Prediction of Heart Disease Using Machine Learning

A PROJECT REPORT

submitted by

Nitesh Kumar (13BCE0789)
Harsh Vardhan (13BCE0454)

in partial fulfillment for the award

of the

B. Tech

degree in

Computer Science and Engineering

School of Computing Science and Engineering

VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ▪ CHENNAI
www.vit.ac.in

School of Computer Science and Engineering

DECLARATION

We hereby declare that the project entitled **"Prediction of heart disease using machine learning"** submitted by us to the School of Computer Science and Engineering, VIT University, Vellore-14 in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out by us under the supervision of **Prof. Meenakshi S.P, Senior Assistant Professor.** I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or university.

Signature

Nitesh Kumar (13BCE0789)

Signature

Harsh Vardhan (13BCE0454)

VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ■ CHENNAI
www.vit.ac.in

School of Computer Science and Engineering

CERTIFICATE

The project report entitled "**Prediction of heart disease using machine learning**" is prepared and submitted by **Nitesh Kumar (13BCE0789) and Harsh Vardhan (13BCE0454) .** It has been found satisfactory in terms of scope, quality and presentation as partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** in VIT University, India.

Guide

(Name & Signature)

Internal Examiner

(Name & Signature)

External Examiner

(Name & Signature)

ACKNOWLEDGEMENT

We would like to express my special thanks of gratitude to my Project Guide **Prof. Meenakshi S.P**, head of our department **Prof. Senthilkumar R** as well as our Dean **Prof. Arunkumar T** who gave me the golden opportunity to do this wonderful project on the topic **"Prediction of heart disease using machine Learning"**, which also helped me in doing a lot of Research and we came to know about so many new things. We are really thankful to them.
Secondly we would also like to thank our University "**VIT University, Vellore**" which gave us such a magnificient platform to show our skills and also we get to know many things here.

CONTENTS

TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations	**Expansion**
PAC	Probabilistic Analysis and Classification
HDPS	Heart Disease Prediction System
DST	Decision Tree
IHDPS	Intelligent Heart Disease Prediction System
MAFIA	Maximal Frequent Item set Algorithm
IHD	Ischemic Heart Disease
BPNN	Back-propagation neural network
BNN	Bayesian neural network
PNN	Probabilistic neural network
SVM	Support vector machine
CVD	Cardiovascular disease

Abstract

Today's world heart disease is the top cause to the numbers of death happening all around the globe. determining the cardiovascular disease is too tough as it is highly complex and doctors need to have great expertise in that field. In today world clinics and hospital has a lot of data and information that is hidden. But then also to find from what a person is suffering it takes too much time. It is so much that mostly it takes the life of a person or he/she is in the last stage of his/her life. One of the most knowing loop hole is that doctors are able to save life of very less persons. But the day by day patients are rising and specialist are less and number of deaths has been increasing rapidly. By analyzing these things and to make a solution so that even a nurse can help in prediction the heart disease by just entering some required inputs and then we can take care of disease by consulting with specialists. Various machine learning algorithms and techniques can be used to predict the cardiovascular disease in an individual. Some which we will be using is k-nearest neighbor, naïve bayes and Decision Tree. The research states that the accuracy level of prediction of heart disease using machine learning algorithms is about 96 percent. Data mining algorithms can be used to generate the pattern among the hidden data present in the health care industry. By gathering data's from various health care and applying inputs for various inputs in this algorithm we can get the results in which we get maximum results true. We can further implement some more algorithm which can give more accuracy than these so that it can be circulated to everywhere in this world so with heart disease no one has to die.

1. INTRODUCTION

1.1 Theoretical Background

Categorization techniques of Data Mining and Machine Learning Algorithms play important part in predicting as well as data mining. Also, symptoms of cardio disease may not be important and thus are many times ignored. Big Data like records of patients is handled using Hadoop Map Reduce technique. Different types of study of the machine learning algorithms is done and graphical representation of the results is provided for easier understanding. This application i.e HDPS is made globally available and accessed by deploying it on Cloud Platform and can be available through any browser in any part of the world.

1.2 Motivation

In today's world there are many scientific technologies which helps doctors in taking important decisions but they might not be accurate. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. Predicting of the heart disease is a very complex task in this world. Prediction of heart disease is a major challenge faced by hospitals and medical centers, especially when it comes to accuracy.

The hospitals collects huge amounts of data which is not possible to handle manually. With the huge growing population, the doctors and experts available are not in proportion with the population.

1.3 Aim of the proposed work

Today in this world there are many people who dies because of heart attacks and some of them because they find it too late to be cured. Aim of this work is to predict the heart disease patient if any so that there will be ample time to cure it or to find the cure.

1.4 Objective of the proposed work

The main objective of this project is predicting the cardio vascular disease risk level of a patient using machine learning algorithms, Introduction of Probabilistic Analysis and Classification (PAC), Comparison of Probabilistic Analysis and Classification (PAC) with existing Machine Language Algorithms and creating a specific System for both doctors and patients to login and view the data on Cloud platform.

1.5 Report Organization

Here, to gather data so that it can be utilised for our testing against the various inputs we took it from Cleveland health care centre in United States.

2. LITERATURE SURVEY

Cardiovascular disease is a term that specifies to a large number of health problems related to heart. These improper conditions describe the illness that directly impact the heart and all its parts. Cardio disease is a major health problem in today's time.

2.1 Existing Work

A model Intelligent Heart Disease Prediction System worked with the help of information mining procedures specifically, Neural Network, Naïve Bayes, and Decision Tree. Comes about demonstrate that every system has its rare quality in understanding the destinations of the characterized mining objectives. IHDPS can answer complex "imagine a scenario in which" questions which customary choice emotionally supportive networks can't be proposed by Sellappan Palaniappan .. The outcomes delineated the awkward quality of each of the techniques in appreciating the objective of the predetermined mining goals.

IHDPS was fit for reacting questions that the customary choice emotionally supportive networks were not ready to. It encouraged the establishment of significant information, for example, designs, connections in the midst of therapeutic components associated with coronary illness. IHDPS stays prosperity electronic, easy to use, dependable, adaptable and expandable. The conclusion of Heart Disease, Blood Pressure and diabetes with the guide of neural systems was presented by Niti Guru et al. . Trials were done on an examined informational collection of patient's records. The Neural Network is prepared and tried with 13 input factors, for example, Blood Pressure, Age, Angiography's report and so forth. The managed arrange has been educated for analysis with respect to heart infections. Preparing was done with the assistance of back engendering calculation. At whatever point new information was embedded by the specialist, the framework recognized the obscure information from correlations with the prepared information and delivered a list of plausible ailments that the patient is defenseless against.

Maximal Frequent Item set Algorithm (MAFIA) is connected for mining maximal regular model in coronary illness database. The standard examples can be characterized into various classes utilizing the calculation as preparing calculation utilizing the idea of data entropy. The outcome exhibits that the outlined forecast framework is fit for foreseeing the heart assault effectively.

In year 2010, a review was directed for prescient model for the Ischemic Heart Disease (IHD); they connected Back-spread neural system (BPNN), the Bayesian neural system (BNN), the probabilistic neural system (PNN) and the bolster vector machine (SVM) to create order models for distinguishing IHD patients on an information gotten from estimations of cardiovascular attractive field at 36 areas (6 × 6 grids) over the torso6. The outcome demonstrates that BPNN and BNN gave the most astounding characterization exactness of 78.43 %, while RBF bit SVM gave the least grouping precision of 60.78 %. BNN exhibited the best affectability of 96.55 % and RBF piece SVM showed the most reduced affectability of 41.38 %. Both polynomial part SVM and RBF bit SVM exhibited the base and most extreme specificity of 45.45 % and 86.36 %, separately.

In 2012, T.John Peter and K. Somasundaram Professor, Department of CSE introduced a paper, "An Empirical Study on Prediction of Heart Disease utilizing arrangement information mining strategy". In this examination paper, the utilization of example acknowledgment and information digging systems are utilized for expectation of hazard in the therapeutic area of coronary illness pharmaceutical is proposed here.

In 2014, M.A.Nishara BanuB.Gomathy Professor, Department of Computer Science and Engineering has distributed an examination paper "Infection Forecasting System Using Data Mining Methods".In this article, the pre-prepared information is grouped utilizing bunching calculations as K-intends to assemble applicable information in a database.

Another review probed a specimen database of patients' records. The Neural Network is tried and prepared with 13 input factors, for example, Age, Blood Pressure, Angiography's report and so forth. The administered arrange has been suggested for analysis of heart diseases4. Preparing was done with the guide of back engendering calculation. At whatever point obscure information was encouraged by the specialist, the framework recognized the obscure information from correlations with the prepared information and produced a rundown of likely illnesses that the patient is defenseless against. The achievement rate for uncertain contributions to recover the coveted yield is nearest to 100%.

In another review the issue of distinguishing compelled affiliation rules for coronary illness forecast was considered by5. The basic dataset incorporated restorative records of individuals having coronary illness with qualities for hazard elements, heart perfusion estimations and supply route narrowing. Three limitations were acquainted with lessening the quantity of examples. Initial one requires the credits to show up on just a single side of the run the show. The second one isolates qualities into uninteresting gatherings. A definitive imperative confines the quantity of qualities in a rule9. Tests delineated that the requirements diminished the quantity of found guidelines strikingly close to diminishing the running time. Two gatherings of guidelines conceived the nearness or nonattendance of coronary illness in four particular heart conduits.

2.2 Gaps In Survey

A portion of the constraints of the customary medicinal scoring frameworks are that there is a nearness of inborn direct blends of factors in the information set, and henceforth they are not gifted at demonstrating nonlinear complex associations in restorative spaces. This confinement is dealt with in this examination by utilization of characterization models which can verifiably distinguish complex nonlinear connections amongst autonomous and ward factors and in addition the capacity to recognize every single conceivable association between indicator factors.

In 2013, Shamsher Bahadur Patel, Pramod Kumar Yadav, and Dr. D. P.Shukla introduced an examination paper,"Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques ".In this exploration paper, the human services industry, the information digging is essentially used for the forecast of coronary illness. The target of our attempts to anticipate the analysis of coronary illness with a decreased number of qualities utilizing Naïve Bayes, Decision Tree.

3. OVERVIEW OF THE PROPOSED SYSTEM

3.1 Introduction

Categorization techniques of Data Mining and Machine Learning Algorithms play important part in predicting as well as data mining. Also, symptoms of cardio disease may not be important and thus are many times ignored. Big Data like records of patients is handled using Hadoop Map Reduce technique. Different types of study of the machine learning algorithms is done and graphical representation of the results is provided for easier understanding.

3.1.1 Related Concepts:

k-Nearest Neighbor algorithm

The k-closest neighbors calculation (k-NN) is a non-parametric technique utilized for order and revert. In both cases, the data comprises of the k nearest preparing cases in the component space . The output relies on upon whether k-NN is utilized for grouping or revert, k-NN is a sort of example based learning, where the capacity is just approximated locally and all calculation is conceded until arrangement. The k-NN calculation is among the easiest of all machine learning calculations. Both for categorization and revert, it can be important enough to allocate weight to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones. For instance, a typical weighting plan comprises in giving each neighbor a weight of 1/d, where d is the separation to the neighbor.

The neighbors are taken from an arrangement of items for which the class (for k-NN order) or the protest property estimation (for k-NN relapse) is known. This can be considered as the preparation set for the calculation, however no unequivocal preparing step is required.

Naïve Bayesian Classifier-

In information mining we utilize credulous Bayesian arrangement, in which we take different sorts of informational collection and as indicated by that suspicion we get the yield required most. It is a basic probabilistic classifier. This classification requires some extreme and adaptable direct parameters in type of variable. The Bayesian Classification delineates a directed learning strategy and a measurable technique for order .In this we can utilize most extreme probability preparing technique which should be possible in assessing a shut shape expression which takes straight time as opposed to iterative guess.

Bayes rule:

It is a contingent likelihood which expresses that there is a probablity of some yield when given some information where a reliance relationship exists amongst perception and conclusion.

Give yield a chance to be "c" and info be "e" then likelihood is indicated by **p(c/e) = (p(e/c)p(c))/p(e)**

In basic terms:- probability= (prior*likelihood)/occasion

Decision Tree-

A tree which utilizes a tree like chart structure which implies it incorporates a root hub, branches and a leaf hub. Each inner hubs indicates a test on a characteristic, each branch signifies the result of a test, and each leaf hub holds a class name. By and large DST are utilized as a part of settling on expository choices which can contact us to our objective in most productive way additionally a compelling device for machine learning . The ways from root to leaf speaks to order runs the show. Tree models where the objective variable can take a limited arrangement of qualities are called characterization trees; in these tree structures, leaves speak to class marks and branches speak to conjunctions of components that prompt those class names. DST where the objective variable can take ceaseless qualities are called relapse trees. It takes after voracious calculation i.e it picks the productive way which can be seen at first and furthermore it takes after recursive and partition and overcome strategy. This calculation begins with entire arrangements of preparing informational indexes among which DST chooses the one which has most informations for grouping and creates a test hub. It partitions the present arrangement of tuples as per their estimations of the present test quality. Classifier era won't work if all the preparation informational indexes is from a similar class or on the off chance that it is not worth to continue promote with an extra detachment when we see that after further characterization likewise it produces order just with pre-grouping limit. For grouping of the preparation informational indexes into various classes we utilizes a particular measure called 'data pick up'. DST calculation processes data pick up for each characteristic and each round, the one with the most data pick up is chosen for the testing reason.

Information Gain:

Most perplexing thing in DST is picking the best credit and to that we utilize data pick up. To do this we have to get another component called entropy. Entropy is the measure of irregularity. Give "s" a chance to be set of comprising set of information tests. What's more, assume class tests has "m" diverse classifications, Ck. Give "si" a chance to be the quantity of tests of "S" in 'Ck'. At that point the data expected to order can be: I(s1,s2… ., sm)= - entirety of(Pk log2(Pk)) ; here , Pk :- likelihood that a subjective example has a place with class Ck and is evaluated by sk/s Now, let quality A have v diverse qualities, {a1,a2,… ,av}.Attribute A can be utilized to parcel S into v subsets, {S1,S2,… ,Sv}, where Sj contains those specimens in S that have esteem aj of A. Let s kj be the quantity of tests of class Ck in a subset Sj. The entropy, or expected data in view of the parceling into subsets by An, is given by:

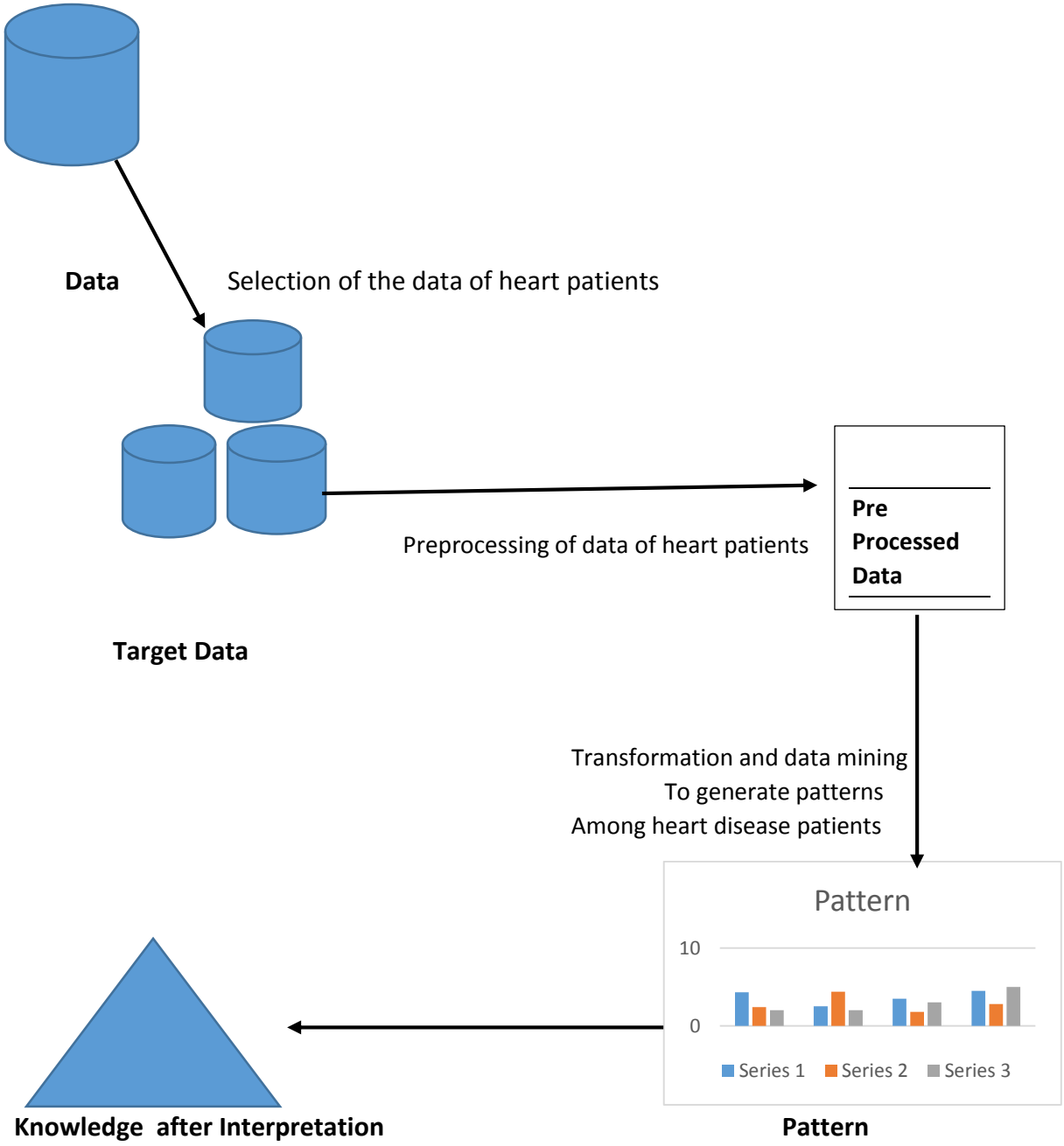
E(A)=(sum of(S1j+…+Smj)/s)I(S1j,…,Smj)

Finally ,**Information gain(A)=I(s1,s2,…,sm) – E(A)**

3.2 System Architecture

Crude information is available in the information stockroom. The information is about the cardio patients having inputs like age, circulatory strain, sex, cholesterol and so on. These information is chosen out from the information distribution center to exhibit the objective information. Presently the preprocessing of the objective information happens to produce the preprocessed information.

Presently on the preprocessed information of the cardio persistent change by utilizing different calculation and information mining happens to create design among cardio ailment quiet. After this information is translated in the wake of investigating the example. At the point when the data is assembled, we can now anticipate a man's opportunity to have cardio malady in view of the estimation of information sources he postures.



After this prediction of a person having heart disease

Is done

3.3 PROPOSED SYSTEM MODEL

3.3.1 Knn model:

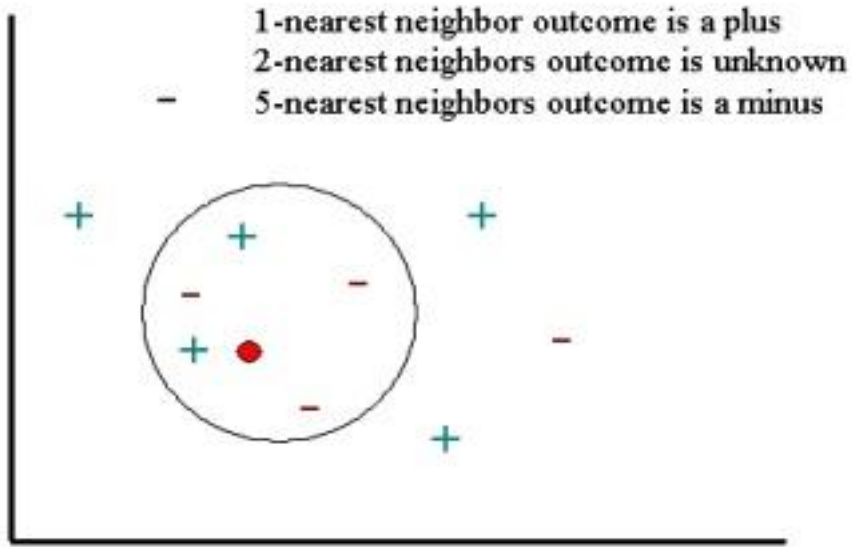
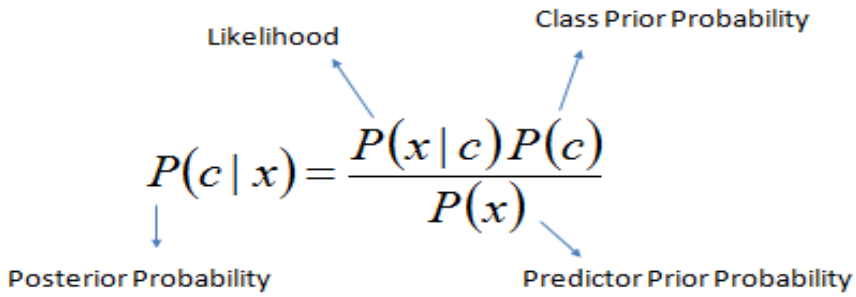


Fig.1

3.3.2 Bayes model:



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

3.3.3Decision Tree:

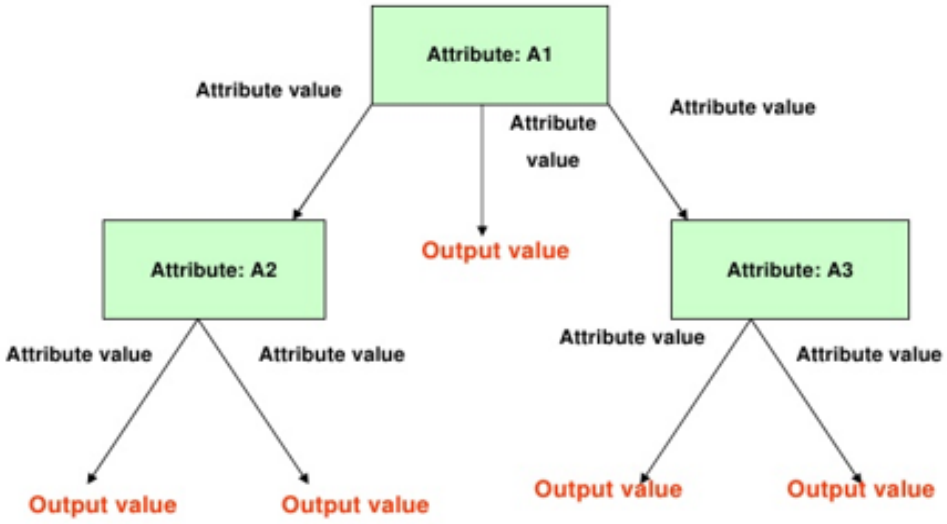


Fig.2

4. PROPOSED SYSTEM DESIGN AND ANALYSIS

4.1 Introduction

The fundamental target of this examination is to build up a Health Care Prediction System utilizing, K-closest neighbors, Naive Bayes and Decision Tree .The can find and concentrate concealed data related with illness (cardiovascular) from a coronary illness information distribution center. It can answer extreme inquiries for diagnosing infection and consequently help specialists to settle on keen clinical choices which conventional choice emotionally supportive networks can't. By giving compelling medicines, it likewise decreases treatment costs.

4.2 Requirement Analysis

For the development of various suitable techniques for the customer ease there are two types of requirements functional and non-functional requirements:

4.2.1 FUNCTIONAL REQUIREMENTS

4.2.1.1 Product Perspective:

Nowadays, heart attack has been common and due to lack of time many people die due to this. So to avoid these many deaths this solution has been implemented so that people will be able to know at their first stage and can get proper treatment.

4.2.1.2 Product Features:

This product takes into account various inputs like age, blood pressure of the person, chest pain, blood sugar, sex and some more and after evaluating all these inputs it calculates the probability of getting heart attack or some other heart disease.

4.2.1.3 User Characterstics:

Charactestic	Description
Sex	Male or female
Chest Pain Type	value 1: typical type 1 angina, value 2: typical type angina, value 3: non- angina pain; value 4: asymptomatic
Fasting Blood Sugar	value 1: > 120 mg/dl; value 0:< 120 mg/dl
Age	In year
Height	In cms
Weight	In kg
Serum Cholesterol	In mg/dl
Thalach	maximum heart rate achieved
Oldpeak	ST depression induced by exercise relative to rest
Blood Pressure	In mm hg
Electrocardiograph	value 0:normal, value 1:having st t wave abnormality, value 2: showing definite left ventricular
Induced Angina	value 0: no, value 1: yes

4.2.1.4 Assumption and Dependencies:

The information that we are getting is from UCI Cleveland Machine Learning Repository and suspicion is done that information is gathered precisely. Since we are evacuating the information of the patient which is copied and where a few qualities are absent as opposed to putting some irregular qualities it will give the exactness of the expectation. Since there are a considerable measure of traits that together verify that the given individual is experiencing coronary illness or not, we accept that every one of the information are autonomous from each other. Yet, the yield information is result which is gotten from the properties information. Consequently we can state that the given yield is reliant upon the properties which are not related upon each other.

4.2.1.5 Domain Requirements:

The yield gather comprises of all dynamic or proficient makers, designers, clients and clients of an application. The distinctive performing artists will have diverse desires and will see the value of the coronary illness expectation application from alternate points of view. Outlining restorative gadgets requires a top to bottom learning and data of the wellbeing market, end-client necessities, security, and administrative consistence.

The venture enables engineers to create financially savvy, superior encompassing insight applications for heterogenious physical gadgets and designer clients are the fundamental target gathering of clients. Engineers of wellbeing gadgets and items need to watch a large number of necessities for item execution, security, convenience, unwavering quality, and cost. At long last, understanding the necessities of end clients and coordinating those requirements into advancement ventures lies at the heart of creating successful therapeutic application. Measuring and satisfying client prerequisites amid restorative application advancement will bring about fruitful items that enhance tolerant security, enhance gadget viability and lessen item reviews and alterations. From an end-client viewpoint, originators and engineers of frameworks and applications must face this test, which will require tending to genuine medicinal services and homecare needs.

- **Functionality:** The application depends on information mining ideas. Information mining is absolutely needy upon the exactness of the information and the measure of the information. As the volume of the information builds the effectiveness and exactness of the application will increment.

- **Design:** The outline of the application ought to be such a route, to the point that the calculation expected to build up this application ought to give exact outcomes. On the off chance that we are utilizing innocent straights calculation the precision is around ninety six percent however in the event that we are utilizing the choice tree calculation the exactness rate is more than ninety nine percent. Along these lines, picking of the calculation ought to be such an approach to give most precise outcomes.

4.2.1.6 User Requirement:

We can see from the above application that there are stakeholders which we can keep in account while developing this application

- **Hospitals:** The vital partner is the healing facility representatives that will give the informational index that we have to prepare to anticipate the coronary illness in a patient. Along these lines, they ought to guarantee that the information conveyed in steady and blunder free.

- **Patient:** Patient is another real partner in that we can experience while building up this application. Quiet needs to give all the property subtle elements of his in the event that he needs to see if he is experiencing coronary illness or not.

- **Developer:** The individual who is building up this application is the third most essential partner. He needs to ensure that the information that he is preparing in free from irregularity like duplication of information and missing qualities. In the event that he is certain about the consistency of the information then he can do preparing of the information by applying legitimate machine learning calculation like credulous straights and choice tree that would deliver exact outcomes for the patients.

4.2.2 NON-FUNCTIONAL RQUIREMENTS

4.2.2.1 Product Requirements:

For predicting the heart disease it requires some attributes like-chest pain type, fasting blood sugar, age, height, weight, serum cholesterol, sex, maximum heart rate achieved and depression induced by relative to rest.

4.2.2.1.1 Efficiency:

It takes less time to execute because we are filtering the data first the training the data which will lead to correctness and accuracy.

4.2.2.1.2 Reliability:

For the naïve bays the accuracy percentage is about ninety six percent and for the decision tree its accuracy is more than ninety nine percent therefore its quite reliable

4.2.2.1.3 Portability:

Since the device we are developing is web based application. So this application can run across various platforms. The code that we write is in Python, so any device having python within their system will be able to run this application.

4.2.2.1.4 Usability:

It is easy to use as you know that it just needs inputs and you will get your result. It can be executed in any browser.

4.2.3 ORGANIZATIONAL REQUIREMENTS

Implementation Requirements:

The application will provide tools for overcoming the deployment obstacles inherent in previous disease management programs incorporating remote monitoring by providing easy to use tools that allow developers to design and implement advanced solutions based on existing or new home medical devices. For the implementation we need to use data analysis tool using python called pandas after that plotting the pattern using the matplotlib and after that training the data and deriving the result for the patients data using sklearn.

Engineering Standard Requirements:

The new innovations set the prerequisites for restorative office outline and development and also the prequalification of doctor's facility plan advisors. The norms go for enhancing therapeutic

results and limiting blunders brought about by unseemly human services office outline, and is a stage towards the advancement for the medicinal services industry. The new advancements are a thorough instrument that all future Healthcare offices and as of now repaired ones need to conform to be authorized.

- Administrative Provisions; separates the permitting procedure for offices given in the wellbeing focus and the prequalification procedure for plan advisors.

- Health Facility Briefing and Planning; incorporates compositional and human services office arranging Standards.

4.2.3.1 Economic:

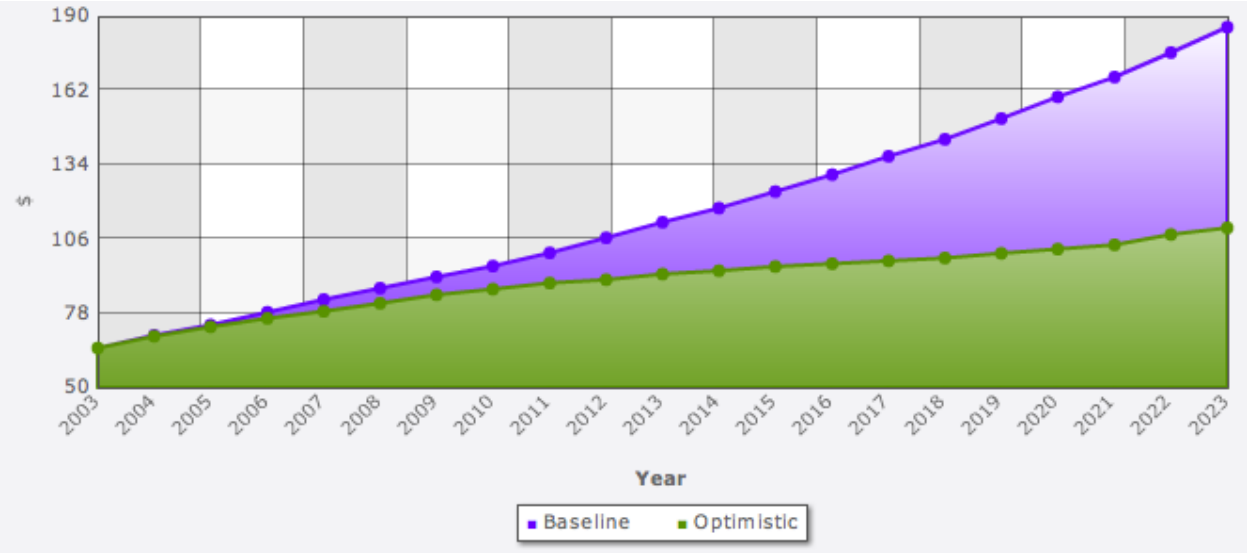


Fig.3

The purple line speaks to the deliberate or anticipated treatment uses in light of the outcome identified with coronary illness, while the green line speaks to the evaluated consumptions that would happen if little alterations were made to enhance coronary illness counteractive action.

Representatives experiencing coronary illness require extra days off and can be less gainful in the work environment. Also, the sudden passing brought on by coronary illness incurs significant injury on the yield of the general workforce. In 2010, an expected $41.7 billion in potential efficiency was lost because of Cardiovascular sickness (CVD)- related worker dismalness, and $137.4 billion was lost because of CVD-related unexpected losses (George and Hong, 2011).

4.2.3.2Environmental:

On the off chance that our proposed arrangement is actualized the majority of the natural effects like ozone layer consumption, impacts of cfc's (chlorofluoro carbons) will diminish as the fundamental hazard variables of CVD abatement. In the event that individuals strolled somewhat more set up of driving everybody's carbon impression would diminish a tiny bit and the ozone would be less in risk. Facilitate, if individuals devoured less fast food, in a perfect world essentially less, there would be a noteworthy reduction in the measure of litter in our nation. In the event that individuals are not expending bundled substances there is nothing to litter, and waste items would likewise consume up less room in landfills. At last, if the quantity of individuals who smoke diminishes so will the quantity of stogies littered and discarded that unfavorably influence the earth, plants and creatures, and our conduits. These advantages will be certain reactions if our proposed arrangement is fruitful at anticipating Cardio vascular disease.

4.2.3.3Social:

The effects of Cardio Vascular Disease are not restricted to provide aid, but rather can diminish social parts of life also. Since cardio disease is an endless disease and not something which can be cured so easily. A site made for coronary illness self improvement cautions patients that they should conform to an existence with no strenuous exercise and that they should do without drinking mixed refreshments (Improve Heart Health, 2009). Contingent upon the kind of way of life a patient had beforehand been utilized to, these confinements may speak to critical changes throughout their life. It is no big surprise that with physical impediments, social constraints, and temperamental wellbeing, numerous CVD patients end up plainly discouraged. While rates of melancholy inside the US are evaluated to be under 10% (Centers for Disease Control, 2011), rates of despondency among CVD patients are as high as 27% (European College of Neuropsychopharmacology, 2007). Concentrates found that the onset of major clinical wretchedness is basic after intense coronary disorder, a common kind of coronary illness (European College of Neuropsychopharmacology, 2007). These reviews demonstrate that coronary illness largy affects psychological wellness alongside general physical wellbeing.

4.2.3.4 Political:

There are many doctors who plays politics in their hospital. Though they don't struggle with doctors and they do what they are told to do and by this way doctor earns more money but if we have this solution then we can do checkup for ourselves and if it is negative then we can go to doctors to attain proper care. By this way our cost also gets reduced.

4.2.3.5Ethical:

By implementing this technology it will have an great impact in our life as well as in ethics because when a person knows from what he or she is suffering that person follows proper rule so that he or she can remove that problem on what he or she is suffering.

4.2.3.6 Health and Safety:

If we will be able to know about our heart disease before any attacks then there are many chances that we will be able to avoid the death or any other major accident that would have been possible if this solution wouldn't have been there. It can be a great boon for our world which can save many lives.

4.2.3.7 Sustainability:

This has been developed by searching too many data throughout the world therefore its accuracy of predicting heart disease is at an approximate level of 96% and it remains it consistency.

4.2.3.8 Legality:

Since there is not any false requirements therefore after proper testing of the software it can be approved by the government of their respective country and can be legalized.

4.2.3.9 Inspectability:

Proper testing and inspection of the software should be made before revealing it in the market**.** Accuracy is one of the most important thing to be considered while testing.

4.2.4 SYSTEM REQUIREMENTS

4.2.4.1 Hardware Requirements:

Since we are using Anaconda as our IDE for this project, hence the minimum hardware requirement will be:

- Processor needed: 1.6 GHz
- RAM: 4GB
- Space required: 2GB

4.2.4.2 Software Requirements:

• **Anaconda**: ("Anaconda Distribution") is a free, simple to-introduce bundle supervisor, condition administrator, Python dissemination, and gathering of more than 720 open source bundles with free group bolster. Hundreds more open source bundles and their conditions can be introduced with a basic "conda introduce [packagename]". It is stage skeptic, can be utilized on Windows, OS X and Linux. Or, then again significantly less demanding, with new Anaconda Navigator for point and snap introduce of situations and bundles.

• **Pandas:** Pandas is a product library composed for the Python programming dialect for information control and investigation. Specifically, it offers information structures and operations for controlling numerical tables and time arrangement. Pandas is free programming discharged under the three-statement BSD permit. The name is gotten from the expression "Board information", an econometrics term for multidimensional organized informational collections.

• **Matplotlib:** Matplotlib is a plotting library for the Python programming dialect and its numerical arithmetic expansion NumPy. It gives a question arranged API to installing plots into applications utilizing universally useful GUI toolboxs like wxPython, Qt, or GTK+. There is additionally a procedural "pylab" interface in view of a state machine (like OpenGL), intended to nearly look like that of MATLAB, however its utilization is debilitated.

• **Scikit-Learn:** Scikit-learn (once in the past scikits. learn) is a free programming machine learning library for the Python programming dialect. It highlights different order, relapse and bunching calculations including bolster vector machines, arbitrary backwoods, slope boosting, k-means and DBSCAN, and is intended to interoperate with the Python numerical and logical libraries NumPy and SciPy.

5. RESULTS AND DISCUSSIONS

5.1 Sample test cases

We have gathered raw data from a clinic of cleaveland and arranged all the data and mined it properly according to its various inputs like age, sex, thalach, old peak etc. We took a record of 300 more records from cleaveland health clinic. In this we took 250 records as training data and the rest for testing data.

(Refer Annexure-1 for the data)

5.2 Summary

From these outcomes and discourse it is presumed that despite the fact that there are many machine learning strategies like, neural system, SVM, KNN and parallel discretization with pick up proportion choice tree applying in coronary illness forecast has bring down exactness rate than when we apply gullible bayes and choice tree method.by this we can likewise say that choice tree strategy beat innocent bayes calculation over expectation.

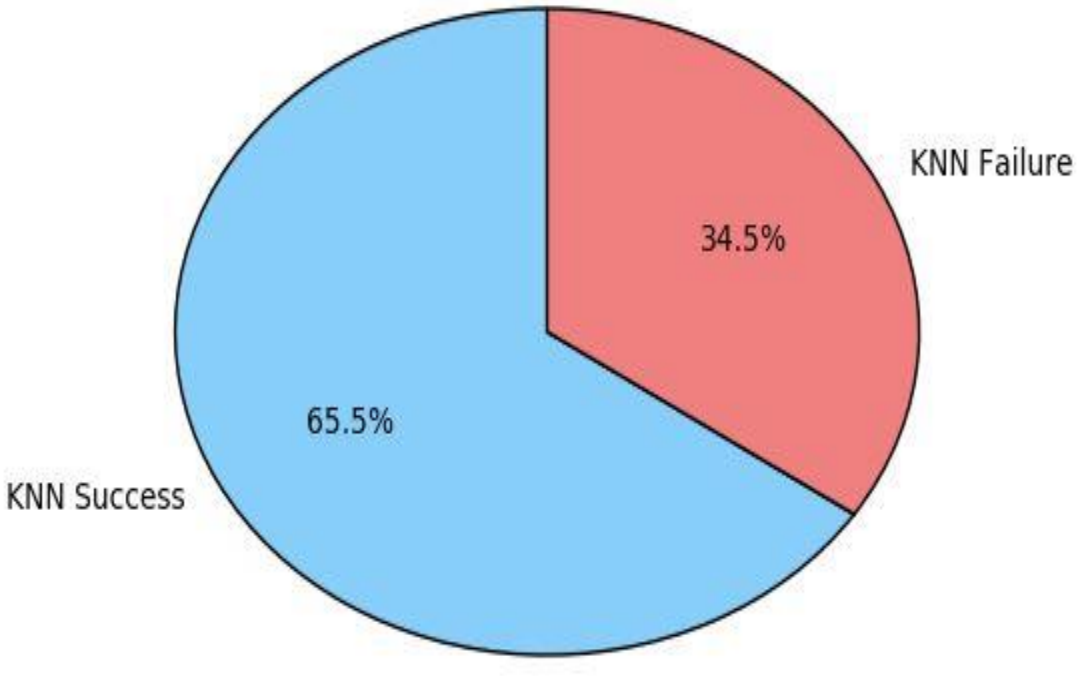


Fig.4

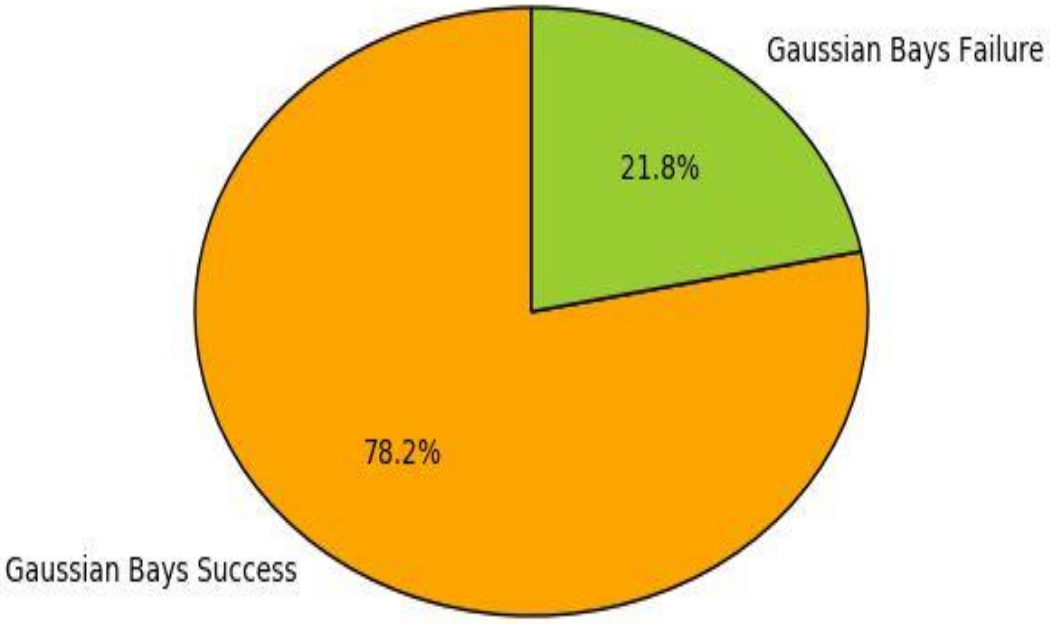

Gaussian Bays Failure
21.8%
78.2%
Gaussian Bays Success

Fig.5

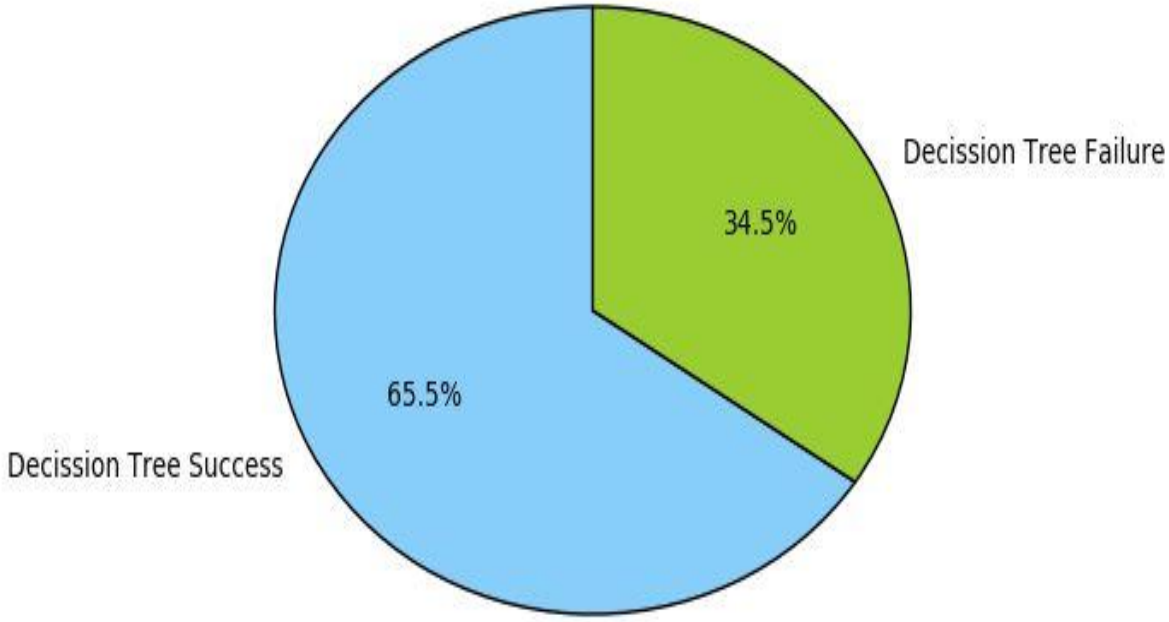

Decission Tree Failure
34.5%
65.5%
Decission Tree Success

Fig.6

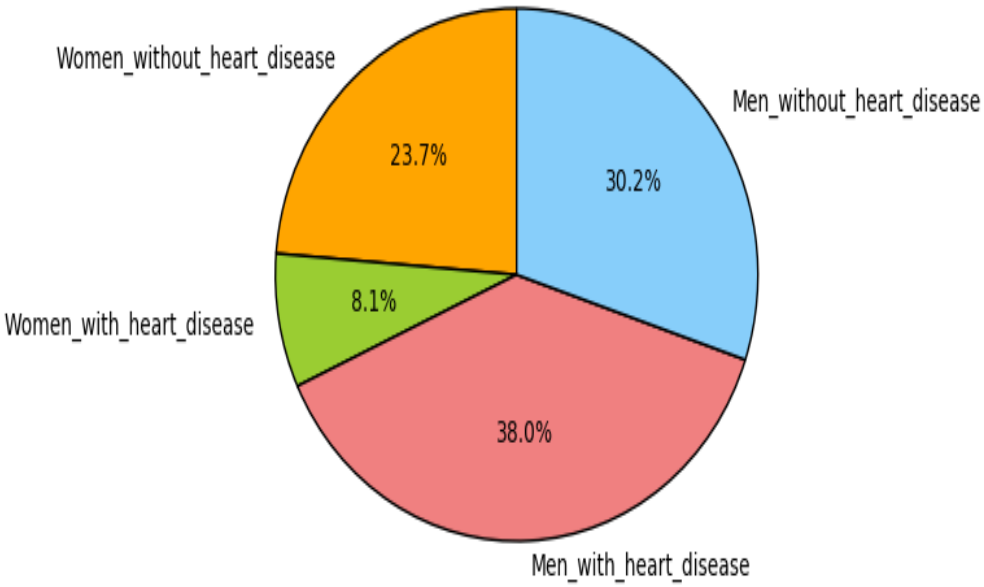

Women_without_heart_disease
23.7%
Men_without_heart_disease
30.2%
Women_with_heart_disease
8.1%
38.0%
Men_with_heart_disease

Fig.7

6. CONCLUSION AND FUTURE WORK

The goal of our work is to give an investigation of various information mining strategies that can be utilized in computerized coronary illness expectation frameworks. Different strategies and information mining classifiers are characterized in this work which has developed as of late for productive and powerful coronary illness conclusion. The examination demonstrates that distinctive advances are utilized as a part of the considerable number of papers with taking diverse number of properties. In this way, unique advancements utilized demonstrated the distinctive exactness to each other. In a few papers it is demonstrated that that Decision Tree has performed well with 96.2% precision by utilizing 13 properties. Thus, extraordinary advancements utilized and demonstrated the diverse precision relies on number of properties taken and instrument utilized for usage. Propelled by the overall expanding mortality of coronary illness patients every year and the accessibility of gigantic measures of information, scientists are utilizing information mining procedures in the finding of coronary illness. In spite of the fact that applying knowledge depth strategies to provide human services professionals in the finding of cardio diseases is having some achievement, the utilization of information mining systems to distinguish an appropriate treatment for coronary illness patients has gotten less consideration.

We can include a few components so that it's precision rate can be expanded and it's extent of region can likewise be broadened. These can be extremely useful in forthcoming days of our future as it will lessen the work of specialists patients can likewise be recuperated speedier. Be that as it may, there is likewise a few slip-ups if which is done it can cost incredibly.

APPENDIX

```
In [23]: import pandas as pd
```

```
In [24]: data=pd.read_csv(r'C:\Users\Harsh Vardhan\Desktop\Heart-Disease-Prediction-using-Machine-Leaning-master\cleveland_data.
         csv')
```

```
In [25]: data.head()
```

Out[25]:

	63.0	1.0	1.0.1	145.0	233.0	1.0.2	2.0	150.0	0.0	2.3	3.0	0.0.1	6.0	0
0	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2.0
1	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1.0
2	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0.0
3	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0.0
4	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0.0

```
In [26]: data.tail()
```

Out[26]:

	63.0	1.0	1.0.1	145.0	233.0	1.0.2	2.0	150.0	0.0	2.3	3.0	0.0.1	6.0	0
302	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0	0.0	7.0	1.0
303	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0	2.0	7.0	2.0
304	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0	1.0	7.0	3.0
305	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0	1.0	3.0	1.0
306	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0	0.0	3.0	0.0

Fig.8

```
In [27]: column_name=['Age','Sex','Chest_Pain','Blood_Pressure','Cholestoral','Blood_Sugar','Electrocardiographic','Heart_Rate',
         'Induced_Angina','ST_Depression','Slope_ST_Segment','Vessels_Colored','Thal','Diagnosis']
```

```
In [28]: data=pd.read_csv(r'C:\Users\Harsh Vardhan\Desktop\Heart-Disease-Prediction-using-Machine-Leaning-master\cleveland_data.
         csv',names=column_name)
```

```
In [29]: data.head()
```

Out[29]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope_S
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0

```
In [30]: data.tail()
```

Out[30]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
303	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0
304	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0
305	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0
306	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0
307	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0

Fig.9

```
In [31]: data.shape
Out[31]: (308, 14)

In [32]: data.dtypes
Out[32]: Age                    float64
         Sex                    float64
         Chest_Pain             float64
         Blood_Pressure         float64
         Cholestoral            float64
         Blood_Sugar            float64
         Electrocardiographic   float64
         Heart_Rate             float64
         Induced_Angina         float64
         ST_Depression          float64
         Slope_ST_Segment       float64
         Vessels_Colored        float64
         Thal                   float64
         Diagnosis              float64
         dtype: object

In [33]: type(data)
Out[33]: pandas.core.frame.DataFrame

In [34]: data.isnull().sum()
```

Fig.10

```
Out[34]: Age                    1
         Sex                    0
         Chest_Pain             2
         Blood_Pressure         1
         Cholestoral            1
         Blood_Sugar            0
         Electrocardiographic   0
         Heart_Rate             0
         Induced_Angina         0
         ST_Depression          2
         Slope_ST_Segment       0
         Vessels_Colored        0
         Thal                   0
         Diagnosis              3
         dtype: int64

In [35]: data[data.Age.isnull()]
```

Out[35]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
228	NaN	1.0	4.0	112.0	204.0	0.0	0.0	143.0	0.0	0.1	1.0

```
In [36]: data[data.Chest_Pain.isnull()]
```

Out[36]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
26	60.0	1.0	NaN	130.0	206.0	0.0	2.0	132.0	1.0	2.4	2.0
174	59.0	0.0	NaN	174.0	249.0	0.0	0.0	143.0	1.0	0.0	2.0

Fig.11

In [37]:
```
data[data.Blood_Pressure.isnull()]
```

Out[37]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
232	52.0	0.0	3.0	NaN	196.0	0.0	2.0	169.0	0.0	0.1	2.0

In [38]:
```
data[data.Cholestoral.isnull()]
```

Out[38]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope_
19	54.0	1.0	4.0	140.0	NaN	0.0	0.0	160.0	0.0	1.2	1.0

In [39]:
```
data[data.ST_Depression.isnull()]
```

Out[39]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope_
24	58.0	1.0	2.0	120.0	284.0	0.0	2.0	160.0	0.0	NaN	2.0
28	58.0	0.0	3.0	120.0	340.0	0.0	0.0	172.0	0.0	NaN	1.0

In [40]:
```
data[data.Diagnosis.isnull()]
```

Out[40]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope_
6	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0
11	53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0
16	52.0	1.0	3.0	172.0	199.0	1.0	0.0	162.0	0.0	0.5	1.0

Fig.12

```
In [41]: data.dropna(how="any",inplace=True)

In [42]: data.shape

Out[42]: (298, 14)

In [43]: data.isnull().sum()

Out[43]: Age                    0
         Sex                    0
         Chest_Pain             0
         Blood_Pressure         0
         Cholestoral            0
         Blood_Sugar            0
         Electrocardiographic   0
         Heart_Rate             0
         Induced_Angina         0
         ST_Depression          0
         Slope_ST_Segment       0
         Vessels_Colored        0
         Thal                   0
         Diagnosis              0
         dtype: int64

In [44]: data.duplicated().sum()

Out[44]: 3

In [45]: data[data.duplicated()]
```

Fig.13

```
In [45]: data[data.duplicated()]
```

Out[45]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
295	44.0	1.0	4.0	120.0	169.0	0.0	0.0	144.0	1.0	2.8	3.0
296	44.0	1.0	4.0	120.0	169.0	0.0	0.0	144.0	1.0	2.8	3.0
297	44.0	1.0	4.0	120.0	169.0	0.0	0.0	144.0	1.0	2.8	3.0

```
In [46]: data.drop_duplicates(keep='first',inplace=True)

In [47]: data.duplicated().sum()
```

Out[47]: 0

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

Out[48]: (295, 14)

```
In [49]: data[(data.Diagnosis>=1) & (data.Sex==0.0)]
```

Fig.14

```
In [50]: data[(data.Diagnosis>=1) & (data.Sex==0.0)].shape
```

Out[50]: (24, 14)

```
In [51]: data[(data.Diagnosis==0) & (data.Sex==0.0)]
```

Out[51]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0
8	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0
13	56.0	0.0	2.0	140.0	294.0	0.0	2.0	153.0	0.0	1.3	2.0
20	48.0	0.0	3.0	130.0	275.0	0.0	0.0	139.0	0.0	0.2	1.0
23	58.0	0.0	1.0	150.0	283.0	1.0	2.0	162.0	0.0	1.0	1.0
27	50.0	0.0	3.0	120.0	219.0	0.0	0.0	158.0	0.0	1.6	2.0
29	66.0	0.0	1.0	150.0	226.0	0.0	0.0	114.0	0.0	2.6	3.0
32	69.0	0.0	1.0	140.0	239.0	0.0	0.0	151.0	0.0	1.8	1.0
44	71.0	0.0	2.0	160.0	302.0	0.0	0.0	162.0	0.0	0.4	1.0
50	65.0	0.0	3.0	140.0	417.0	1.0	2.0	157.0	0.0	0.8	1.0
52	41.0	0.0	2.0	105.0	198.0	0.0	0.0	168.0	0.0	0.0	1.0

Fig.15

```
In [52]: data[(data.Diagnosis==0) & (data.Sex==0.0)].shape

Out[52]: (70, 14)

In [53]: data[(data.Diagnosis>=1) & (data.Sex==1.0)]
```

Fig.16

```
In [56]: data[(data.Diagnosis==0) & (data.Sex==1.0)].shape

Out[56]: (89, 14)

In [57]: Women_without_heart_disease=data[(data.Diagnosis==0) & (data.Sex==0.0)].shape
         Women_with_heart_disease=data[(data.Diagnosis>=1) & (data.Sex==0.0)].shape
         Men_with_heart_disease=data[(data.Diagnosis>=1) & (data.Sex==1.0)].shape
         Men_without_heart_disease=data[(data.Diagnosis==0) & (data.Sex==1.0)].shape

In [58]: name=["Women_without_heart_disease","Women_with_heart_disease","Men_with_heart_disease","Men_without_heart_disease"]
         count=[Women_without_heart_disease[0],Women_with_heart_disease[0],Men_with_heart_disease[0],Men_without_heart_disease[0
         ]]
         colors = ['orange', 'yellowgreen', 'lightcoral', 'lightskyblue']

In [59]: import matplotlib.pyplot as plt


         plt.pie(count,labels=name,autopct='%1.1f%%',colors=colors,startangle=90)
         plt.show()
```

Fig.17

```
In [60]: data.describe()
```

Out[60]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_De
count	295.000000	295.000000	295.000000	295.000000	295.000000	295.000000	295.000000	295.000000	295.000000	295.0(
mean	54.423729	0.681356	3.152542	131.518644	246.884746	0.149153	0.996610	149.457627	0.328814	1.0471
std	9.133931	0.466742	0.965626	17.434976	51.873737	0.356844	0.994879	23.062708	0.470580	1.167(
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000(
25%	47.500000	0.000000	3.000000	120.000000	212.000000	0.000000	0.000000	133.000000	0.000000	0.000(
50%	56.000000	1.000000	3.000000	130.000000	242.000000	0.000000	1.000000	152.000000	0.000000	0.800(
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600(
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200(

```
In [61]: import numpy as np

In [62]: new_data=data.iloc[:,:-1]

In [63]: new_data.head()
```

Fig.18

In [62]:
```
new_data=data.iloc[:,:-1]
```

In [63]:
```
new_data.head()
```

Out[63]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope_S
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0

In [64]:
```
new_data.tail()
```

Out[64]:

	Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
303	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0
304	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0
305	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0
306	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0
307	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0

In [65]:
```
heart_patient=[]
```

Fig.19

```
In [65]: heart_patient=[]

In [66]: for i in data.Diagnosis:
             if i==0:
                 heart_patient.append(0)
             if i>0:
                 heart_patient.append(1)

In [67]: len(heart_patient)
Out[67]: 295

In [68]: from sklearn.naive_bayes import GaussianNB

In [69]: import numpy as np

In [70]: X=np.array(new_data)

In [71]: type(X)
Out[71]: numpy.ndarray

In [72]: Y=np.array(heart_patient)

In [73]: type(Y)
Out[73]: numpy.ndarray

In [74]: Y
```

Fig.20

```
Out[74]: array([0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
                0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0,
                0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1,
                1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1,
                1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1,
                1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0,
                1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0,
                0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1,
                0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0,
                1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1,
                0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
                1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0,
                1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0])

In [75]: model=GaussianNB()

In [89]: model=model.fit(X,Y)

In [105]: bayspredicted=model.predict([[56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0]])

In [106]: bayspredicted
Out[106]: array([0])

In [107]: for i in bayspredicted:
              if i==1:
                  print("Person has heart disease")
              else:
                  print("Person doesn't has heart disease")
          Person doesn't has heart disease
```

Fig.21

```
In [108]: from sklearn import tree

In [112]: decision=tree.DecisionTreeClassifier()

In [114]: decision=decision.fit(X,Y)

In [115]: dtreepredicted=decision.predict([[63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0]])

In [117]: dtreepredicted

Out[117]: array([1])

In [120]: for i in dtreepredicted:
              if i==1:
                  print("Person is having heart disease")
              else:
                  print("Person is not having heart disease")

          Person is having heart disease
```

Fig.22

ANNEXURE –I

Training Data set

Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0
37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0
56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0
57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0	0.0	0.4	2.0
44.0	1.0	2.0	120.0	263.0	0.0	0.0	173.0	0.0	0.0	1.0
57.0	1.0	3.0	150.0	168.0	0.0	0.0	174.0	0.0	1.6	1.0
49.0	1.0	2.0	130.0	266.0	0.0	0.0	171.0	0.0	0.6	1.0
64.0	1.0	1.0	110.0	211.0	0.0	2.0	144.0	1.0	1.8	2.0
43.0	1.0	4.0	150.0	247.0	0.0	0.0	171.0	0.0	1.5	1.0
59.0	1.0	4.0	135.0	234.0	0.0	0.0	161.0	0.0	0.5	2.0
44.0	1.0	3.0	130.0	233.0	0.0	0.0	179.0	1.0	0.4	1.0
42.0	1.0	4.0	140.0	226.0	0.0	0.0	178.0	0.0	0.0	1.0
61.0	1.0	3.0	150.0	243.0	1.0	0.0	137.0	1.0	1.0	2.0
40.0	1.0	1.0	140.0	199.0	0.0	0.0	178.0	1.0	1.4	1.0
59.0	1.0	3.0	150.0	212.0	1.0	0.0	157.0	0.0	1.6	1.0
51.0	1.0	3.0	110.0	175.0	0.0	0.0	123.0	0.0	0.6	1.0
53.0	1.0	3.0	130.0	197.0	1.0	2.0	152.0	0.0	1.2	3.0
65.0	1.0	4.0	120.0	177.0	0.0	0.0	140.0	0.0	0.4	1.0
44.0	1.0	2.0	130.0	219.0	0.0	2.0	188.0	0.0	0.0	1.0
54.0	1.0	3.0	125.0	273.0	0.0	2.0	152.0	0.0	0.5	3.0
51.0	1.0	1.0	125.0	213.0	0.0	2.0	125.0	1.0	1.4	1.0

50.0	0.0	2.0	120.0	244.0	0.0	0.0	162.0	0.0	1.1	1.0
50.0	0.0	4.0	110.0	254.0	0.0	2.0	159.0	0.0	0.0	1.0
64.0	0.0	4.0	180.0	325.0	0.0	0.0	154.0	1.0	0.0	1.0
64.0	0.0	3.0	140.0	313.0	0.0	0.0	133.0	0.0	0.2	1.0
37.0	0.0	3.0	120.0	215.0	0.0	0.0	170.0	0.0	0.0	1.0
46.0	0.0	2.0	105.0	204.0	0.0	0.0	172.0	0.0	0.0	1.0
46.0	0.0	4.0	138.0	243.0	0.0	2.0	152.0	1.0	0.0	2.0
64.0	0.0	4.0	130.0	303.0	0.0	0.0	122.0	0.0	2.0	2.0
41.0	0.0	3.0	112.0	268.0	0.0	2.0	172.0	1.0	0.0	1.0
54.0	0.0	3.0	108.0	267.0	0.0	2.0	167.0	0.0	0.0	1.0
39.0	0.0	3.0	94.0	199.0	0.0	0.0	179.0	0.0	0.0	1.0
34.0	0.0	2.0	118.0	210.0	0.0	0.0	192.0	0.0	0.7	1.0
67.0	0.0	3.0	152.0	277.0	0.0	0.0	172.0	0.0	0.0	1.0
74.0	0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0
54.0	0.0	3.0	160.0	201.0	0.0	0.0	163.0	0.0	0.0	1.0
49.0	0.0	2.0	134.0	271.0	0.0	0.0	162.0	0.0	0.0	2.0
41.0	0.0	2.0	126.0	306.0	0.0	0.0	163.0	0.0	0.0	1.0

49.0	0.0	2.0	134.0	271.0	0.0	0.0	162.0	0.0	0.0	2.0
41.0	0.0	2.0	126.0	306.0	0.0	0.0	163.0	0.0	0.0	1.0
49.0	0.0	4.0	130.0	269.0	0.0	0.0	163.0	0.0	0.0	1.0
60.0	0.0	3.0	120.0	178.0	1.0	0.0	96.0	0.0	0.0	1.0
51.0	0.0	3.0	120.0	295.0	0.0	2.0	157.0	0.0	0.6	1.0
42.0	0.0	3.0	120.0	209.0	0.0	0.0	173.0	0.0	0.0	2.0
67.0	0.0	4.0	106.0	223.0	0.0	0.0	142.0	0.0	0.3	1.0
76.0	0.0	3.0	140.0	197.0	0.0	1.0	116.0	0.0	1.1	2.0
44.0	0.0	3.0	118.0	242.0	0.0	0.0	149.0	0.0	0.3	2.0
60.0	0.0	1.0	150.0	240.0	0.0	0.0	171.0	0.0	0.9	1.0
71.0	0.0	4.0	112.0	149.0	0.0	0.0	125.0	0.0	1.6	2.0
66.0	0.0	3.0	146.0	278.0	0.0	2.0	152.0	0.0	0.0	2.0
39.0	0.0	3.0	138.0	220.0	0.0	0.0	152.0	0.0	0.0	2.0
58.0	0.0	4.0	130.0	197.0	0.0	0.0	131.0	0.0	0.6	2.0
55.0	0.0	2.0	132.0	342.0	0.0	0.0	166.0	0.0	1.2	1.0

41.0	0.0	2.0	105.0	198.0	0.0	0.0	168.0	0.0	0.0	1.0
46.0	0.0	3.0	142.0	177.0	0.0	2.0	160.0	1.0	1.4	3.0
54.0	0.0	3.0	135.0	304.0	1.0	0.0	170.0	0.0	0.0	1.0
65.0	0.0	3.0	155.0	269.0	0.0	0.0	148.0	0.0	0.8	1.0
65.0	0.0	3.0	160.0	360.0	0.0	2.0	151.0	0.0	0.8	1.0
51.0	0.0	3.0	140.0	308.0	0.0	2.0	142.0	0.0	1.5	1.0
53.0	0.0	4.0	130.0	264.0	0.0	2.0	143.0	0.0	0.4	2.0
53.0	0.0	3.0	128.0	216.0	0.0	2.0	115.0	0.0	0.0	1.0
53.0	0.0	4.0	138.0	234.0	0.0	2.0	160.0	0.0	0.0	1.0
51.0	0.0	3.0	130.0	256.0	0.0	2.0	149.0	0.0	0.5	1.0
44.0	0.0	3.0	108.0	141.0	0.0	0.0	175.0	0.0	0.6	2.0
63.0	0.0	3.0	135.0	252.0	0.0	2.0	172.0	0.0	0.0	1.0
57.0	0.0	4.0	128.0	303.0	0.0	2.0	159.0	0.0	0.0	1.0
71.0	0.0	3.0	110.0	265.0	1.0	2.0	130.0	0.0	0.0	1.0
35.0	0.0	4.0	138.0	183.0	0.0	0.0	182.0	0.0	1.4	1.0
45.0	0.0	2.0	130.0	234.0	0.0	2.0	175.0	0.0	0.6	2.0
62.0	0.0	4.0	124.0	209.0	0.0	0.0	163.0	0.0	0.0	1.0
43.0	0.0	3.0	122.0	213.0	0.0	0.0	165.0	0.0	0.2	2.0
55.0	0.0	2.0	135.0	250.0	0.0	2.0	161.0	0.0	1.4	2.0
60.0	0.0	3.0	102.0	318.0	0.0	0.0	160.0	0.0	0.0	1.0

54.0	1.0	3.0	150.0	232.0	0.0	2.0	165.0	0.0	1.6	1.0
48.0	1.0	2.0	130.0	245.0	0.0	2.0	180.0	0.0	0.2	2.0
45.0	1.0	4.0	104.0	208.0	0.0	2.0	148.0	1.0	3.0	2.0
39.0	1.0	3.0	140.0	321.0	0.0	2.0	182.0	0.0	0.0	1.0
52.0	1.0	2.0	120.0	325.0	0.0	0.0	172.0	0.0	0.2	1.0
44.0	1.0	3.0	140.0	235.0	0.0	2.0	180.0	0.0	0.0	1.0
47.0	1.0	3.0	138.0	257.0	0.0	2.0	156.0	0.0	0.0	1.0
66.0	1.0	4.0	120.0	302.0	0.0	2.0	151.0	0.0	0.4	2.0
62.0	1.0	3.0	130.0	231.0	0.0	0.0	146.0	0.0	1.8	2.0
...	...	...	...	...	...	...	...	...	...	...
53.0	1.0	3.0	130.0	246.0	1.0	2.0	173.0	0.0	0.0	1.0
42.0	1.0	1.0	148.0	244.0	0.0	2.0	178.0	0.0	0.8	1.0
59.0	1.0	1.0	178.0	270.0	0.0	2.0	145.0	0.0	4.2	3.0
42.0	1.0	3.0	120.0	240.0	1.0	0.0	194.0	0.0	0.8	3.0
50.0	1.0	3.0	129.0	196.0	0.0	0.0	163.0	0.0	0.0	1.0
69.0	1.0	1.0	160.0	234.0	1.0	2.0	131.0	0.0	0.1	2.0
57.0	1.0	3.0	150.0	126.0	1.0	0.0	173.0	0.0	0.2	1.0

Testing Data Set

Age	Sex	Chest_Pain	Blood_Pressure	Cholestoral	Blood_Sugar	Electrocardiographic	Heart_Rate	Induced_Angina	ST_Depression	Slope
41.0	1.0	3.0	130.0	214.0	0.0	2.0	168.0	0.0	2.0	2.0
56.0	1.0	1.0	120.0	193.0	0.0	2.0	162.0	0.0	1.9	2.0
59.0	1.0	4.0	138.0	271.0	0.0	2.0	182.0	0.0	0.0	1.0
42.0	1.0	2.0	120.0	295.0	0.0	0.0	162.0	0.0	0.0	1.0
41.0	1.0	2.0	110.0	235.0	0.0	0.0	153.0	0.0	0.0	1.0
62.0	1.0	2.0	128.0	208.0	1.0	2.0	140.0	0.0	0.0	1.0
57.0	1.0	4.0	110.0	201.0	0.0	0.0	126.0	1.0	1.5	2.0
64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0
43.0	1.0	4.0	115.0	303.0	0.0	0.0	181.0	0.0	1.2	2.0
70.0	1.0	2.0	156.0	245.0	0.0	2.0	143.0	0.0	0.0	1.0
44.0	1.0	3.0	120.0	226.0	0.0	0.0	169.0	0.0	0.0	1.0
42.0	1.0	3.0	130.0	180.0	0.0	0.0	150.0	0.0	0.0	1.0
66.0	1.0	4.0	160.0	228.0	0.0	2.0	138.0	0.0	2.3	1.0
64.0	1.0	1.0	170.0	227.0	0.0	2.0	155.0	0.0	0.6	2.0
47.0	1.0	3.0	130.0	253.0	0.0	0.0	179.0	0.0	0.0	1.0
35.0	1.0	2.0	122.0	192.0	0.0	0.0	174.0	0.0	0.0	1.0
58.0	1.0	2.0	125.0	220.0	0.0	0.0	144.0	0.0	0.4	2.0
56.0	1.0	2.0	130.0	221.0	0.0	2.0	163.0	0.0	0.0	1.0
56.0	1.0	2.0	120.0	240.0	0.0	0.0	169.0	0.0	0.0	3.0
41.0	1.0	2.0	120.0	157.0	0.0	0.0	182.0	0.0	0.0	1.0

54.0	0.0	3.0	160.0	201.0	0.0	0.0	163.0	0.0	0.0	1.0
49.0	0.0	2.0	134.0	271.0	0.0	0.0	162.0	0.0	0.0	2.0
41.0	0.0	2.0	126.0	306.0	0.0	0.0	163.0	0.0	0.0	1.0
49.0	0.0	4.0	130.0	269.0	0.0	0.0	163.0	0.0	0.0	1.0
60.0	0.0	3.0	120.0	178.0	1.0	0.0	96.0	0.0	0.0	1.0
51.0	0.0	3.0	120.0	295.0	0.0	2.0	157.0	0.0	0.6	1.0
42.0	0.0	3.0	120.0	209.0	0.0	0.0	173.0	0.0	0.0	2.0
67.0	0.0	4.0	106.0	223.0	0.0	0.0	142.0	0.0	0.3	1.0
76.0	0.0	3.0	140.0	197.0	0.0	1.0	116.0	0.0	1.1	2.0
44.0	0.0	3.0	118.0	242.0	0.0	0.0	149.0	0.0	0.3	2.0
60.0	0.0	1.0	150.0	240.0	0.0	0.0	171.0	0.0	0.9	1.0
71.0	0.0	4.0	112.0	149.0	0.0	0.0	125.0	0.0	1.6	2.0
66.0	0.0	3.0	146.0	278.0	0.0	2.0	152.0	0.0	0.0	2.0
39.0	0.0	3.0	138.0	220.0	0.0	0.0	152.0	0.0	0.0	2.0
58.0	0.0	4.0	130.0	197.0	0.0	0.0	131.0	0.0	0.6	2.0
55.0	0.0	2.0	132.0	342.0	0.0	0.0	166.0	0.0	1.2	1.0

REFERENCES

[1] Prerana THM, Shivaprakash N C, Swetha N , Prediction of heart disease using machine learning algorithms-Naïve bayes , Introduction to PAC algorithm, Comaprison of algorithms and HDPS. International journal of science and engineering . November 2, 2015

[2] Sonam Nikhar, A.M Karandikar International journal of advanced engineering, Management and Science . June 6,2016

[3]Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel ,Heart disease prediction using machine learning technique and data mining,IJCSC,MARCH-2016

[4]S.Florence, N.G.Bhuvaneswari Amma, G.Annapoorani, K.Malathi, Predicting the Risk of Heart Attacks using Neural Network and Decision Tree International Journal of Innovative Research in Computer and Communication Engineering-NOVEMBER 2000

[5]Vikas Chaurasia, Saurabh Pal, Early Prediction of Heart Diseases Using Data Mining Techniques, Caribbean Journal of Science and Technology, september-2016

[6] R. Chitra and Dr.V. Seenivasagam, Heart Disease Prediction System Using Supervised Learning Classifier, Bonfring International Journal of Software Engineering and Soft Computing,March-2013

[7]R.Subha, Study On Cardiovascular Disease Classification Using Machine Learning Approaches, International Journal Of Engineering And Computer Science, Dec 2015

[8]Aditya Methaila , Prince Kansal , Himanshu Arya , Pankaj Kumar ,Early Heart Disease Prediction Using Data Mining Techniques,IJSCE.nov-2105

[9]Ankur Makwana , Jaymin Patel ,Decision Support System for Heart Disease Prediction using Data Mining Classification Techniques, International Journal of Computer Applications ,May 2015

[10]V.A. Kanimozhi, Dr. T. Karthikeyan, A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease ,International Journal of Advanced Research in Computer and Communication Engineering , April 2016

[11] V.Kirubha, S.Manju Priya ,Survey on Data Mining Algorithms in Disease Prediction ,International Journal of Computer Trends and Technology ,August 2016

[12] R.Subha, K.Anandakumar, A.Bharathi, Study on Cardiovascular Disease Classification Using Machine Learning Approaches, International Journal of Applied Engineering Research November 6

[13] J.Vijayashree and N.Ch.Sriman , NarayanaIyengar, Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques, International Journal of Bio-Science and Bio-Technology April 2016

[14] Mai Shouman, Tim Turner, and Rob Stocker, Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients, International Journal of Information and Education Technology, June 2012

[15] Nidhi Bhatla, Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques ,International Journal of Engineering Research & Technology, October 2012

[16]Dr. T. Karthikeyan, V.A.Kanimozhi,Deep Learning Approach for Prediction ofHeart Disease Using Data mining Classification Algorithm Deep Belief Network,International Journal of Advanced Research in Science, Engineering and Technology,January 2017
[17]A.J. Aljaaf, D.Al-Jumeily, A.J. Hussain, T. Dawson, P Fergus and M.Al-Jumaily, Predicting the Likelihood of Heart Failure with a Multi Level Risk Assessment Using Decision Tree, International Conference on Technological Advances in Electrical, Electronics and Computer Engineering,march - 2015
[18]Lakshmi Devasena C ,Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction ,International Journal of Computing Academic Research ,february-2016

Web link

[19] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[20] stackoverflow.com/questions/.../a-simple-explanation-of-naive-bayes-classification
[21] machinelearningmastery.com/naive-bayes-for-machine-learning/
[22] *https://en.wikipedia.org/wiki/Decision_tree_learning*
[23] https://dzone.com/articles/a-tutorial-on-using-the-big-data-stack-and-machine