

HEART DISEASE PREDICTION MODEL BASED ON DATA MINING

Submitted in partial fulfillment of the requirements for the award of Bachelor
of Engineering Degree in Computer Science and Engineering

By

VIJAY PRASANNA.S (Reg. No:39290126)

NAVIN.S (Reg. No:39290070)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**

APRIL 2022



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **VIJAY PRASANNA.S (Reg. No. 39290126)** and **NAVIN.S (Reg. No.39290070)** who carried out the project entitled “**HEART DISEASE PREDICTION MODEL BASED ON DATA MINING**” under my supervision from November 2021 to April 2022.

Internal Guide

Dr.JANCY.

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on

Internal Examiner External Examiner

DECLARATION

I **VIJAY PRASANNA.S (Reg. No. 39290126)**, hereby declare that the Project Report entitled "**HEART DISEASE PREDICTION MODEL BASED ON DATA MINING**" done by me under the guidance of **Dr. JANCY** is submitted in partial fulfillment of the requirements for the award of Bachelor of Science degree in Computer Science. **DATE:**

**PLACE: CHENNAI SIGNATURE OF THE CANDIDATE
ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA, M.E., Ph.D.**, Dean, School of Computing and **Dr. S. VIGNESHWARI, M.E., Ph.D.**, and **Dr. L. LAKSHMANAN, M.E., Ph.D.**, Head of the Department, Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.JANCY** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways For the Completion of the project.

ABSTRACT

Diagnosing the Heart disease in its initial stage is really one of the essential and difficult tasks to be done. Hence this paper is mainly based on prediction of Heart disease based on the systemic efforts with a higher accuracy rate. During this work, six machine learning classification algorithms are studied namely Logistic Regression, Gaussian Naive Bayes algorithm, SVM, DTrees, K-NN and Random Forest Classifier are evaluated on various measures like finding test

data and train data accuracy. Our dataset is Cleveland Heart Disease Dataset. Python code is implemented in the Jupyter platform under Anaconda Navigator. In future works, we can apply advanced classifiers such as neural networks over this dataset. From this study, we can provide some rules and regulations for the improvisation of methods; thereby further studies will be gained to predict the patients even before their diagnosis.

v
Table of Contents

Bona-fide Certificate

ii Declaration

iii Acknowledgements

iv Abstract

v List of Figures

viii List of Tables ix

1. Introduction

1.1. Major Hazards of illness

1.2. Outcomes and Significant Chances of Heart disease

1.3. Motivation

2. Objective

3. Methodology

3.1. Preprocessing

3.2. Dataset Selection

3.3. Jupyter Lab Tool

3.4. Data Preprocessing

3.5. Data Splitting

3.6. Classifier

3.6.1. Random Forest Algorithm

3.6.2. KNN Algorithm

3.7. Achieving Accuracy

3.7.1. Test Accuracy

3.7.2. Train Accuracy

3.7.3. Confusion Matrix

4. Results and Discussion

4.1 Dataset

4.2 Classification Of Algorithms

4.2.1 List of Algorithms

4.2.1.1 KNN

4.2.1.2 Random Forest

4.3 Classification Table

4.4 Mean Accuracy

4.5 Accuracy Obtained

4.6 Bar Plot of the prediction model

5. Conclusion and Further Work 32

6. References 33

vii

List of Figures

Figure No.	Title	Page No.
3.1	Workflow of prediction model	08
3.2	Cleveland Heart Disease Dataset	08
3.3	Barplot of affected and unaffected patients	10
4.1	Confusion matrix of Logistic Regression	19
4.2	Confusion matrix of Gaussian Naive Bayes	20
4.3	Confusion matrix of Support Vector Machine	22
4.4	Confusion matrix of K-Nearest Neighbors	24
4.5	Confusion matrix of Decision Tree	26

4.6	Confusion matrix of Random Forest	28
4.7	Barplot of a prediction model	31

viii
List of Tables

Table No.	Table Name	Page No.
4.1	Confusion Matrix for Logistic Regression	19
4.2	Confusion Matrix For Gaussian Naive Bayes	21
4.3	Confusion Matrix for Support Vector Machine	22
4.4	Confusion Matrix For K-Nearest Neighbors	24
4.5	Confusion Matrix For Decision Tree	26
4.6	Confusion Matrix For Random Forest	28
4.7	Classification Table	29
4.8	Accuracy of model	30

ix
CHAPTER 1

1.INTRODUCTION

1.1.Major Hazards of illness

In everyday life numerous components that influence a human heart. Numerous issues are

happening at a fast pace and new heart sicknesses are quickly being distinguished. In this day and age of pressure Heart, being an basic organ in a human body which pumps blood through the body for the blood dissemination is fundamental and its well being is to be saved for a sound living. As indicated by the World Health Organization, consistently in excess of 12 million passings are happening worldwide because of the different kinds of heart illnesses which is additionally known by the term cardiovascular infection. The term Heart infection incorporates numerous illnesses that are differing also, explicitly influencing the heart and the courses of a person. Indeed, even youthful matured individuals around their 20-30 years of life expectancy are getting influenced by heart maladies. The expansion in the chance of coronary illness among youthful might be because of the terrible dietary patterns, absence of rest, fretful nature, despondency and various other factors, for example, corpulence, less than stellar eating routine, family ancestry, high blood pressure, high blood cholesterol, inert conduct, family ancestry, smoking and hypertension. A danger of a heart attack or the chance of the coronary illness whenever recognized early, can help the patients play it safe and take administrative measures. The human services industry has been creating immense measures of information about patients and their infection conclusion reports are by and large particularly taken for the expectation of coronary failures around the world. At the point when the information about coronary illness is tremendous, the classification techniques can be implemented for the examination.

Data Mining is a process of removing the crucial dynamic data from a group of past records for future examination or on the other hand expectation. The data might be covered up and isn't recognizable without the utilization of data mining. The order is one data mining method through which the future result or expectations can be made dependent on the chronicled information that is accessible. The clinical information mining made a potential answer for incorporating the grouping methods and giving electronic preparation on the dataset that further prompts investigating the covered up designs in the clinical informational indexes which is utilized for the forecast of the patient's future state. In this way, it is conceivable to give bits of knowledge on a patient's history and can offer clinical help through the examination.

1.2 Outcomes and Significant Chances of Heart Disease

Traditional classification algorithms cluster the disease with visibly classified symptoms and there is no way to process the complications and difficulties faced due to the inadequate information. Sometimes, due to inadequate information, the treatment data provided by the related algorithms can be erroneous and inappropriate.

In some cases a disease can show different symptoms at different stages but because of lack of risk differentiation, the traditional classification algorithm can diagnose it as a different disease. So if the patient continues with the same treatment it can cause adverse effects with implication and spawning of a syndrome possibly with pestilential effects. Compared to other clustering algorithms the DPCA works efficiently even in enormous datasets and provides much reliable results. The Apriori algorithm is used to make the association rules and analysis of the relationship between the disease symptoms cluster and the treatment recommended to it after pre-processing. The minimum support and confidence values are pre-programmed and updated in quick succession. If the support and confidence values are greater than the minimum support and confidence rules, the frequent item is obtained. The odds of low density data point processing are very low because it takes global data density distribution only into the account. It fails to identify clusters in the case like varying density and highly sparse dataset. The partition becomes complex for multiprocessor systems. Because of dealing with a large number of candidate item set generation, Apriori algorithm requires large memory space Compared to the other

algorithms, its computational time is very high . To overcome the accuracy precision problem and those space consumption Many new classification algorithms came into the data mining field. Using those as a key, the target is to identify certain new patterns and perceive these patterns to provide users with meaningful and useful relevant data. Heart disease leads to respiratory disease , Breathing disorder, emotional disability and blockage leads to increased tension in the blood thereby increasing the risk of blood pressure. This lowers the data on cardio illness in a critical problem is a productive route. Techniques and methods for data mining are found to find the right solution for Heart disease dataset, the extraction of valuable pattern methods and strategies. A scientific analysis in bioinformatics to forecast cardio illness has been performed. UCI has obtained the Cleveland heart disease Site Repository for scientific purposes. The dataset was researched and evaluated to construct an accurate predictive and diagnostic model. The illness of Cardiovascular diseases. Within this analysis we plan to use the resampling technique to increase precision. Next, add and evaluate the results of Logistic Regression, Naïve Bayes, SVM, K Nearest Neighbors (kNN), Decision Trees and Random Forest. We have two sections in the Cleveland heart disease problems in adult patients whether or not optimistic. The dataset was purchased from the server directory of the UCI computer learning. The complete instance consists of 300 and those qualities, including, Blood pressure diastolic (mm Hg), Cholesterol level, obesity, body fat, amount of occasions score, Fasting Blood sugar ,chest pain, age (years). Next Data instances preprocessing is that. PCA used to decrease the dataset size. PCA has reported six attributes using PCA attributes for recognition instruction. Then we have to deal with it. Application of no replacement resample filter that heart disease the replicated details.

The effects of classification are measured by analogy in Conditions and were graded and listed properly. Such steps are carried out except for precision, precise and reminder. When assuming independent predictors is true, a classifier for the Naive Bayes performs better than other models. Naive Bayes uses a limited volume of testing information to forecast the test results. There is also less preparation to prepare. Naive bayes can therefore be quickly introduced. The primary imitation in Naive Bayes is the presumption of the leading indicators being autonomous. Naive Bayes mistakenly assumes that all features are consensually identical. If a primary key has an experimental data array aspect not found in the training phase, otherwise it will be 0 (zero) and would not be difficult to forecast the model would be provided. It is nearly impossible that we would get an extreme predictive model. Also regarded as the zero value. Random forest works fine for a wider array of data items than single decision trees. Random forests are highly versatile and of very high precision. It also preserves consistency even though a large proportion of the data is incomplete. The main drawback of Random Forests is the complexity of these. They are much harder to build than decision trees, and time-consuming to create. It also needs more computing space, and is less intuitive. Compared to other algorithms, the process of forecasting utilizing random forests requires time.

The normal distribution is also simpler to apply, evaluate and educate the results with better performance. The more robust confusion matrix becomes harder to incorporate, test and educate.

The main downside to a statistical analysis is the assumption to linearity between the network monitoring and the endogenous variable. The prediction model is that the analysis is more accurate. Evidence is rarely linearly definable in the material realm. Unless the number of measurements is higher than the number of mechanisms, a regression model can not be used as it may induce overfitting. Any of the specifics are a jumbling chaos. With regression analysis, only isolated capabilities can be expected. The vector based on the discriminant analysis is then restricted to the range of independent numbers. The restriction itself is perilous, since continuous data estimation is onerous. prediction model is the most reliable and effective mode of inference. The simulation's efficiency in selecting the most suitable investment potential including its hyperparameters is increased. In which we split the data set into several folds and train the algorithm through multiple folds. This stops our model from overfitting the data set for the preparation. Logistic regression significantly improves training time. logistic regression is very costly computationally as regards the computing power required. In Logistic regression different scores are used to measure the accuracy, this stands as a major disadvantage in this algorithm.

The decision-making tree requires fewer preparation for pre-processing in order to generate output as other methodologies. A decision tree doesn't need any more runtime standardization and scaling.

Moreover, missing data values have little drastic impact on the decision tree design process. Just one small shift in the statistics can trigger a large change in the decision tree layout, which disenfranchises voters. Often computation might be more hard for a feature selection with regard to those kinds of methodologies.

KNN classifier is referred to as a Lazy Learner (instance-based learning). During the teaching era, little is learned. The training data does not extract any discriminative function. In other terms, there is no time for it. It retains the testing data collection, and then learns from it as observations are produced in real time. It makes the KNN algorithm slightly quicker than those classifiers.

Since the KNN algorithm needs no preparation until predictions are made, new data may be introduced automatically which does not change the algorithm's accuracy. KNN is incredibly simple to incorporate. The specification of KNN needs only two parameters. Doesn't cooperate all that well in massive amounts of data: the cost for estimating the distance within each juncture in larger datasets is huge which deteriorates the performance of the proposed algorithm. Doesn't compete very well with large dimensions: The KNN algorithm does not fit in very well with high-dimensional problems because the algorithm can't demonstrate the separation of substantial numbers of bland or else, KNN can generate incorrect forecasts.

1.3 MOTIVATION

The key goal is to improve precision by resampling on a well-known data collection of heart disease collected from UCI's machine learning library, Cleveland heart Dataset. In future predicting heart disease is more important than curing. The motivation of our main project is predicting Heart(Cardiovascular) disease in patients with an accuracy rate of highest precision. The improved accuracy rate is achieved by using classifier methods in Machine Learning(ML). The JupyterLab tool was used for each and every classification method. The

overview of the data collection, the preprocedure and the classification techniques are included in this portion. The accuracy of the data impacts the forecast to a great degree. A Detailed research was also conducted on the treatment of Cleveland's Heart disease and the findings were collected and inspired. Information preprocessing thus plays a prominent part in the design. Our project also adapts one or more datasets. So our prediction model helps several medical fields and predicting diseases early and precisely.

CHAPTER 2

OBJECTIVE

The Key purpose of this research is to enhance predictive model accuracy. All precision can be improved by boosting the effectiveness of data, algorithms or even tuning algorithms. We improve precision through improvement of the data step of preprocessing that functions really well. Implementation of Imputer technique to the Heart(cardiovascular) disease dataset increases virtually all classifiers' precision, however the Trees of choice prevail over others. The conclusion is also that the reliability of a sample relies significantly on the dataset.

Therefore, Cleveland Heart disease dataset fits really well this strategy but the same conclusions may not be guaranteed for another Information collection of datasets. The core issues we have overcome are to boost the predictive performance design and application adherence of multiple datasets. We assume that our model proposed a high accuracy rate of predictions of racism about the scientific findings by other scholars. To do so it was important to carry out a fair analysis with the findings of others using the same Cleveland Heart disease dataset using the JupyterLab toolkit platform. Suitable filters have been used to increase the expiration and the dataset's pragmatism. The model suggested, comprising all Enhanced classification accurate cluster and grade process guaranteed ratio. The actual and the latest hospital reports need to be introduced. Prolonged preparation and improvement of medical details our plan.

The standard and the data set should be sufficiently large for training and foreseeing. All sophisticated models and algorithms will be available in the Heart disease report. Grading requirements are also provided for patients with potential Cardiac illness. Development of a number of rules and Norms are a legitimate method for preventing the formation of Heart valve's cholesterol blocks. Strong Foundation and much more efficacious Cardio illness and grading potential predicting model. It is presented to patients. It helps to reduce Heart related block, infection and disease development And the chance of experiencing Heart attacks due to instant

blocks is gradually decreased.

CHAPTER 3

EXPERIMENTAL WORK

PROPOSED METHODOLOGY

This is significant that we focus on a high-chance network of individuals managing Heart sicknesses. The innovation of data mining is a proper field of study for us. Build up a prescient model and the objective is to manufacture the models of the estimate utilizing a few calculations of machine investigation. Utilizing the machine is a sort of virtual understanding that makes Machine Learning without unequivocal programming. The strategy for order creates a prescient model progressively solid as it is the most generally utilized procedure in AI analyzes and produces a derivation highlight of the preparation results, which can be utilized for inventive or unnoticeable model planning. The pattern of seeking is applicable for preparing the information patterns and additionally it is significantly distinguished as Knowledge Discovery in Databases (KDD). This incorporates certain stages, for example, information choice, pre-handling, Processing, characterization and is characterized as the way toward registering.

Discover relationships in wide databases that remember approaches for the man-made brainpower assembly, AI, figures and Systems in the site. The key focuses of these methodologies are patterns Recognition, estimation, mergers and groups. Mining Data contains an assortment of steps that are discarded consequently or semi-naturally for Extract and reveal intriguing, obscure, shrouded highlights Data Quantities. High information proficiency and the effectively applied two basic aspects of data mining are methods. Data mining has been applied broadly to different tangible fields Society, for example, climate estimates, business research, specialized administration, and purchaser cooperations. In any case, the code there is consistently space for sickness avoidance and clinical information preparing Enhancements. As we all know, there are huge quantities of cardio sickness casualties, and this is nonstop growing-up.

Steps involved in this Process :

- Selection of Dataset
- JupyterLab toolkit (Anaconda Navigator)
- Preprocessing of Cleveland Heart Disease dataset •

Splitting of data into trained and test data

- Classifiers used for classification process i.e.

Random Forest,

K-Nearest Neighbour

- Achieving Accuracy

Test Accuracy,

Train Accuracy,

Confusion Matrix Accuracy.

Fig 3.1. Workflow of a Prediction model

3.2 DATASET SELECTION:

The major features of the person having Heart disease will suffer from obesity, narrowing and blocking of blood vessels, pain in the regions of chest and left hand and deposition of cholesterol in the inner layers of the vessels carrying blood that leads to insufficiency in the level of blood flow. Depending on the diagnostic methods, we can predict the onset of Heart disease. This was basically from Cleveland Heart disease dataset. This was designated to assume whether a patient has Heart (cardiovascular) disease or not. These are the characteristics to describe this from most populations. Our study was based on characteristic information from 303 records. From which, the detailed database with attributes includes blood pressure(diastolic), cholesterol level, etc.

ATTRIBUTE DETAILS:

- Age - > age of patient
- Sex - > 1-male,2-female
- cp - > chest pain type,1:Typical angina, 2:Atypical angina, 3:Non-anginal pain, 4:Asymptomatic
- trestbps - > Resting blood pressure
- chol - > Serum cholesterol in mg/dl
- fbs - > fasting blood sugar > 120 mg/dl
- restecg - > Resting electrocardiographic results (values 0,1,2)
- thalach - > maximum heart rate achieved
- exang - > Exercise induced angina
- oldpeak - > oldpeak = ST depression induced by exercise relative to rest ➤ slope - > the slope of the peak exercise ST segment
- ca - > number of major vessels(0-3) hued by fluoroscopy
- thal - > thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
- target - > 0 - suffered; 1 - not suffered

3.3 JUPYTERLAB TOOL:

Jupyter Lab is used in our study. It provides various accesses from data connection and manipulates the data. It allows and creates, modifies and runs the data experiment at every instance. Anaconda Navigator is a GUI in Anaconda distribution and allows various platform software and control Anaconda products, environments and channels quickly without using command line commands. Advantage of this software is free open source and friendly graphical interface. It gives better results compared to other tools.

3.4 DATA PREPROCESSING:

In the real world mostly the dataset contains redundant and some missing values. Our dataset contains some duplicate, missing values and irrelevant attributes. Preprocessing technique helps in less timing consumption during classifier implementation. So we want to improve our data before applying a classifier. Data preprocessing is one of the major techniques to clean and

improve our datasets. It is otherwise called an elimination technique or noise reducing technique. Our dataset contains some duplicate, missing value and irrelevant attributes.

After Preprocessing technique our data gives higher accuracy than before. In our study we have some optimization techniques such as transfer of numerical value into nominal value, Outcome of data indicates numerical value like True=1 and False=0 which means true(1) experiences Heart ailment and false(0) not experience any coronary illness. It will reduce the complexity of the process. For this process we will use Imputer preprocessing technique and fill to fit transform technique. Total Number of True(1) and False(0) values have been counted for better understanding of the dataset.

3.5 DATA SPLITTING :

This method is used for finding inaccurate prediction values. Splitting data is achieved by **train_test_split** function under sklearn import selection mode. It is going to split the existing dataset into two parts. The ratio of the two parts is flexible but the ideal split is **80:20** . So we used this ratio for separation. 80% is considered as Trained dataset and 20% considered as a Test dataset. Here random_state is used for random two sub parts. Data size and parameters should be set by us.

3.6 CLASSIFICATION MODEL:

The classifier is selected based on their performance and uses in other researcher's studies. Random forest and Logistic Regression are the best suit model for the heart disease dataset. Classifier is a one of the tools in data mining and it is used to classify and forecast the data. The main objective of the classifier is increasing accuracy by using algorithms like Logistic regression algorithm, Gaussian Naive Bayes algorithm, Support vector machine, K Neighbors algorithm, Decision Trees and Random forest.

3.6.1 RANDOM FOREST ALGORITHM

Random forests are highly versatile and of very high precision. Random forests is an ensemble methodology of learning for classification of many tasks. It creates a valuable thing of decision making trees from an arbitrarily choosing substring of datasets which is trained. It then picks the votes from the non identical Decision trees to make a correct decision for the tested object's final class.

It is a half eliminator that lays a number of Decision tree classifiers on non identical sub-samples of the dataset and uses average to upgrade the accuracy of the prediction made and over-fitting control. Random forest works fine for a wider array of data items than single Decision trees.

The main drawback of Random Forest is the complexity which occurs while creating a lot of trees. They are much harder to build than decision trees, and time-consuming to create. It also needs more computing space, and is less intuitive. The method of predicting using random forests is consuming a lot of time compared with other algorithms.

3.6.2 K-NEAREST NEIGHBORS

K-NN gives higher accuracy and a versatile model. It needs a huge memory area to store data values. **No Training Period** : K-NN is called Lazy Learner (Learning based on Instance). During the teaching era, little is learned. It is the simplest model compared to others. It is mainly used for data points injected into several classes and to predict accuracy rate. It is called a non-parametric method. This method is opposed to eager technique. K-NN applications are credit rating, bank related fields, etc. It is not important to learn the other methods. The training data does not extract any discriminative function. In other terms, there is no time for it. It retains the testing data collection, and then learns from it as observations which are produced in real time. It makes the KNN algorithm slightly quicker.

Since the KNN algorithm needs no preparation until predictions are made, new data may be introduced automatically which does not change the algorithm's accuracy. KNN is incredibly simple to incorporate. The specification of KNN needs only two parameters .

It will not properly fit in larger data values. Finding distance between new and existing values in a bigger dataset is high in cost and it reduces the overall performance. The KNN algorithm doesn't fit well with high dimensional data as it is impossible for the method to measure in dimension distance for a vast number of dimensions. feature scaling is needed: Before applying K-NN method in every data value, we will do scaling features between normalization values and standardization values. If we do not, KNN could generate false predictions.

3.7 ACHIEVING ACCURACY:

In our project accuracy is determined by inbuilt function. Accuracy is defined as a measure of the number of correct decisions divided by the number of test examples. So it is easy to find the accuracy of our classifier. Here, we want to find three accuracies like 1)Test accuracy 2)Train Accuracy 3)Confusion Matrix. Achieving high accuracy is our main motive.

It is known as the number of perfect patterns that are classified to the total patterns and it is the ratio of added the true +ve values and true -ve values to the total trials. Incase of accuracy, we apply equal cost to false +ve values and false -ve values. Because of the imbalanced dataset it has (99%)instances in one class and(1 %)in others.

3.7.1 TEST ACCURACY:

Test accuracy is an accuracy achieved by a test dataset that is 20% test data. If we find test accuracy then we want to compare this accuracy in trained accuracy. Test accuracy is used for testing our dataset and improving data.

3.7.2 TRAIN ACCURACY:

Train accuracy is an accuracy achieved by a trained dataset that is 80% trained data. If we find train accuracy and it acts as an actual dataset. In this dataset we want to find the accuracy in the test data.

3.7.3 CONFUSION MATRIX:

The uncertainty matrix provides different paths in which the concept of classification is uncertain as the predictions are made. This provides us a glimpse of the classifier errors and its types.

Here,

Class 1 : +ve

Class 2 : -ve

Definition Terms:

+ve (P) : Actual state is +ve

-ve(N) : Actual state is -ve

11

True +ve (TP): Actual state is +ve, and is predicted to be +ve.

False -ve (FN) : Actual state is +ve, but is predicted -ve.

True -ve (TN) : Actual state is -ve, and is predicted to be -ve.

False +ve (FP) : Actual state is -ve, but is predicted +ve.

TRUE POSITIVE	FALSE POSITIVE
FALSE NEGATIVE	TRUE NEGATIVE

RECALL

Recall implies to the total percentage similar values perfectly divided by the proposed theory. For most of the problems, a trade-off is the first need, and a decision should be made to maximize precision or recall. Recall is the measure of those qualities based on mistakes done.

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

PRECISION

Precision is the count of true positive. Precision is the calculation of the quality of the predictions based on the positive prediction and implies to the percentage of the results that are similar.

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

F-MEASURE

The F score is a calculation of the accuracy of tests and is known as harmonic mean weightage of test's precision accuracy and recall values. It gives a unique score that combines accuracy in a single number as well as recall .

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

12

CLASSIFICATION ACCURACY

It is known as the number of perfect patterns that are classified to the total patterns, and it is the ratio of addition of true positive values and true negative values to the total trials. In case of accuracy, we apply equal cost to false positive values and false negative values. In case of an imbalanced dataset it has (99%) instances in one class and (1 %) in another.

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

CHAPTER 4

RESULTS AND DISCUSSION

4.1 DATASET

DATA : Cleveland Heart Disease Dataset

No. of rows & columns : 303, 14

4.2 CLASSIFICATION ALGORITHMS

A classification algorithm is a function which evaluates the input features so that the output separates one class into positive values and the other into negative values.

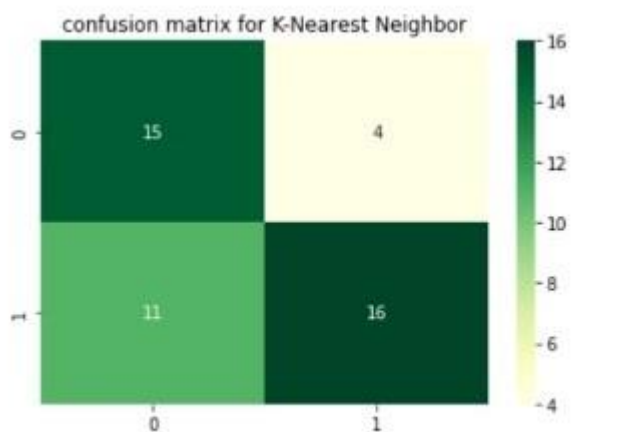
4.2.1 LIST OF ALGORITHMS

These are the algorithms/classifiers which we used in our prediction model. We use Six classifiers to classify the Heart(Cardiovascular) disease dataset which gives efficient accuracy of heart diseases earlier to the person.

- ❖ KNN algorithm
- ❖ Random Forest algorithm

13

4.2.1.4 K-NEAREST NEIGHBORS:



The accuracy score achieved using KNN is: 67.39 %

Fig4.4. Confusion Matrix of K-Nearest Neighbors

Finally we achieve **67.21%** accuracy for K-Nearest Neighbors. From the above table we obtained the values of **True Positive**, **False Positive**, **True Negative**, and **False Negative**. Thus the values of precision, recall, support, f1 score and the accuracy of K-Nearest Neighbors algorithm is noted for future reference.

4.2.1.6 RANDOM FOREST :

//To print the confusion matrix of Random Forest Algorithm

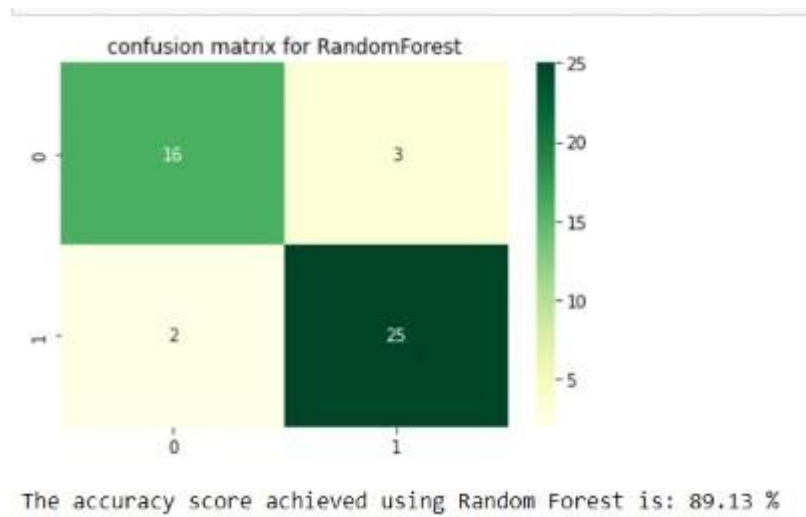


Fig 4.6. Confusion Matrix of Random Forest

Finally we achieve **88.52%** accuracy for the Random Forest algorithm. From the above table we obtained the values of **True Positive, False Positive, True Negative, and False Negative**. Thus the values of precision, recall, support, f1 score and the accuracy of the Random Forest algorithm is noted for future reference.

14

4.3 CLASSIFICATION TABLE:

CLASSIFIER	TP	FP	FN	TN	ACCURACY	PRECISION	RECALL
KNN	23	11	09	18	0.67	0.72	0.68
Random forests	30	04	03	24	0.89	0.91	0.88

Table 4.7. Classification table of prediction model

4.4 MEAN ACCURACY

//To get the mean accuracy values of each classifier

Here we have the mean accuracy values of the six classifiers. The **cross_val_score** appraisals the anticipated accuracy of the model and reduces the overfitting by train the algorithms on different folds. This results the mean accuracy for K-Nearest Neighbors is 73.333% and Random Forest is 73.939%. The variance of the cross_val_score is also calculated in this prediction model.

4.5 ACCURACY OBTAINED

Finally, this prediction model shows that the Random **Forest** algorithm is the best classifier for the **Cleveland Heart Disease** dataset compared to other classifiers. By these results, we also find the mean value of the accuracy score and the value is 81.696%.

15

PROPOSED METHOD	ACCURACY ACHIEVED
KNN	67.21%
Random Forest	88.52%

Table 4.8. List of accuracies achieved by the model
CHAPTER 5

5.1 CONCLUSION AND FUTURE PLANS

This is a prediction model which involves diagnosing and prognosing the Heart disease. We applied this method on six classification algorithms K-Nearest Neighbors, Random Forest. Our dataset is Cleveland Heart disease Dataset. Python code is implemented in the JupyterLab platform under Anaconda Navigator. The main aim of the paper is to define a highly accurate model for the prediction of Heart Disease. For this first we had done the Dataset preprocessing which includes removing the unwanted data and imputer (finding mean value for missing data). By this preprocessing, we can enhance the accuracy of the data.

After preprocessing, we went with splitting the dataset into two parts namely trained dataset and test dataset. Next we apply those six classifiers over with two datasets and then we find “Test Accuracy”, “Trained Accuracy”, “Confusion Matrix”, “Mean Accuracy”. According to this model, the highest accuracy is found to be given by “Random Forest” about 88.52%. This algorithm is found to have less “false positive value” compared to other methods. Hence this classifier is called OPTIMISTIC solution. And the least accuracy is given by “K-Nearest Neighbors” about 67.21% which is considered to be a PESSIMISTIC solution.

In future works, we can apply advanced classifiers such as neural networks over this dataset. We planned to extend this prediction model towards a large number of datasets to obtain the different accuracy values. We can further optimize this dataset by adding new hospital patient details. From this prediction model, we can provide some advanced treatments and regulations to the patients even before they suffer a lot.

REFERENCE

- [1] **R. El-Bialy, M.A. Salamay , O.H. Karam, M.E. Khalifa** "Feature analysis of coronary artery heart disease data sets", *Procedia Comput. Sci* **2019; 65 : 459-468**
- [2] **Ujma Ansari JyotiSoni, Dipesh Sharma** "Predictive data mining for medical diagnosis: an overview of heart disease prediction", *Int. J. Comput. Appl* **2018;17 :(0975 – 8887)**
- [3] **K. Thenmozhi, P. Deepika** "Heart disease prediction using classification with different decision tree techniques", *Int J Eng Res Gen Sci* , **2019;2:6.**
- [4] **Mohammed Shafennor Amin, et al.** " Identification of Significant features and data mining techniques in predicting heart disease", *Telematics Inf* **2019; 82-93.**
- [5] **K. Sudhakar** "Study of heart disease prediction using data mining" **2020; 1157-1160.**
- [6] **Jagwant Singh, Rajinder Kaur** "Cardiovascular disease classification ensemble optimization using genetic algorithm and neural network" *Indian J. Sci. Technol* . **2020; 9.**
- [7] **K. Sudhakar and M. Manimekalai** , "Study of Heart Disease Prediction using Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering* ,**2019;4:1157-60.**
- [8] **G. Purusothaman and P. Krishnakumari** , "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", *Indian Journal of Science and Technology* , **2019;8:12.**

[9] **S. Chaitrali, Dangare and S. Apte Sulabha**, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications* (0975-888) , **2019; 47: 44-48** .

[10] **Beant Kaur and H. Williamjeet Singh** , "Review on Heart Disease Prediction System using Data Mining Techniques", *International Journal on Recent and Innovation Trends in Computing and Communication* , **2019;2:3003-08**.

17

[11] **C. Ordonez** , "Association rule discovery with the train and test approach for heart disease prediction", *IEEE transactions on Information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society* , **2018;10:334-43**.

[12] **R. Chitra and V. Seenivasagam** , "Review of heart disease prediction system using data mining and hybrid intelligent techniques", *ICTACT JOURNAL ON SOFT COMPUTING* , **2019;3:605-09**.

[13] **Beant Kaur and H. Williamjeet Singh** , "Review on Heart Disease Prediction System using Data Mining Techniques", *International Journal on Recent and Innovation Trends in Computing and Communication* , **2020;2:3003-08**.

[14] **L. Verma, S. Srivastava, P.C. Negi** "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data " , *J Med Syst* , **2020;40:1-7**.

[15] **Niti Guru, Anil Dahiya, NavinRajpal** "Decision support system for heart disease diagnosis using neural network, *Delhi Business Review*", **2018;1:8**.

