest and classification via decision tree.

Code

[

Scores = score_lr, score_svm,score_knn,

Algorithms = [\"Naïve bayes\",\"SVM"\,\"K-Nearest

Neighbours\",\"decision tree\",\"random forest\"] \n"

\n,

For I in range(len(algorithms)):\n",

Print("the accuracy score achived using \algorithm[i]+\"is :\"+str(score[i])+\")"

]

5.2 Random Forest Algorithm

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Random forests is great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features

But as stated, a random forest is a collection of decision trees. ... With that said, random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results.

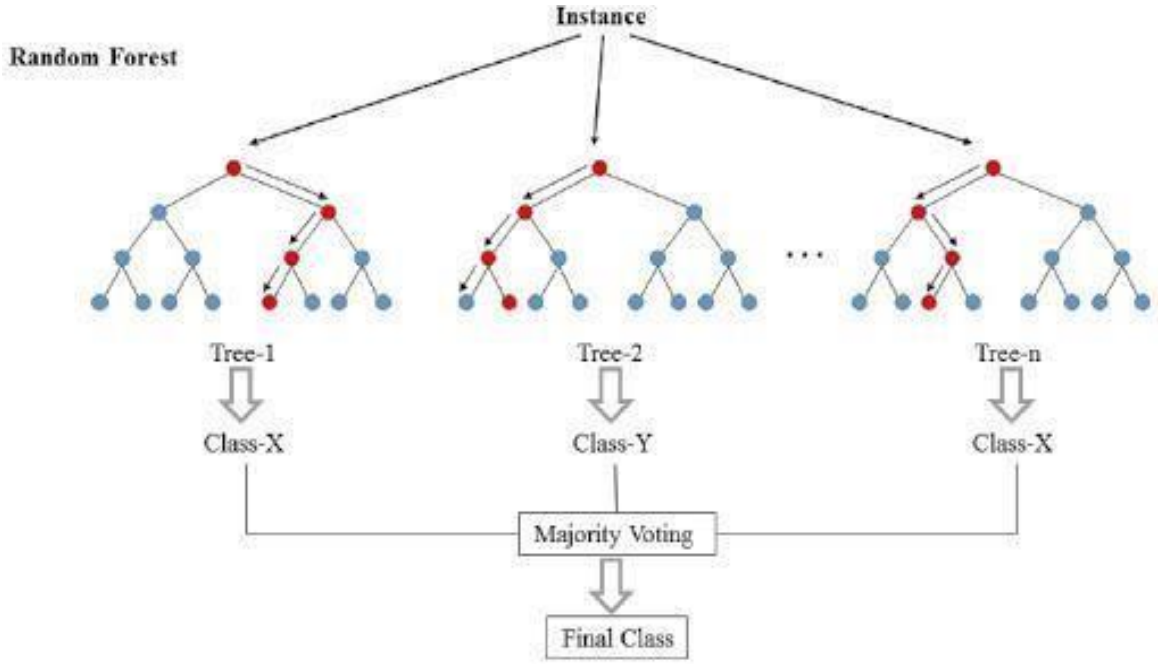


Fig 5.1 Random Forest Architecture

Techniques and Algorithm

Input: Dataset(Heart_csv), Multiple Algorithms

Output: Accuracy of the testing multiple Algorithms

1. Import files
2. Import Dataset
3. Read Dataset

4. Preprocessing
5. Test SVM algorithm and others(Knn, XGboost…)
6. Random Forest training
7. Random Forest Testing,
8. Cross Validation
9. Calculate\Predict the accuracy.
10. Compare all algorithms accuracy
11. End

Pseudo Code

```
Import numpy as np
Import pandas as panda
Import matplotlib.pyplot as plt
Import seaborn as sns
%matplotlib inline
Import os
Print(os.listdir())
Import warnings
Warnings.filter warnings(“ignore”)”
]
```

Dataset= pd.read_csv(\"heart.csv"\)

Text/plain:[pandas.core.frame.Dataframe"

]

}

<style scoped>\n,

.dataframe tbody tr th {\n,vertical-align: middle;\n"

Info=["age","1":male, "0":female,"chest pain type",1: typical angina, 2:atypical angina, 3:non-anginal pain, 4:asymptomatic", "resting blood pressure", "serum cholestral in mg\dl", "fasting blood sugar>120 mg\dl", "resting electrocardiographic results(values 0,1,2)", "maximum heart rate achived", exercise induced angina", "oldpeak = st depression induced by excersise relative to rest", "the slope of the peak exercise ST segment", "number of majar vessels(0-3) colored by flourosopy", "thal:3=normal; 6=fixed defect; 7=reversible defect"]\n",

For I in range(len(info)):\n",

Print(dataset.coloumns[i]+\":\\t\\t\\t"+info[i])"

Text: [6

Target 1.000000\n,

Exang 0.436757\n,

Cp 0.4333798\n,

Oldpeak 0.430696\n,

Thalach 0.421741\n,

Ca 0.391724\n,

Slope 0.345877\n,

Thal 0.344029\n,

Sex 0.280937\n,

Age 0.225439\n,

Trestbps 0.144931\n,

Restecg 0.137230\n,

Chol 0.85239\n,

Fbs 0.028046\n,

Name : target, dtype:float64\n.

Print(dataset.corr()[\"target\"].abs().sort_values(ascending=False))"

From sklearn.ensemble import RandomForestClassifier\n,

\n,

Max_accuracy=0\n,

\n,

For x in range(2000):\n,

rf=RandomForestClassifier(random_state=x)\n,

rf.fit(x_train , Y_train)\n,

Y_pred_rf = rf.predict(X_test)\n,

Current_accuracy = round(accuracy_score(Y_pred_rf, Y_test)*100,2)\n,

If (current_accuracy >max_accuracy):\n,

Max_accuracy = current_accuracy\n,

Best_x = x\n,

\n,

#print(max_accuracy)\n,

#print(best_x)\n,

rf=RandomForestClassifier(random_state=best_x)\n,

[

Scores = score_lr, score_svm,score_knn,score_dt,score_rf,score_xgb,score-nn]\n

Algorithms = [\"Logistic Regression\", \"Naïve bayes\",\"SVM"\,\"K-Nearest

Neighbours\",\"decision tree\",\"random forest\",\"XGboost\",\"neural network\"] \n"

\n,

For I in range(len(algorithms)):\n",

Print("the accuracy score achived using \algorithm[i]+\"is :\"+str(score[i])+\")"

]

CHAPTER 6

TESTING

CHAPTER-6

TESTING

6.1 TESTING PROCESS

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to heck the functionality of components, sub assemblies, and/or a finished product.

It's the process of exercising software with intent of ensuring the software system meets its requirements and user expectations and does not fall in an unacceptable manner.

6.1.1 Test Objectives

Predict heart diseases

Using Random forest algorithm

Get high accuracy

6.1.2 Features to be tested

Huge amount of data tested

6.2TYPES OF TESTING

A test case is an asset of data that the system will process as normal input.

6.2.1.Unit Testing

Unit testing involves the design of test case that validate the internal logic is functioning properly, and that program inputs produce valid outputs. All

decision branches and internal code flow should be validated. It is the testing of individuals software units of the application. It is done after the completion of an individual units before integration. This is a structural testing, that replies on knowledge of its construction and is invasive. unit test perform basic tests at component level .

6.2.2 Integration Test

Integration test are designed to test individual software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screen or fields. Integration tests demonstrate that although the components were individually satisfacation, as shown by successfully unit testing, the combination of components is correct and consistent.

6.2.3 Functional Testing

Functional test provides systematic demonstration that functions tested are available as specified by the business and technical requirements ,system documentation and user manuals.

Functional testing is centered on the following items:

Valid Input : Identified classes of valid input must be accepted.

Invalid Input : Identified classes of valid input must be rejected.

Functions : identifications function must be exercised.

Output : Identified classes of application outputs must be exercised.

Procedure : interfacing systems or procedure must be invoked.

6.2.4 System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known predictable results. An example of testing is the configuration oriented systems integration test. System testing is based on process description and flows, emphasizing pre-driven process links and integration points.

6.2.5 Acceptance Testing

User acceptance testing is a critical phase of any project and requires significant participations by the end user. It also ensures that the systems meets the functional requirements.

6.3 TEST CASES

6.3.1 Admin

Sl.no	Testcase	Input	Expected Output	Actual Output	Result
1	Dataset upload	CSV file(attributes)	Import successfully	Import successfully	Pass
2	Train Split	Attributes assign	Final Ranking R	Final Ranking R	Pass
3	Split points	Set of attributes	Return values	Return values	Pass
4	Algorithm	R.F algorithm	Accuracy 95%	Accuracy 95%	Pass

Table 6.3.1 Admin Testcase

6.3.2 Customer

Sl.no	Testcase	Input	Expected Output	Actual Output	Result
1	Dataset Upload	Dataset	2000 data and 14 attributes in the data set	2000 data and 14 attributes in the data set	pass
2	Preprocess	Attributes	Null values removed	Null values removed	pass
3	Training	Confusion Matrix	The number of correcting entries by the classifier	The number of correcting entries by the classifier	pass
4	Algorithm	SVM	Accuracy given	Accuracy given	pass
5	Result	Random Forest	95% Accuracy	95% Accuracy	pass

Table 6.3.1 Custmor Testcase

CHAPTER 7

CONCLUSION AND FUTURE WORK

CHAPTER-7

CONCLUSION AND FUTURE WORK

This paper presented a new approach to heart disease classification, using the Random Forest machine learning algorithm and attributes based on clinical data and patient test results. It reached an overall accuracy of 84.448%. The highest accuracy was reached while using an additional 10 times cross-validation in the process and it outperforms other machine learning techniques using the same database. Using the Random Forests algorithm without the cross validation secured an overall accuracy of 95%.

While we intensively evaluated other traditional machine learning approaches such as clustering and also most modern convolutional neural networks, which are outstanding in other domains such as image recognition, we achieved the best results here with decision trees. However, the method of choice always limits scientific understanding. Hence our study has these limitations:

To further improve the accuracy of the algorithm, updating the database with more information and attributes could help to increase the level of accuracy already achieved. As mentioned in the introduction to this paper, it is already known that Age, Gender, Body Mass Index, Smoking Condition, Homocysteine, Reactive C-Protein, Fibrinogen, Previous familiar cases, Diet, Cholesterol HDL Triglycerides Lipoprotein, Sedentary Condition, Glucose Tolerance and Metabolic System, High blood pressure are risk factors for heart attacks. But some of that relevant information is not available in the used data set .For example, beside the information about the Serum cholesterol in mg/dl, there is no information about the constitution of the patient. Adding information like weight or the Body Mass Index (BMI) could increase the information level

of the database. Also, the information about the Smoking Condition is missing. Both, tobacco use and an unhealthy diet are significant reasons for heart attacks and strokes, based on the information provided by World Health Organizations key messages. Other important information which is missing from the database is a patient's family history and other individual differences such as personality .Although all the listed reasons for heart diseases are known, their epidemiological relevance is different from case to case. Therefore, the attributes probably need to be weighed correctly. In order to estimate the risk of suffering from a heart disease, a global evaluation should be added to the information in the database as well. One solution could be the Anderson Table

In future work we will triangulate simple ECG sensor data with other physiological sensor data (i.e., heart rate variability, electroencephalography ,electrodermal activity , eye fixation , eye pupil diameter). Furthermore, we will experimentally evaluate whether our novel approach is also robust under various conditions of a user's cognitive workload , concentration , and mindfulness . In addition, we will report common method bias evaluations and the results of transferring our novel spectral method to ECG, where we already achieved outstanding results in predicting diseases such as schizophrenia , epilepsy , and sleep disorder based on electroencephalographic data. Finally, we will conduct an empirical implementation study to evaluate acceptance and trust by physicians and patients and if the automated approach improves the coordination between physicians more efficiently.

APPENDICES

APPENDIX 1

SAMPLE CODING

APPENDIX 1- SAMPLE CODING

Source code:

```
{

Text/plain:[pandas.core.frame.Dataframe”]

}

<div>

<style scoped>\n,

    .dataframe tbody tr th {\n,vertical-align: middle;\n”

Info=[“age”,”1”:male, “0”:female,”chest pain type”,1: typical angina, 2:atypical angina, 3:non-anginal pain, 4:asymptomatic”, “resting blood pressure”, “serum cholestral in mg\dl”, “fasting blood sugar>120 mg\dl”, “resting electrocardiographic results(values 0,1,2)”, “maximum heart rate achived”, exercise induced angina”, “oldpeak = st depression induced by excersise relative to rest”, “the slope of the peak exercise ST segment”, “number of majar vessels(0-3) colored by flourosopy”, “thal:3=normal; 6=fixed defect; 7=reversible defect”]\n”,

For I in range(len(info)):\n”,

Print(dataset.coloumns[i]+\”:\\t\\t\\t”+info[i])”

]

}

Dataset= pd.read_csv(\”heart.csv”\)
```

```
    dataset.shape
data:{
text/html:[
<div>\n,
<style scoped>\n
    .dataframe tbody tr th {\n,
     Vertical-align:middle;\n,
    }\n,
\n,
<tbody>\n,
</thread>\n,
<tr style=\"text-align:right;\">\n",
<th></th>\n
<th>age</th>\n
<th>sex</th>\n
<th>cp</th>\n
<th>trestbps</th>\n
<th>chol</th>\n
<th>fbs</th>\n
```

<th>restecg</th>\n

<th>thalach</th>\n

<th>exang</th>\n

<th>oldpeak</th>\n

<th>slope</th>\n

<th>ca</th>\n

<th>thal</th>\n

<th>target</th>\n

<\tr>\n,

</thread></body>

Name : “stdout”,

Text: [

Target 1.000000\n,

Exang 0.436757\n,

Cp 0.4333798\n,

Oldpeak 0.430696\n,

Thalach 0.421741\n,

Ca 0.391724\n,

Slope 0.345877\n,

Thal 0.344029\n,

Sex 0.280937\n,

Age 0.225439\n,

Trestbps 0.144931\n,

Restecg 0.137230\n,

Chol 0.85239\n,

Fbs 0.028046\n,

Name : target, dtype:float64\n.

Print(dataset.corr()[\"target\"].abs().sort_values(ascending=False))" From sklearn.ensemble import RandomForestClassifier\n,

\n,

Max_accuracy=0\n,

\n,

For x in range(2000):\n,

rf=RandomForestClassifier(random_state=x)\n,

rf.fit(x_train , Y_train)\n,

Y_pred_rf = rf.predict(X_test)\n,

Current_accuracy = round(accuracy_score(Y_pred_rf, Y_test)*100,2)\n,

If (current_accuracy >max_accuracy):\n,

Max_accuracy = current_accuracy\n,

Best_x = x\n,

\n,

#print(max_accuracy)\n,

#print(best_x)\n,

rf=RandomForestClassifier(random_state=best_x)\n,

rf.fit(x_train,Y_train)\n,

y_pored_rf=rf.predict(x_test)

[

Scores = score_lr, score_svm,score_knn,

Algorithms = [\"Naïve bayes\",\"SVM"\,\"K-Nearest

Neighbours\",\"decision tree\",\"random forest\"] \n"

\n,

For I in range(len(algorithms)):\n",

Print("the accuracy score achived using \algorithm[i]+\"is :\"+str(score[i])+\")"

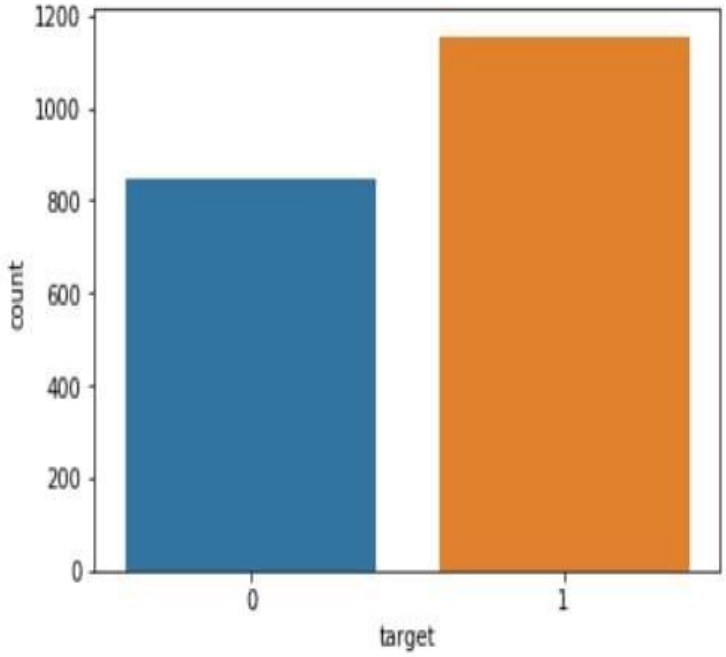
]

APPENDIX 2

SCREENSHOTS

APPENDIX 2-SCREENSHOTS

```
1    1155
0     844
Name: target, dtype: int64
```



I. Importing essential libraries

In [4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

import os
print(os.listdir())

import warnings
warnings.filterwarnings('ignore')
```

```
['.ipynb_checkpoints', 'heart.csv', 'Heart_disease_prediction-checkpoint.ipynb', 'Heart_disease_predi
ction.ipynb', 'README.md']
```

II. Importing and understanding our dataset

```
In [5]: dataset = pd.read_csv("heart.csv")
```

Printing out a few columns

```
In [5]: dataset.head(5)
```

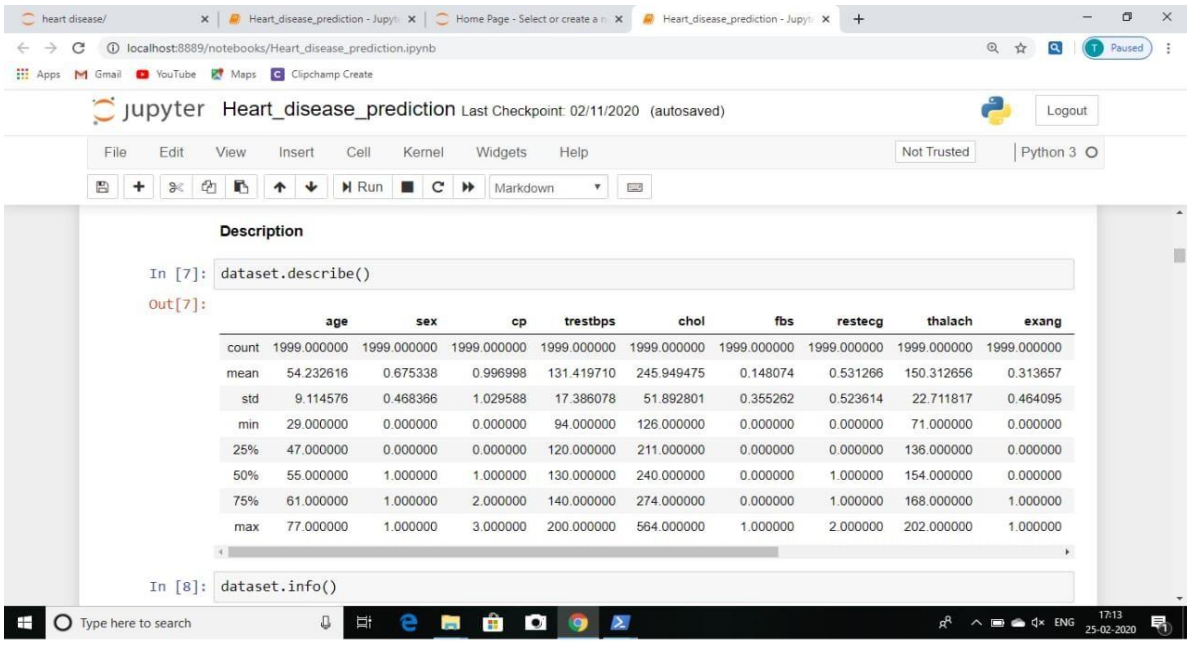
Out[5]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Verifying it as a 'dataframe' object in pandas

```
In [6]: type(dataset)
```

```
Out[6]: pandas.core.frame.DataFrame
```



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
count	1999.000000	1999.000000	1999.000000	1999.000000	1999.000000	1999.000000	1999.000000	1999.000000	1999.000000
mean	54.232616	0.675338	0.996998	131.419710	245.949475	0.148074	0.531266	150.312656	0.313657
std	9.114576	0.468366	1.029588	17.386078	51.892801	0.355262	0.523614	22.711817	0.464095
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000
25%	47.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	136.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	154.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.000000	0.000000	1.000000	168.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000

```
In [8]: dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1999 entries, 0 to 1998
Data columns (total 14 columns):
age         1999 non-null int64
sex         1999 non-null int64
cp          1999 non-null int64
trestbps    1999 non-null int64
chol        1999 non-null int64
fbs         1999 non-null int64
restecg     1999 non-null int64
thalach     1999 non-null int64
exang       1999 non-null int64
oldpeak     1999 non-null float64
slope       1999 non-null int64
ca          1999 non-null int64
thal        1999 non-null int64
target      1999 non-null int64
dtypes: float64(1), int64(13)
memory usage: 218.8 KB
```

In [10]:
```
info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non-ang

for i in range(len(info)):
    print(dataset.columns[i]+":\t\t\t"+info[i])
```

```
age:                    age
sex:                    1: male, 0: female
cp:                     chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain,
4: asymptomatic
trestbps:                       resting blood pressure
chol:                    serum cholestoral in mg/dl
fbs:                    fasting blood sugar > 120 mg/dl
restecg:                        resting electrocardiographic results (values 0,1,2)
thalach:                         maximum heart rate achieved
exang:                  exercise induced angina
oldpeak:                        oldpeak = ST depression induced by exercise relative to rest
slope:                  the slope of the peak exercise ST segment
ca:                     number of major vessels (0-3) colored by flourosopy
thal:                   thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
```

Analysing the 'target' variable

In [11]:
```
dataset["target"].describe()
```

Out[11]:
```
count    1999.000000
mean        0.577789
std         0.494035
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: target, dtype: float64
```

In [12]:
```
dataset["target"].unique()
```

Out[12]:
```
array([1, 0], dtype=int64)
```

In [13]:
```
print(dataset.corr()["target"].abs().sort_values(ascending=False))
```

```
target      1.000000
exang       0.439366
oldpeak     0.435927
cp          0.430455
thalach     0.420095
ca          0.390669
slope       0.344712
thal        0.344665
sex         0.279063
age         0.222820
trestbps    0.142456
restecg     0.140056
chol        0.083862
fbs         0.028589
Name: target, dtype: float64
```

In [15]:
```
y = dataset["target"]

sns.countplot(y)


target_temp = dataset.target.value_counts()

print(target_temp)
```

```
1    1155
0     844
Name: target, dtype: int64
```

In [21]:
```
print("Percentage of patience without heart problems: "+str(round(target_temp[0]*100/2121,2)))
print("Percentage of patience with heart problems: "+str(round(target_temp[1]*100/2121,2)))

#Alternatively,
# print("Percentage of patience with heart problems: "+str(y.where(y==1).count()*100/303))
# print("Percentage of patience with heart problems: "+str(y.where(y==0).count()*100/303))

# #Or,
# countNoDisease = len(df[df.target == 0])
# countHaveDisease = len(df[df.target == 1])
```

```
Percentage of patience without heart problems: 45.54
Percentage of patience with heart problems: 54.46
```

IV. Train Test split

In [38]:
```
from sklearn.model_selection import train_test_split

predictors = dataset.drop("target",axis=1)
target = dataset["target"]

X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20,random_state=0)
```

Logistic Regression

In [44]:
```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train,Y_train)

Y_pred_lr = lr.predict(X_test)
```

In [45]:
```
Y_pred_lr.shape
```

Out[45]: (61,)

In [46]:
```
score_lr = round(accuracy_score(Y_pred_lr,Y_test)*100,2)

print("The accuracy score achieved using Logistic Regression is: "+str(score_lr)+" %")
```

```
The accuracy score achieved using Logistic Regression is: 85.25 %
```

Naive Bayes

In [47]:
```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb.fit(X_train,Y_train)

Y_pred_nb = nb.predict(X_test)
```

In [48]:
```
Y_pred_nb.shape
```

Out[48]: (61,)

In [49]:
```
score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")
```

```
The accuracy score achieved using Naive Bayes is: 85.25 %
```

SVM

In [50]:
```
from sklearn import svm

sv = svm.SVC(kernel='linear')

sv.fit(X_train, Y_train)

Y_pred_svm = sv.predict(X_test)
```

In [51]:
```
Y_pred_svm.shape
```

Out[51]: (61,)

In [52]:
```
score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")
```

```
The accuracy score achieved using Linear SVM is: 81.97 %
```

K Nearest Neighbors

In [53]:
```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train,Y_train)
Y_pred_knn=knn.predict(X_test)
```

In [54]:
```
Y_pred_knn.shape
```

Out[54]: (61,)

In [55]:
```
score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")
```

```
The accuracy score achieved using KNN is: 67.21 %
```

Decision Tree ¶

In [56]:
```
from sklearn.tree import DecisionTreeClassifier

max_accuracy = 0


for x in range(200):
    dt = DecisionTreeClassifier(random_state=x)
    dt.fit(X_train,Y_train)
    Y_pred_dt = dt.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_dt,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)
```

```
dt = DecisionTreeClassifier(random_state=best_x)
dt.fit(X_train,Y_train)
Y_pred_dt = dt.predict(X_test)
```

In [57]:
```
print(Y_pred_dt.shape)
```

```
(61,)
```

In [58]:
```
score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")
```

```
The accuracy score achieved using Decision Tree is: 81.97 %
```

Random Forest

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

max_accuracy = 0


for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)
```

```
rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

In [60]:
```
Y_pred_rf.shape
```

Out[60]: (61,)

In [61]:
```
score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_rf)+" %")
```

```
The accuracy score achieved using Decision Tree is: 95.08 %
```

XGBoost

In [62]:
```
import xgboost as xgb

xgb_model = xgb.XGBClassifier(objective="binary:logistic", random_state=42)
xgb_model.fit(X_train, Y_train)

Y_pred_xgb = xgb_model.predict(X_test)
```

In [63]:
```
Y_pred_xgb.shape
```

Out[63]:
```
(61,)
```

In [64]:
```
score_xgb = round(accuracy_score(Y_pred_xgb,Y_test)*100,2)

print("The accuracy score achieved using XGBoost is: "+str(score_xgb)+" %")
```

```
The accuracy score achieved using XGBoost is: 85.25 %
```

VI. Output final score

In [72]:
```
scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn]
algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Deci

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 95.08 %
The accuracy score achieved using XGBoost is: 85.25 %
The accuracy score achieved using Neural Network is: 80.33 %
```

REFERENCES

[1]Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction using Machine Learning sand Data Mining Techniques",2017.

[2]M.Ilayaraja ,T.Meyyappann."Medical Data Mining Method to Predict Risk Factor of Heart Attack and Raise Early Warning to Patient",2015.

[3]Vikas Chaurasia,Saurabh Pal,"Data Mining Approach to Detect Heart Diseases",2013.

[4]K.Gomathi,Dr.Shanmugapriya,"Heart Disease Prediction Using Data Mining Classification",2013.

[5]Priya R.Patil,S.A.Kinariwala,"Automated Diagnosis of Heart Disease using Data Mining Techniques",201

Certificate of Publication

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY

The Board of
International Journal Of Innovative Research In Technology
is hereby awarding this certificate

ABINAYA.B

In recognition of the Publication of the paper entitled

PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHM

Publication In e-Journal

Volume 6 Issue 12 May 2020

PAPER ID: 149428

EDITOR IN CHIEF

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY | IJIRT
website : www.ijirt.org | email ID : editor.@ijirt.org | ISSN : 2349 - 6002

Certificate of Publication

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY

The Board of
International Journal Of Innovative Research In Technology
is hereby awarding this certificate

AKSHAYA PRIYA.R.K

In recognition of the Publication of the paper entitled

PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHM

Publication In e-Journal

Volume 6 Issue 12 May 2020

PAPER ID: 149428

EDITOR IN CHIEF

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY | IJIRT
website : www.ijirt.org | email ID : editor.@ijirt.org | ISSN : 2349 - 6002

Certificate of Publication

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY

The Board of
International Journal Of Innovative Research In Technology
is hereby awarding this certificate

CHANDRALEKA.P

In recognition of the Publication of the paper entitled

PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHM

Publication In e-Journal

Volume 6 Issue 12 May 2020

PAPER ID: 149428

EDITOR IN CHIEF

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY | IJIRT
website : www.ijirt.org | email ID : editor.@ijirt.org | ISSN : 2349 - 6002

Certificate of Publication

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY

The Board of
International Journal Of Innovative Research In Technology
is hereby awarding this certificate

HARNI.R

In recognition of the Publication of the paper entitled

PREDICTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHM

Publication In e-Journal

Volume 6 Issue 12 May 2020

PAPER ID: 149428

EDITOR IN CHIEF

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY | IJIRT
website : www.ijirt.org | email ID : editor.@ijirt.org | ISSN : 2349 - 6002