A

Seminar Report

On

**“Heart Disease Prediction System”**

**( Using Machine Learning)**

Submitted in partial fulfillment of the requirements for the

Degree of

**BACHELOR OF TECHNOLOGY**

**(Semester -V)**

**In**

**Computer Science and Engineering**

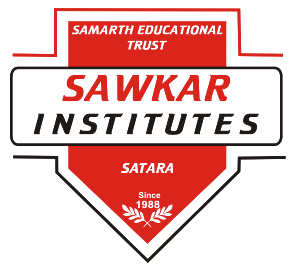
SUBMITTED BY

MR. SHRIDHAR PRAKASH AWARE

MISS. SAKSHI CHANDRASHEKHAR SHINDE

MISS. HARSHADA DATTATRAY JADHAV

MISS. ANAMIKA DHANANJAY GULUMKAR



Department of Computer Science and Engineering

**ARVIND GAVALI COLLEGE OF ENGINEERING, SATARA**

**2021-22**

Certificate

This is to certify that the Seminar report entitled **“Heart Disease Prediction System”**

is a bonafide work carried out by:

MR. SHRIDHAR PRAKASH AWARE

MISS. SAKSHI CHANDRASHEKHAR SHINDE

MISS. HARSHADA DATTATRAY JADHAV

MISS. ANAMIKA DHANANJAY GULUMKAR

under our supervision, during the year 2021-22 and submitted to the Faculty of Computer

Science and Engineering, AGCE, Satara in partial fulfilment of the requirements for the award of the

Degree of Bachelor of Technology in Computer Science and Engineering .

**(Project Coordinator) (HOD) (R&D Coordinator) (Principal)**

**(Internal Examiner) (External Examiner)**

**INDEX Page No**

[**ABSTRACT 4-5**](#_TOC_250003)

**CONTENTS 6**

CHAPTERS

1. INTRODUCTION 7-9

* General introduction
* Motive and backgrounds of the present work
* [Need](#_TOC_250002)
* LITERATURE REVIEW 11-14
* Literature review

1. SYSTEM DIAGRAMS 15-16
2. [System Architecture](#_TOC_250000) 19-21
3. Purpose of the system 17
4. Features 18-20

* [System Architecture](#_TOC_250000) 21-27
* Objective of study 28

1. Approach 29-30
2. Application 31
3. RESULT AND CONCLUSION 32-35

* **Result**

### Conclusion

* **Future Scope**

### 8.REFERENCES 36

* 1. **Research Papers**
  2. **Websites**

**ABSTRACT**

Cardiovascular disease is a major health burden worldwide in the

21st century. Human services consumptions are overpowering

national and corporate spending plans because of asymptomatic

infections including cardiovascular ailments. Consequently,

there’s an urgent requirement for early location and treatment of

such ailments. The information which is gathered by data

analysis of hospitals is utilizing by applying different blends of

calculations and algorithms for the early-stage prediction of

Cardiovascular ailments. Machine Learning is one among the

slanting innovations utilized in numerous circles far and wide

including the medicinal services application for predicting

illnesses. The proposed project is predicated on a typical

machine learning algorithm k-nearest algorithm

The healthcare domain is one of the prominent research fields in the current scenario with the rapid improvement of technology and data. It is difficult to handle the huge amount of data of the patients. It is easier to handle this data through Big Data Analytics. There are a lot of procedures for the treatment of multiple diseases across the world. Machine Learning is an emerging approach that helps in prediction, diagnosis of a disease. This paper depicts the prediction of disease based on symptoms using machine learning.

Machine Learning algorithms such as Naive Bayes, Decision Tree and Random Forest are employed on the provided dataset and predict the disease. Its implementation is done through the python programming language. The research demonstrates the best algorithm based on their accuracy. The accuracy of an algorithm is determined by the performance on the given dataset.

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis from [Kaggle](https://www.kaggle.com/ronitf/heart-disease-uci). My complete project is available at [Heart Disease Prediction](https://github.com/kb22/Heart-Disease-Prediction). In this article, I’ll discuss a project where I worked on predicting potential Heart Diseases in people using Machine Learning algorithms. The algorithms included K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier.

The dataset has been taken.Cardiovascular disease is a major health burden worldwide in the 21st century. Human services consumptions are overpowering national and corporate spending plans because of asymptomatic infections including cardiovascular ailments. Consequently, there’s an urgent requirement for early location and treatment of such ailments.

The information which is gathered by data analysis of hospitals is utilizing by applying different blends of calculations and algorithms for the early-stage prediction of Cardiovascular ailments. Machine Learning is one among the slanting innovations utilized in numerous circles far and wide including the medicinal services application for predicting illnesses. The proposed project is predicated on a typical machine learning algorithm k-nearest algorithm Cardiovascular disease is a major health burden worldwide in the 21st century.

Human services consumptions are overpowering national and corporate spending plans because of asymptomatic infections including cardiovascular ailments. Consequently, there’s an urgent requirement for early location and treatment of such ailments.

The information which is gathered by data analysis of hospitals is utilizing by applying different blends of calculations and algorithms for the early-stage prediction of Cardiovascular ailments. Machine Learning is one among the slanting innovations utilized in numerous circles far and wide including the medicinal services application for predicting illnesses. The proposed project is predicated on a typical

## Content

**Chapter-1**

## INTRODUCTION

## GENERAL INTRODUCTION ABOUT PRESENTATION :-

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models.

Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease.

Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India.

Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body.

Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge.

This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

## MOTIVE ABOUT PRESENT WORK :-

## A major Challenge facing health care organisation hospitals medical centres is the provision of quality services at affordable costs qualities service implies diagnosting patients correctly and administrating treatments that are effective poor clinical decisions can lead to digestive stores consequences which are there for an acceptable hospitals must also minimise the cost of clinical test they can achieve these results by employing appropriate computer based information and or decision support system.

## Most hospitals today employee some sort of hospital information system to message their healthcare or patient data this system typically generate shoes amount of data which taken the form of members takes charge and images unfortunately this data are rarely used to support clinical decision making there is wealth of hidden information in this data that is largely untapped these raises and important question how we can turn data into useful information that can unable to make intelligent clinical decisions this is the main motivation for this research.

## Heart disease (HD) is one of the most common diseases nowadays, due to number of contributing factors, such as high blood pressure, diabetes, cholesterol fluctuation, exhaustion and many others. An early diagnosis of such disease has been sought for many years, and many data analytics tools have been applied to help health care providers to identify some of the early signs of HD. Many tests can be performed on potential patients to take the extra precautions measures to reduce the effect of having such a disease [[1](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03626-y#ref-CR1)], and reliable methods to predict early stages of HD, such as the methods proposed in this paper, can be a crucial task for saving lives. Number of Machine Learning (ML) algorithms, such as, Naïve Bayes, Stochastic Gradient Descents (SGD), Support Vector Machine (SVM), K- Nearest Neighbor (K-NN), Adaboost, JRip, Decision tree J48, and others were applied for the purpose of classification and prediction of HD dataset, and many promising results were presented in the literature [[2](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03626-y#ref-CR2)].

## 

**Chapter-2**

## LITERATURE REVIEW

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

[1] “Efficient Heart Disease Prediction System”

Using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

[2] Prediction of Heart Disease Using Machine Learning Algorithms”

Using decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables. But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[3] Training and Testing of Dataset

“Prediction of Heart Disease Using Machine Learning”, in which training and testing of dataset is

performed by using neural network algorithm multi-layer perceptron. In this algorithm there will

be one input layer and one output layer and one or more layers are hidden layers between these

two input and output layers. Through hidden layers each input node is connected to output layer.

This connection is assigned with some random weights. The other input is called bias which is

assigned with weight based on requirement the connection between the nodes can be

feedforwarded or feedback.

[4]“Heart Disease Prediction using Evolutionary Rule Learning”.

Data is directly retrieved from electronic records that reduce the manual tasks. The amount of

services are decreased and shown major number of rules helps within the best prediction of heart

disease. Frequent pattern growth association mining is performed on patient’s dataset to generate

strong association.

[5]“Heart Attack Prediction Using Deep Learning”

In which heart attack prediction system by using Deep learning techniques and to predict the

Probable aspects of heart related infections of the patient Recurrent Neural System is used. This

model uses deep learning and data mining to give the best precise model and least blunders. This

paper acts as strong reference model for another type of heart attack prediction models

.

[6]Most of the papers have implemented several Machine learning techniques

For prediction of heart disease such as classification trees, naive bayes, neural network, support vector machine and K-nearest neighbor algorithm. Some of the papers also implemented feature selection techniques such as wavelet transformation, principle component analysis and information gain module for identifying important attributes for efficient performance of classifier for prediction of heart disease. One of the bases on which the papers differ is the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. [15]In this paper, decision support system for heart disease prediction is described. Integer coded Genetic Algorithm is used to find optimized feature subset for maximizing classification accuracy with a reduced number of features

## Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models.

## Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.

## The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease.

## Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India.

## Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases.

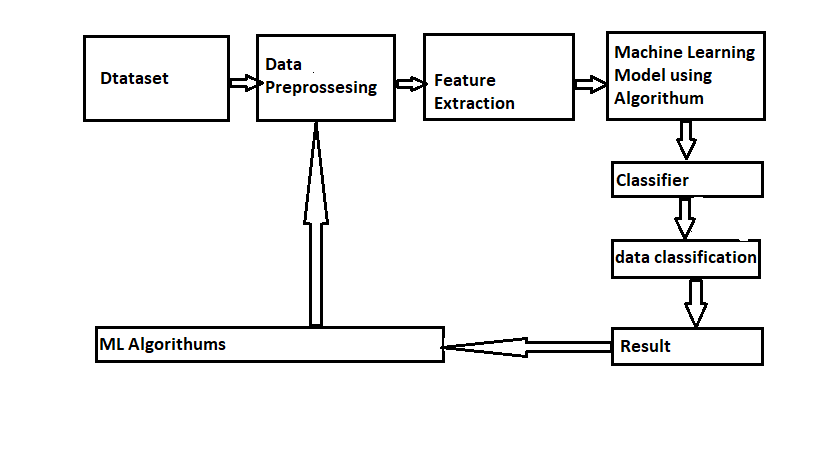
## The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body.

**Chapter 3**

**SYSTEM DESINING/Block Diagram**

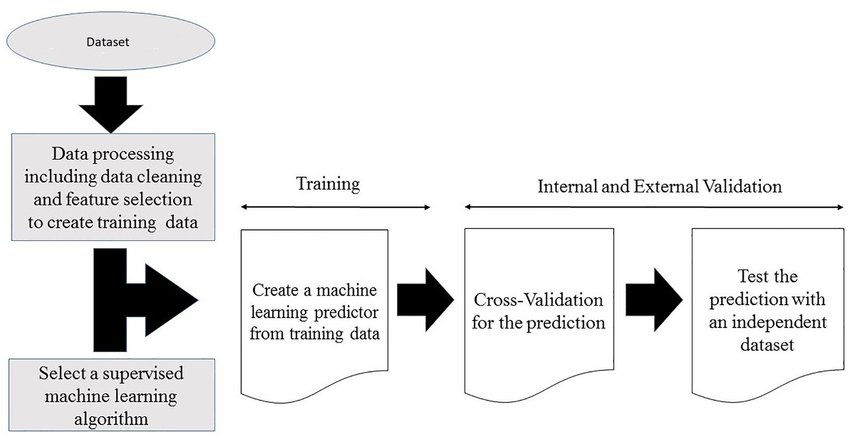
**1] Heart Disease Prediction System:**

**(a)Block Diagram -**



**2] Heart Disease Prediction System:-**

**(b)System Design -**



## **Purpose of Heart Disease Prediction System**

## What is the purpose of HDPS(Heart Disease Prediction System)?

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

This system is implemented using the following modules.

1. Collection of Dataset
2. Selection of attributes
3. Data Pre-Processing

4.) Balancing of Data

5.) Disease Prediction

Collection of dataset

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

2 Selection of attributes:-

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

## .3 Pre-processing of Data:-

## Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

## .4 Balancing of Data:-

## Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

## 5 Prediction of Disease:-

## Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

**Chapter 3**

## **Working of Heart Disease Prediction System**

The system architecture gives an overview of the working of the system.

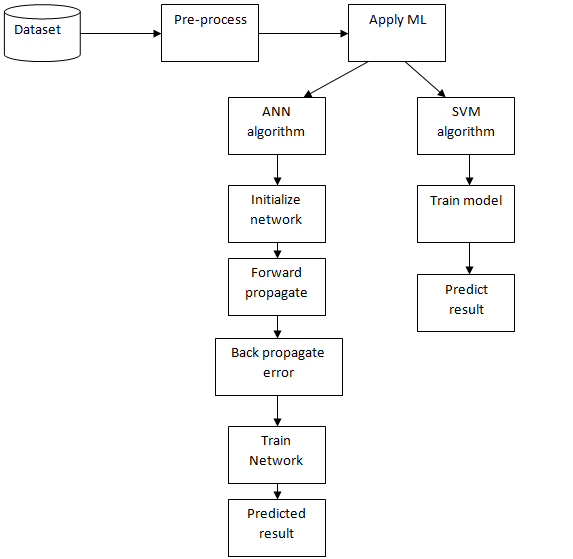
The working of this system is described as follows:

Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.

The Proposed Architecture:-

* In this system we are implementing effective heart attack prediction system using Naïve Bayes algorithm. We can give the input as in CSV file or manual entry to the system.
* After taking input the algorithms apply on that input that is Naïve Bayes. After accessing data set the operation is performed and effective heart attack level is produced.
* The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts.
* The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Diagram:-System Architecture



Definitions, Acronyms, and Abbreviations

Definitions:

The stud

Definitions, Acronyms, and Abbreviations

Definitions:

The stud

f

## Functional Requirements of the Heart Disease Prediction System:-

The functional requirements of this system are:

* In this system have two actors are there admin and user.
* Admin has to login using his username and password.
* Train the machine using machine learning approach.
* In user part they have to register themselves.
* User has to login using user id and password.
* Then user has to input patient data, and then based on.
* Trained model user input data will check and it will give the output.

## Non- Functional Requirements of the Heart Disease Prediction System:

## Nonfunctional necessities describe however a system should behave and establish constraints of its practicality. This type of requirements is also known as the system’s quality attributes. Attributes such as performance, security, usability, compatibility are not the feature of the system, they are a required characteristic. They are "developing" properties that emerge from the whole arrangement and hence we can't compose a particular line of code to execute them. Any attributes required by the customer are described by the specification. We must include only those requirements that are appropriate for our project

## Reliability - The structure must be reliable and strong in giving the functionalities. The movements must be made unmistakable by the structure when a customer has revealed a couple of enhancements. The progressions made by the Programmer must be Project pioneer and in addition the Test designer.

## Performance -The framework will be utilized by numerous representatives all the while. Since the system will be encouraged on a single web server with a lone database server outside of anyone's ability to see, execution transforms into a significant

## System Architecture

## **C:\Users\Harshada Jadhav\Downloads\sys architecture.png**

The architecture of the proposed system is as displayed in the figure below. The major components of the architecture are as follows: patient database, preprocessing, tokenization, training the model, test the model, design fitness function, application of genetic algorithm, results collection and prediction of heart disease.

1)Patient database - The dataset as provided by University of California, Irvine Machine Learning Repository is initially imported for the analysis of this system work. The dataset consists of the following attributes: age, chest pain, blood pressure, cholesterol, diabetes, ECG, heart rate, physical activity, slope, thalassemia,

3)Preprocessing -

Preprocessing is a significant stage in the knowledge discovery process. Real world data tends to be noisy and inconsistent. Data processing techniques like data cleaning etc help in overcoming these drawbacks. Normalization of the dataset helps in classify the data which further makes the data to smoothly allow algorithms to execute with efficient results.

4)Tokenization-

In tokenization, the data will be clubbed into set of meaning sentences or chunks for further processing. This will further enhance the efficiency of the data that has undergone preprocessing.

5) Training the model -

In the training part, the backpropagation algorithm as mentioned above will be implemented. backpropagation helps in finding a better set of weights in short amount of time. The training is done on basis of the dataset input to the system. Herein ‘min max’ function is implemented so as to gain a matrix of minimum and maximum values as specified in its argument. This function is applied for training of the network. The efficiency of the system can be improved every instance as many times the model is trained, the number of iterations etc. The whole dataset provided which consists of 13 attributes and 872 rows will help the model undergo training. Training can also be implemented by splitting the data in equalized required amount of data partitions. In the user interactive GUI, as the user will select train network option after entering his data at the backend the .csv file of UCI dataset will be read and normalization will be carried out so as to classify the data into classes which becomes easier to be fed onto the neural network. the neural network that is created here will be consisting of three layers namely: input layer, hidden layer and output layer. Hidden layers can be customized to 2 or 3 as per users requirements. To generate a network, train() function is implemented so as to pass the inputs. this network will be stored in .mat file. After the network is generated, we check for mean square error.

6)Testing the model -

Testing will be conducted so as to determine whether the model that is trained is providing the desired output. As the data is entered for testing, the .csv file will be retrieved to crosscheck and then compare and the results of the newly entered data will be generated. On basis of how the model is trained with the help of the dataset, the user will input values of his choice to the attributes specified and the results will be generated as the whether there is a risk of heart disease or not.

7) Design fitness function of genetic algorithm – The genetic algorithm is applied so as to initialize neural network weight. The genetic algorithm is used to evaluate and calculate the number of layers in the neural network along with the total number of weights used and bias. The initial population is generated at random. Bias is used such that the output value generated will not be 0 or negative. On basis of the mean square error calculated during testing, the fitness function of each chromosome will be calculated. Ater selection and mutation is carried out in genetic algorithm, the chromosome consisting of lower adaptation are replaced with optimizedd one that is better and fitter chromosomes

8) Prediction of heart disease – The genetic algorithm is applied so as to initialize neural network weight. The genetic algorithm is used to evaluate and calculate the number of layers in the neural network along with the total number of weights used and bias. The initial population is generated at random. Bias is used such that the output value generated will not be 0 or negative. On basis of the mean square error calculated during testing, the fitness function of each chromosome will be calculated. Ater selection and mutation is carried out in genetic algorithm, the chromosome consisting of lower adaptation are replaced with optimizedd one that is better and

This component will help in predicting the severity of the cardiovascular disease. When user will input data, the weights will be cross checked with the given inputs. The prediction neural network will consist of 13 nodes as a part of input layer considering that 13 attribute values will be input to the system. Then the hidden layer and one node in the output layer which will provide the result. The predicted will be generated in the form of a ‘yes’ or ‘no’ format considering all the risk factors whether they lie in the criteria as per the model is trained.

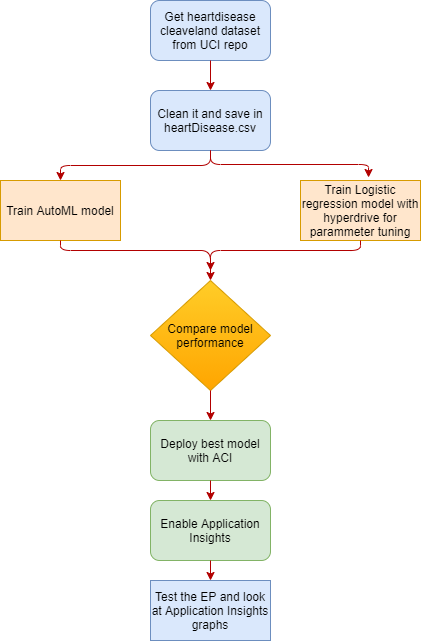
## What Heart Disease Prediction System Indicate:-

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

Among various life-threatening diseases, heart disease has garnered a great deal of attention in medical research. The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise.

A major challenge faced by health care organizations, such as hospitals and medical centers, is the provision of quality services at affordable costs.[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/#b1-ijn-13-121) The quality service implies diagnosing patients properly and administering effective treatments. The available heart disease database consists of both numerical and categorical data. Before further processing, cleaning and filtering are applied on these records in order to filter the irrelevant data from the database.[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/#b2-ijn-13-121) The proposed system can determine an exact hidden knowledge, ie, patterns and relationships associated with heart disease from a historical heart disease database. It can also answer the complex queries for diagnosing heart disease; therefore, it can be helpful to health care practitioners to make intelligent clinical decisions. Results showed that the proposed system has its unique potency in realizing the objectives of the defined mining goals.

Heart Disease Prediction System Software:



## Methods:-

## The experiment was carried out on a publicly available database for heart disease. The dataset contains a total of 303 records that were divided into two sets, training set (40%) and testing set (60%). A data mining tool named Weka 3.6.11 was used for the experiment. Additionally, multilayer perceptron neural network (MLPNN) with backpropagation (BP) was used as the training algorithm.

The overall process of effective heart disease prediction system (EHDPS) is based:

* Data collection
* Data pre-processing and
* The classification of data.

## The data are collected from a standard dataset that contains 303 records. The 15 parameters, such as age, sex, chest pain type (CP), and cholesterol (chol), with some domain values associated with them, considered to predict the probability of heart disease are shown in table.

Functionality-

* The Student Information System allows the students to upload their information, enroll to the courses and view the remarks given by their tutor. This system also allows the tutor to view the students enrolled in his/her course and give them remarks.
* The system also allows parents to view the remarks on their child’s progress given by the tutor.
* The college accountant can register himself to the Student Information System and manage the fees record for the Self-Financed students. Figure 3 show other functionality built in the system.

# The Data:-

The dataset consists of 303 individuals data. There are 14 columns in the dataset, which are described below.

1. Age: Displays the age of the individual.
2. Sex: Displays the gender of the individual using the following format   
    1=male  
    0 = female
3. Chest-pain type: displays the type of chest-pain experienced by the individual using the the following format :  
    1 = typical angina  
    2 = atypical angina  
    3 = non — anginal pain  
    4 = asymptotic
4. Resting Blood Pressure: Displays the resting blood pressure value of an individual in mmHg (unit).
5. Serum Cholestrol: displays the serum cholesterol in mg/dl (unit)
6. Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl.  
   If fasting blood sugar > 120mg/dl then : 1 (true)  
   else : 0 (false)
7. Resting ECG: displays resting electrocardiographic results  
   0 = normal  
   1 = having ST-T wave abnormality  
   2 = left ventricular hyperthrophy
8. Max heart rate achieved: displays the max heart rate achieved by an individual.
9. Exercise induced angina :  
   1 = yes  
   0 = no
10. ST depression induced by exercise relative to rest: displays the value which is an integer or float.
11. Peak exercise ST segment :  
    1 = upsloping  
    2 = flat  
    3 = downsloping
12. Number of major vessels (0–3) colored by flourosopy : displays the value as integer or float.
13. Thal: displays the thalassemia :  
    3 = normal  
    6 = fixed defect  
    7 = reversible defect
14. Diagnosis of heart disease : Displays whether the individual is suffering from heart disease or not :  
    0 = absence  
    1, 2, 3, 4 = present.

## Why these parameters :-

## **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.

1. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman’s risk is similar to a man’s although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
2. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn’t get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.
3. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.
4. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of a heart attack.
5. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body’s blood sugar levels to rise, increasing your risk of a heart attack.
6. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.
7. **Max heart rate achieved**: The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
8. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.
9. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’ test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

## MODULES:

* Admin
* Users
* Disease Prediction

## MODULE DESCRIPTIONS:

Admin**:**

In this Module, Admin can add the Doctor details and the Training datasets. The doctor detail consists of fields such as name, contact and specialist in. Similarly Training Datasets consists of previous analyzed data’s of patient’s history, such as Blood Sugar, Blood Pressure, Heart Rate and etc.

Users**:**

Intended Users who wants to predict the possibility of disease they are suffered, they will give the information about their conditions. The System will show the analyzed Results from the training datasets to the Users.

Disease Prediction**:**

* Disease Prediction will analyzed from training datasets through Data mining techniques:

K-mean clustering: .The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information into k groups, where k is considered as an input parameter. It then assigns each information to clusters based upon the observation’s proximity to the mean of the cluster.

* Prophecy of heart disease using K – Means clustering techniques:

The algorithm arbitrarily selects k points as the initial cluster centers (“means”).

Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.

* Each cluster center is recomputed as the average of the points in that cluster.

Steps 2 and 3 repeat until the clusters converge. Convergence may be explained differently depending upon the performance, but it regularly explains that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

.

## 

## OBJECTIVES OF THE STUDY

## There are enormous amount of data available from medical industry which could be useful for medical practitioners when it is used for discovering hidden pattern with help of existing data mining techniques. The basic medical records from a patient’s profile can be useful in identifying hidden pattern with data mining techniques. In this paper, NaÃ¯ve Bayes algorithm to predict heart disease is implemented with basic records of patients like age, sex, heart rate, blood pressure etc., from a sample dataset. The benefits, limitations, and technical details of this implementation will also be discussed in this paper.

Over these years in medical history, many types of medical problems have been identified and many data are available regarding a particular problem. But not all the medical data are same, but there are many patterns hidden inside those data which needs to be identified. Data mining techniques could help identify these hidden patterns by knowledge discovery. In the medical field, patient’s health issues are predicted by doctor’s intuition or experience [2] where the knowledge rich data is suppressed which results in high medical expenses and unnecessary medical tests. In recent years, there are many researches being conducted in order to find the hidden pattern from basic medical data [1]. Identifying these hidden pattern would result in a developing an efficient decision making system in medical industry which aide as a tool to support doctor’s decision making or at least serve as a prediction system for any medical issues.

In this paper, we have taken into consideration of heart disease and predict it using the set of data that are already in existence with the help of data mining technique. The algorithm that we have chosen is the NaÃ¯ve Bayes algorithm, this algorithm is ideal for a vast amount of database that may contain hundreds and thousands of rows and columns. The NaÃ¯ve Bayes algorithm provides the intended output faster and more accurate as the number of data in the database increase.In this paper, NaÃ¯ve Bayes algorithm to predict heart disease is implemented with basic records of patients like age, sex, heart rate, blood pressure etc., from a sample dataset. Based on the literature survey NaÃ¯ve Bayes algorithm was found to be an effective technique. The probabilistic method helped in finding the converse probability of the conditional relationship. The dependence relation may exist between two attributes of data set which can be determined with this algorithm.

Building and running a Demo of Heartg Disease Prediction System -

TOMCAT server is used to present the output in web based form. The output will run in localhost. The MySQL database is used to identify the patient records. At the execution point, the local host is accessed and 15 questions will be displayed which will be obtained from user and algorithm will be called to calculate and predict the disease possibility on that person. A report will be generated at the end of the demo which says if the person is predicted with heart disease or not.

In general,

1. Obtains the values from user.

2. Reads the data file.

3. Calls the algorithm and calculates mean, deviation, and probability of attributes.

4. Generates a report displaying the values given with the prediction of disease

1 Problem Scenario

There are only few decision support systems available in medical industry whose functionalities are very limited. As mentioned earlier, medical decisions are made with doctor’s intuition and not from the rich data from the medical database. Wrong treatment due to misdiagnosis causes serious threat in medical field. In order to solve these issues data mining solution was with help of medical databases was introduced.

2 Related Work

There are many techniques available to discover knowledge from medical database [1]. Researchers at Southern California used data mining technique to discover the success and failure of back surgery in order to improve medical treatment [3]. Shouman et al [4] implemented predictive data mining to diagnose heart disease of patients. Palaniappan et al [2] developed a prototype Intelligent Heart Disease Prediction System (IHDPS), using data mining techniques.

.

**APPENDIX**

* Python:-

Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant White space. Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

* Sklearn:-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

* Numpy:-

NumPy is a library for the python programming language, adding support for large, multi- dimensional arrays and matrices, along with a large collection of high level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim with contributions from several other developers. In 2005, Travis created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open source software and has many contributors.

* Librosa

Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation(using LSTMs), Automatic Speech Recognition. 67 It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques.

* Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a statemachine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.

* Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a highlevel interface for drawing attractive and informative statistical graphics. Seaborn is a library in Python predominantly used for making statistical graphics. Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

* SciPy

SciPy contains modules for optimization, linearalgebra, integration, interpolation, special functions, FFT, signal and imageprocessing, ODE solvers and other tasks common in science and engineering. SciPy is also a family of conferences for users and developers of these tools: SciPy (in the United States), EuroSciPy (in Europe) and SciPy.in (in India). Enthought originated the SciPy conference in the United States and continues to sponsor many of the international conferences as well as host the SciPy website. SciPy is a scientific computation library that uses NumPy underneath. It provides more utility functions for optimization, stats and signal processing.

**Chapter-4**

Feature Analysis on Related Work

Table 1 below depicts feature analysis on similar systems addressed in the previous section. Collectively, these

systems pose the features as described below:

• Student – manage student record

• Placement – manage student’s practical training (industrial placement)

• Notices – information on various events

• Registration – handles student course registration

• Hostel Management – student hostel management

• e-Library – focuses on collection and management of digital library material

• e-Voting - electronic means to either aid or take care of casting and counting votes

From the above, some features (Student, Notices) are suitable to be adapted in our proposed student activities

management system. Additional features need to be obtained by acquiring user requirements to compliment results

obtained from the feature analysis performe

Feature Analysis on Related Work

Table 1 below depicts feature analysis on similar systems addressed in the previous section. Collectively, these

systems pose the features as described below:

• Student – manage student record

• Placement – manage student’s practical training (industrial placement)

• Notices – information on various events

• Registration – handles student course registration

• Hostel Management – student hostel management

• e-Library – focuses on collection and management of digital library material

• e-Voting - electronic means to either aid or take care of casting and counting votes

From the above, some features (Student, Notices) are suitable to be adapted in our proposed student activities

management system. Additional features need to be obtained by acquiring user requirements to compliment results

obtained from the feature analysis performe

Feature Analysis on Related Work

Table 1 below depicts feature analysis on similar systems addressed in the previous section. Collectively, these

systems pose the features as described below:

• Student – manage student record

• Placement – manage student’s practical training (industrial placement)

• Notices – information on various events

• Registration – handles student course registration

• Hostel Management – student hostel management

• e-Library – focuses on collection and management of digital library material

• e-Voting - electronic means to either aid or take care of casting and counting votes

From the above, some features (Student, Notices) are suitable to be adapted in our proposed student activities

management system. Additional features need to be obtained by acquiring user requirements to compliment results

obtained from the feature analysis performe

## APPROCH

## Approch:

**Experimental Analysis**

System Configuration :-

* Hardware requirements:

Processer : Any Update Processer

Ram : Min 4GB

Hard Disk : Min 100GB

* Software requirements:

Operating System : Windows family

Technology : Python3.7

IDE : Jupiter notebook

## **Result**

## Result

## After performing the machine learning approach for training and testing we find that accuracy of the XG-boost is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 81% accuracy and the comparison is shown below.

## TABLE: Accuracy comparison of algorithms Algorithm Accuracy:-

## 

## Algorithm Accuracy

## XG-boost 81.3%

## SVM 80.2%

## Logistic Regression 79.1%

## Random Forest 79.1%

## Naive Bayes 76.9%

## Decision Tree 75.8%

## Adaboost 73.6%

**Chapter-6**

**Conclusion**

* Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. Alternating decision trees when used with PCA, have performed extremely well but decision trees have performed very poorly in some other cases which could be due to overfitting. Random Forest and Ensemble models have performed very well because they solve the problem of overfitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest). Models based on Naïve Bayes classifier were computationally very fast and have also performed well.SVM performed extremely well for most of the cases.
* Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.
* In this paper, NaÃ¯ve Bayes algorithm is the only algorithm used for calculation of attributes and prediction. Efficiency and accuracy of the algorithm in predicting were discussed. Designing effective models are constrained by size of the datasets and noisy, incorrect, missing data values. The prototype developed so far has been generally tested by computer experts and not by the doctors. For effective understanding of the health issues, medical experts have to work collaboratively and test the prototypes in order to implement the system in real life to support medical experts in taking clinical decision
* All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 81%.

**Chapter 7**

**Future Scope**

Further research should be done to increase classification accuracy through the use of advanced algorithms such as Bagging, Vector Machine Support or table decision etc. Determine the performance of the predictions per algorithm and apply the proposed system to the area of interest. We can add more features needed to improve accuracy implementation of algorithms. Stakeholders should use it as dedicated tool to make better decisions. We did not change parameters in our implementation. In future, it can be improved and adjust by changing the parameters for the experiment. In the future, more work can be done by using more data related to heart disease and by using different data reduction techniques. For better results and predictions of heart disease, high quality oriented datasets can be used which are free from inconsistencies

The algorithm is working well with this sample dataset. Implementing the algorithm with large dataset could give better results which can aid as a supporting tool in making medical decisions. In future, other possible algorithms could be implemented where efficiency of all algorithms could be analyzed to decide on best suitable technique in terms of speed, reliability, and accuracy.

## **Refrences**

## RESEARCH PAPER

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8.

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072

## Website

* https://www.ukessays.com/essays/computer-science/the-heart-disease-prediction-system-computer-science-essay.php
* https://nthu-datalab.github.io/ml/labs/03\_Decision-Trees\_RandomForest/03\_Decis
* https://search.brave.com/search?spellcheck=0&q=objective+of+the+heart+disease+prdiction+system
* https://www.kaggle.com/ronitf/heart-disease-uci