Heart Disease Predicting Using Machine Learning Algorithms and Data Mining Technique

[1]Aman Jain, [2]Simran Sharma
[1,2]Jaypee Institute of Information Technology, Noida

ABSTRACT

Health Care Clinical diagnosis usually ends with the knowledge and practice of the physician. The computer-assisted decision support system plays a major role in the medical field. Provides the extraction of methodological data and technology to change these hills of data into useful information for decision-making. Using data extraction techniques, it takes less time to predict the disease more accurately. Researchers use many data mining techniques to help health professionals diagnose heart disease. But the use of data mining technology can reduce the number of tests required. In order to reduce the number of deaths due to heart disease, a rapid and effective detection technique is essential. The decision tree is one of the effective methods for extracting the data used. This research compares different algorithms. The algorithms tested are the Decision Tree algorithm, the Naive Bays algorithm, the support vector machine and the Random Forest algorithm. Current data sets are being used for heart patients in the Cleveland UC database to test and test the performance of tree resolution algorithms. These datasets include 303 observations and 76 attributes. Later, a classification algorithm offering the ideal potential for use in large data will be proposed. The purpose of this study is to extract hidden patterns by applying significant data extraction techniques for heart disease and predicting the existence of cardiopathy in patients for whom this presence was assessed by the absence of probable presence.

Keywords: Data mining technique, machine learning Algorithm, Decision Support System, Health care, Heart Disease

I. INTRODUCTION

To extract hidden models and relationships from large databases, data mining integrates statistical analysis, automated learning, and database technology. Heart disease refers to various diseases affecting the heart, such as chest pain, shortness of breath, heart attacks and other symptoms. It includes various diseases affecting the heart. Chest pain occurs when the blood received by the heart muscle is insufficient. Heart disease refers to many problems that block the heart and blood vessels in the heart. [2] The term "cardiovascular disease", which represents a category of heart disease, includes various conditions that disrupt the heart and blood vessels and the manner in which blood is pumped and circulates in the body. Heart disease is the leading cause of death in the world in the last decade. The World Health Organization (WHO) has reported that heart disease is the leading cause of death in high- and low-income countries.

Medical diagnosis plays a vital role and a complex task that needs to be implemented effectively and accurately. To reduce the cost of clinical testing, computerized information and decision support must be supported. Data mining involves using software techniques to find patterns and consistency in data sets. In addition, with the emergence of data mining over the past two decades, computers have the ability to directly create and categorize different attributes or categories. Learning the components of heart disease risk helps medical experts identify patients at risk for heart disease. Statistical analysis identified risk factors for heart disease related to age, blood pressure, total cholesterol, diabetes, excessive blood pressure, family history of heart disease, obesity, lack of exercise and fasting glucose, etc [3].

The heart is one of the main organs of the human body. It pumps blood into the blood vessels of the circulatory system. The circulatory system is very important because it transfers blood, oxygen and other substances to the various organs of the body. The heart plays the most important role in the circulation. If the heart does not work properly, it will cause serious health problems, including death.[4].

III. DATA MINING TECHNOLOGY

Data mining technology is useful for extracting non-trivial information from medical databases. It is an intelligent computer analysis of large data sets using a combination of machine learning, statistical analysis and database technology, in order to discover useful models and rules to guide decisions. future activities. The purpose of data mining is to predict and generalize the model of other data. The extraction of medical data has become increasingly important in the field of health care. Data mining is a powerful technology that can help organizations focus on the most important information in their repositories. Data mining tools help predict future trends and behaviours and help organizations make proactive decisions based on knowledge. There are many techniques for extracting data whose relevance depends on the field of application. Health data mining applications can have tremendous potential and benefits. It automates the process of finding predictive information in large databases. The mining data mining technology includes two models, such as the taxonomy model and the evaluation model.

IV. HEART DISEASE

Heart disease or cardiovascular disease (CVD) is a class of diseases that include the heart and blood vessels. Cardiovascular diseases include coronary heart disease such as angina pectoris and myocardial infarction (commonly known as heart attack). Another heart disease is called coronary heart disease (CHD), where a waxy substance called plaque develops in the coronary arteries. It is the arteries that supply the heart muscle with oxygen-rich blood. When plaque starts to accumulate in these arteries, it is called atherosclerosis. The development of plaque occurs over several years. Over time, this paint may harden or break (open fracture). Plasma hardening eventually narrows the coronary arteries, reducing oxygen-rich blood flow to the heart. If this plaque is torn, it can form a blood clot on the surface. A large blood clot can block blood flow into the coronary artery. Over time, the plaque increases and the coronary arteries become narrower. If the stopped blood flow is not restored quickly, the part of the heart muscle begins to die. Without prompt treatment, a heart attack can lead to serious health problems or even death. Heart attack is a common cause of death worldwide [4].

The heart attack occurs when the arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed.

Various types of heart diseases are:

1) Coronary heart disease
2) Cardiomyopathy
3) Cardiovascular disease
4) Ischemic heart disease
5) Heart failure
6) Hypertensive heart disease
7) Inflammatory heart disease
8) Valvular heart disease

❖ **Common risk factors of heart disease include**

1) High blood pressure
2) Abnormal blood lipids
3) Use of tobacco
4) Obesity
5) Physical inactivity
6) Diabetes
7) Age
8) Gender
9) Family generation

V. DATASET DESCRIPTION

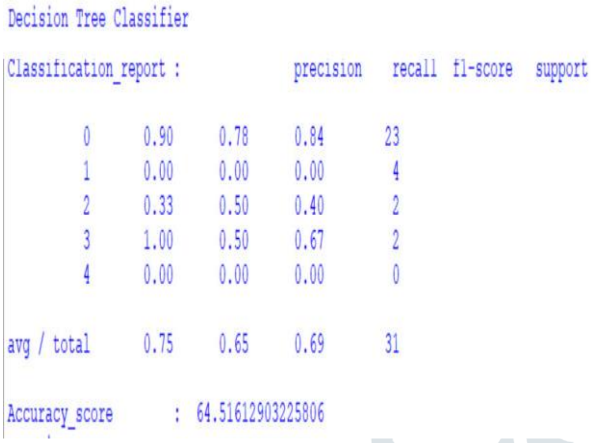
We simulated the computer on a single data set. A dataset is a set of data from the heart. The dataset is available in the UCI learning repository. The dataset contains 303 samples, 14 input functions and one output function. Describe the financial, personal and social characteristics of loan applicants. The exit function is a decision class with a value of 1 for good credit and 2 for bad credit. The data configuration contains 700 cases with good credit, while 300 cases constitute bad credit. The dataset contains characteristics expressed in nominal, ordinal or interval metrics. A list of all these features is given in the table.

Feature No.	Feature Name	Type	Decription
1	Age	Continuous	Age in years
2	Sex	Discrete	Aex (1 = male; 0 = female)
3	cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypicalangina, 3 = non-anginal pain, 4 =asymptom
4	Trestbps		Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol		serum cholestoral in mg/dl
6	Fbs	Discrete	(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	Discrete	Resting electrocardiographic results,0= normal,1=aving ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 V),2= showingprobable or definite left entricularhypertrophyby Estes' criteria
8	Thalach		Maximum heart rate achieved
9	Exang		Exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	Discrete	ST depression induced by exercise relative to rest
11	Slop		the slope of the peak exercise ST segment,1: upsloping,2: flat,3: downsloping
12	Ca	Discrete	Number of major vessels (0-3) coloured byflourosopy
13	Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
14	Num		Diagnosis of heart disease (angiographic disease status),0: < 50% diameter narrowing,1: > 50% diameter narrowing

VI. ALGORITHM USED

1) **Decision Tree**: The decision tree divides the input space of the data set into mutually exclusive areas,

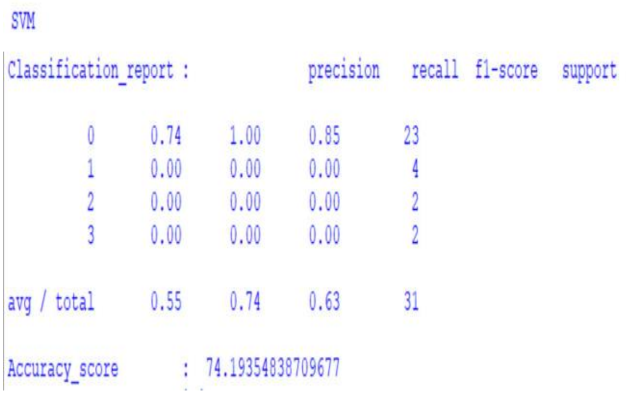
where a tag, a value, or an action is assigned to distinguish their respective data points. The mechanism of the tree is transparent and we can follow its structure easily to learn how to make a decision. A decision tree is a tree structure composed of internal and external nodes connected by branches. The internal node is the decision-making unit that evaluates the decision function to determine the child's node that will visit it later. In contrast, the outer node has no child contract and is associated with a classification or value.



```
Decision Tree Classifier

Classification_report :          precision    recall  f1-score   support

          0       0.90      0.78      0.84        23
          1       0.00      0.00      0.00         4
          2       0.33      0.50      0.40         2
          3       1.00      0.50      0.67         2
          4       0.00      0.00      0.00         0

avg / total       0.75      0.65      0.69        31

Accuracy_score      :  64.51612903225806
```

Figure 1 : Decision Tree Classifier Result

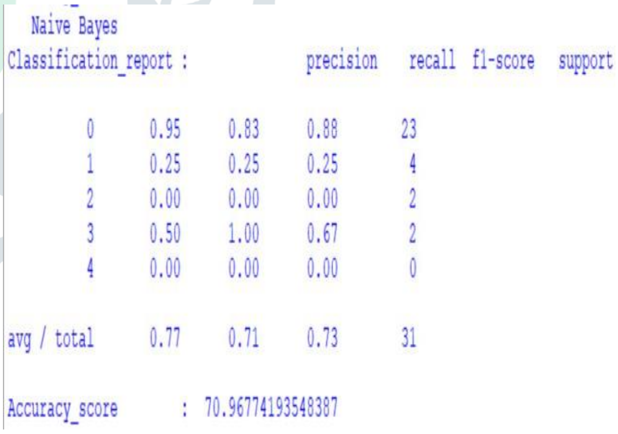
2) **Support Vector Machine**: Support vector machines come in different forms, both linear and non-linear. The vector support machine is a supervised moderator. What is usual in this context, two different data sets are included with SVM, Training and Test Group. Ideally, the layers are detachable in writing. In such a case, a line can be found, which completely divides the two layers. However, a single line is not limited to dividing the dataset accurately, but a whole set of lines does. From these lines, the best is chosen as "chapter line". The best line is found by maximizing the distance to the closest points of the two categories of the training group. Maximizing this distance can be converted into a problem that is minimizing, easy to solve. Data points on the maximum margin lines are called media carriers. In most cases, data sets are not well distributed, so layers can be separated by a higher order line or function. Real data sets contain random or noise errors that produce a less clean data set. Although it is possible to create a model that ideally separates the data, this is not desirable because these models are highly compatible with the learning data. Installation occurs more by combining random errors or noise in the model. Thus, the model is not generic and greatly increases errors in other datasets.



```
SVM

Classification_report :          precision    recall  f1-score   support

          0       0.74      1.00      0.85        23
          1       0.00      0.00      0.00         4
          2       0.00      0.00      0.00         2
          3       0.00      0.00      0.00         2

avg / total       0.55      0.74      0.63        31

Accuracy_score      :  74.19354838709677
```

Figure 2 : SVM Result

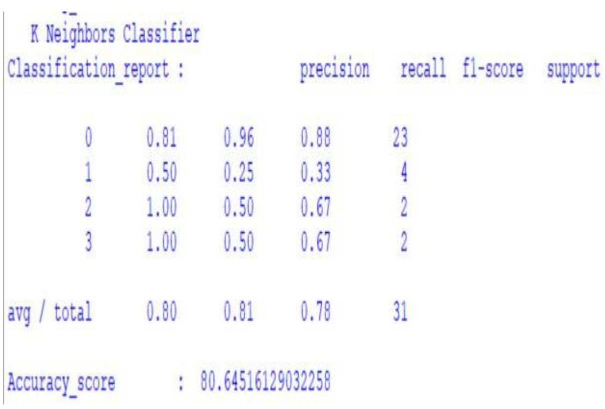
3) **Naive base classifier: -** This work is a strong representation of probability and its use in classification has received a lot of attention. This workbook draws conditional learning data for each attribute Ai with the label C. The classification is then performed by applying the Bays rule to the calculation of the probability C according to the cases A1 and predicting then the category with the highest ascending probability. The objective of the classification is to correctly predict the value of a distinct discrete variable according to the predictor or attribute vector. Naive Bays is in particular a Bayesian network where parents have no class and each class has its own mother. Although the Bayesian Naïve (NB) algorithm is simple, it is very efficient in many real-world data sets because it can offer better predictive accuracy than well-known methods such as C4.5 and BP.



```
Naive Bayes
Classification_report :          precision    recall  f1-score   support

          0       0.95      0.83      0.88        23
          1       0.25      0.25      0.25         4
          2       0.00      0.00      0.00         2
          3       0.50      1.00      0.67         2
          4       0.00      0.00      0.00         0

avg / total       0.77      0.71      0.73        31

Accuracy_score      :  70.96774193548387
```

Figure 3 : Naive Bays Result

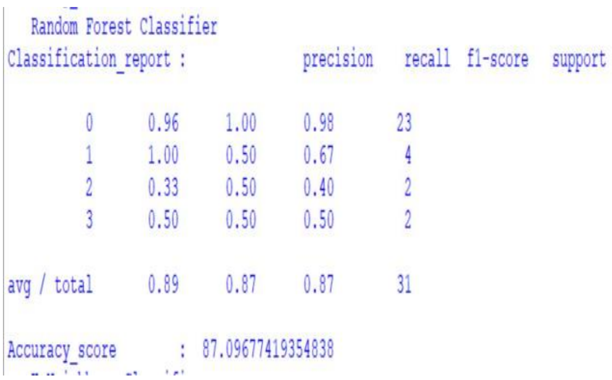
4) **K-Nearest Neighbour**: This workbook is a statistical learning algorithm that is very easy to apply and open to various differences. In short, the training part of the nearest neighbours is little more than storing the data points that are provided to it. When asked to predict an unknown point, the neighbouring workbook finds the closest training point to an unknown point and predicts that training point category based on certain distance measurements. To measure the distance used in the nearest neighbourhood of numerical entities is the simple Euclidean distance.



```
 K Neighbors Classifier
Classification_report :            precision    recall  f1-score   support

          0       0.81      0.96      0.88        23
          1       0.50      0.25      0.33         4
          2       1.00      0.50      0.67         2
          3       1.00      0.50      0.67         2

avg / total       0.80      0.81      0.78        31

Accuracy_score       :  80.64516129032258
```

Figure 4 : K Neighbors Classifier Result

5) **Random Forest Classifier:** Random Forest is a classified binder with many decision trees. Class outings are represented by individual trees. It is derived from a random forest resolution proposed by Bell Labs Tin Cam in 1995. This method combines with the random selection of features to construct resolution trees with controlled variations. The tree is created using an algorithm as discussed. Random Forest (RF) is one of the cases of such actions. RF is a multi-tree form of selection trees where each ht tree is generated from the set of information and takes a number of random numbers that are distributed non-vectorially and without vector. The vectors 1, theta 2, ..., theta t-1 were used to create classifiers h1; H2. ::. Ht-1. Each resolution tree is made from a random subset of the preparation dataset. Random vectors were produced from a certain variable probability diffusion, where the probability rotation is transferred to central samples that are difficult to organize. The random vector can be linked to the tree process of different perspectives. The paper axes are designated for each tree by evaluating the backward accounting of information class names. Each internal axis contains a test of the best parts of the data area to be controlled. Another hidden order is organized by sending it to each tree and aggregating it into the credit sheets. [1]



```
Random Forest Classifier
Classification_report :            precision    recall  f1-score   support

          0       0.96      1.00      0.98        23
          1       1.00      0.50      0.67         4
          2       0.33      0.50      0.40         2
          3       0.50      0.50      0.50         2

avg / total       0.89      0.87      0.87        31

Accuracy_score       :  87.09677419354838
```

Figure 5 : Random Forest Classifier Result

VI. PERFORMANCE COMPARISONS

Table 1: Accuracy Comparison Table

Algorithm classification	Accuracy Score	Inaccuracy Score
Decsion Tree	64.51	35.49
Naïve Bayes	70.96	29.04
Random Forest	87.09	12.91
KNN	80.64	19.36
SVM	74.19	25.81

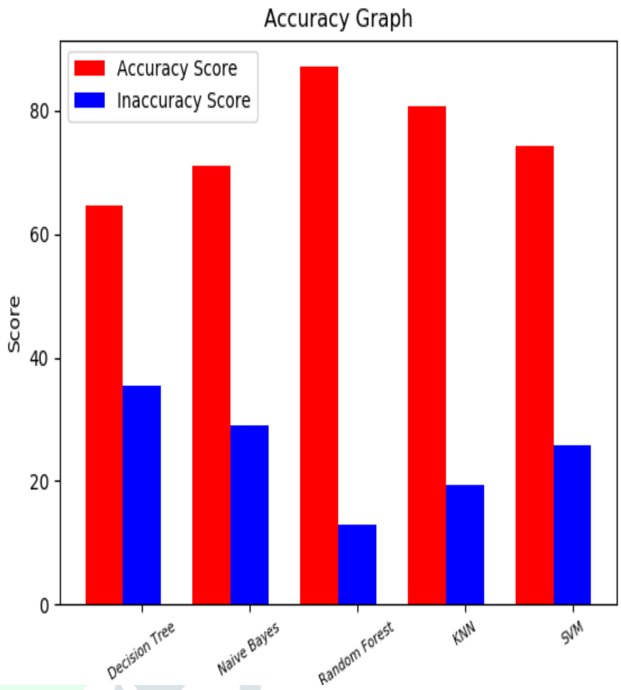


Figure 6 : Accuracy Graph

VII. CONCLUSIONS

The application of data mining to the analysis of medical data is a good way to examine the relationships between variables. From our proposed approach, we have shown that mining can recover the useful link of qualities that are not direct indicators of the category we are trying to predict. In our work, we have tried to predict the risks of heart disease using diagnostic functions of diabetes and we have shown that it is possible to diagnose the weakness of heart disease in diabetics with reasonable accuracy. Exercise books of this type can help early detection of diabetic impairment of heart disease. There by the patients can be warned to change their way of life. This will prevent diabetics from developing heart disease, with low mortality, as well as lower health costs in the state.

REFERENCES

[1]. Jaymin Patel, Prof. TejalUpadhyay and Dr. Samir Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique”, IJCSC, Volume-7, N0.-1, pp-129-137, Sep 2015 – March 2016.

[2]. G. Parthiban and S.K.Srivatsa, “Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients”, International Journal of Applied

Information Systems (IJAIS), Volume 3– No.7, August 2012.

[3]. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 3, Issue 3, pp- 262-269, 2018.

[4]. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”, Advances in Computational Sciences and Technology, Volume 10, Number 7, pp. 2137-2159, 2017.

[5]. Sanjay Kumar Sen, “Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms”, International Journal Of Engineering And Computer Science, Volume 6, Issue 6, pp- 21623-21631, June 2017.

[6]. Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 5, Issue: 8, pp- 99-104, 2017.