HEART DISEASE DATASET COLLECTION CLASSIFICATION AND PREDICTION

SIRIKONDA ANANTHNAG, MANGALI ANIL KUMAR, VIJAYA BHASKAR MADGULA,
ASSISTANT PROFESSOR [1,2,3]

ananthnagsvit9f@gmail.com, anilkumarcse02@gmail.com, vijaya.bhaskar2010@gmail.com

Dept. of CSE, Sri Venkateswara Institute of Technology, N.H 44, Hampapuram,
Rapthadu, Anantapuramu, Andhra Pradesh 515722

Abstract

These days, one of the most pressing problems in the medical field is the prognosis of cardiovascular disease. Because there is a certain amount of individuals who die from heart attacks every minute. Doctors having access to mountains of patient data have a tough time predicting the onset of cardiovascular illness. We need to use the automated hearing illness prediction system to alert the patient and achieve recovery from the condition in order to overcome this complication. Using machine learning methods, we can simply do heart disease prediction with vast data and achieve an autonomous system. Unsupervised learning classifiers and supervised learning classifiers are two examples of the many subsets of machine learning. Methods for making predictions from unstructured data that are part of the unsupervised learning framework. However, these classifiers are best implemented using supervised learning algorithms that operate with structured data. Consequently, supervised machine learning methods such KNN, RF, NN, DT, NB, and SVM classifiers are used in this system.This system is using a training dataset obtained from the UCI machine learning library in order to make predictions about heart disease.In addition, the system displays the accuracy results graphically and compares the performance of different machine learning methods.

Keywords: - Dataset Collection, Classification, Prediction

1. INTRODUCTION

Day by data The healthcare business is seeing a significant increase in the volume of health records. Consequently, it is advised to handle massive amounts of data and transform them into valuable insights for making favourable decisions. Because of this issue, the healthcare sector is looking to adopt an automated system method that can make useful decisions from large datasets. This means that problems like these can be effectively addressed using machine learning approaches. Reason being, it may provide practical ways to extract important data without having to go through an enormous database. Important data may be collected from a variety of patients in the medical business. symptoms and medical records for review by doctors. Many individuals are experiencing signs of heart failure at this point of their lives. But when we look at the generations side by side, we see that the elderly are the ones dealing with these kinds of issues. Nevertheless, ML methods may predict cardiac disease status from training datasets by discovering connections between various parameters. It can identify people with heart problems without the assistance of doctors by employing these training

models. The system may then act as if it were an automated system that can distinguish between patients with positive and negative cardiac illness.patients with precision, which in turn decreases the time and money needed for diagnosis and therapy. Providing high-quality services and making accurate diagnostic status predictions is a major problem in the health care industry. Even when the condition was adequately monitored and controlled by an automated system, a large number of individuals still died from heart disease, according to the poll. Here, the timely identification of any illness is contingent on other factors. Thus, the suggested approach may foretell the heat sickness condition in advance, alert patients, and aid in their recovery. Medical professionals create mountains of medical records in order to study and extract valuable information from that repository. Predicting cardiac illness from the health care database is a laborious process since most of the data is unconnected. In the medical field, for example, illness prediction might benefit from the use of machine learning algorithms, which can decipher structured data. So, in order to prevent patients from suffering from harsh severity, this method suggested an automated technique to doctors for early detection of heart illness. Therefore, supervised classifiers using machine learning methods play a crucial role in the early diagnosis of cardiac illness.

2. RELATED WORK

One of the most challenging tasks in the field of health care is the identification or prognosis of cardiac disease. Some indicators, such as smoking, a high-fat diet, excessive alcohol consumption, etc., help doctors discover heart disease. But doctors can tell whether a patient has an illness based on these symptoms.right on. Patients run the risk of experiencing negative consequences since physicians are unable to treat them in the early stages. Therefore, we need to create a technology that can identify any illness in its early stages so that doctors can treat patients to avert adverse outcomes if we want to detect or forecast cardiac disease. In this case, supervised machine learning methods may be used to develop a tool for predicting the occurrence of cardiac disease. Massive amounts of unstructured medical records are produced by many clinical or hospital settings. Therefore, it is easy to gather relevant data for the training dataset with the help of this prediction tool. This training dataset will be sent into the machine learning algorithms, which will then use the current patient information as the testing dataset to forecast the condition of the heart disease. Here, we provide a method for predicting the occurrence of cardiac problems using six different machine learning classification algorithms, and we compare their respective levels of accuracy. The primary objective of this approach is to properly determine the onset of cardiac symptoms in patients. In terms of implementation, we construct the train model file using a heart disease training dataset and machine learning classifiers. Then, doctors may input the values obtained from patients' medical records into the training model to make predictions about heart illness. In this case, the medical department specialists at UCI have submitted the heart disease training dataset to the repository. We used the Python programming language, which provides access to machine learning via pre-trained modules and packages, to construct our system. category detectors. You can find all six algorithms that forecast heart disease with the highest accuracy here.

3. IMPLEMENTATION

System Model

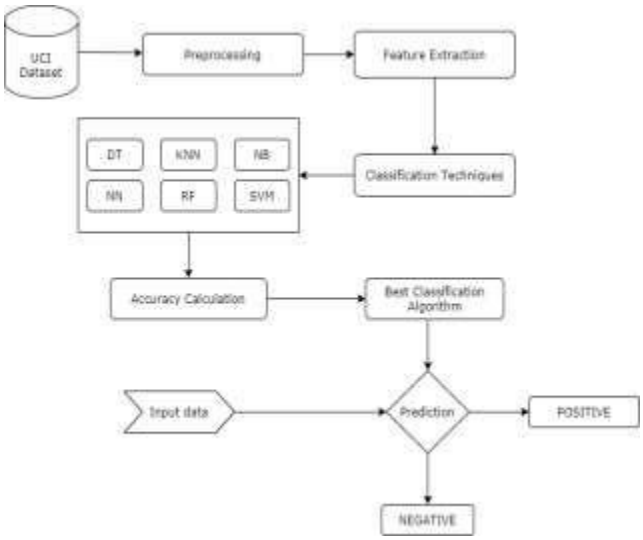

Figure.1 System Architecture

Our suggested system model is shown in figure.1. The heart disease training dataset was retrieved from the UCI library and used in this system model. After some preliminary processing, it can read the training dataset, extract features to separate independent and dependent attributes, construct a training model using a classification algorithm to predict the occurrence of heart disease from input data, and then compare the performance of six ML classifiers.

Dataset Collection

In this system we are using UCI heart disease dataset shown in figure.2 which is accessing from Kaggle web repository (https://www.kaggle.com/ronitf/heart-disease-uci?select=heart.csv). This training dataset contain 14 attributes or features which are defined in Table.1 as well as it contains 303 records among them 164 records belong to NEGATIVE and 139 records belong to POSITIVE classes or targets.

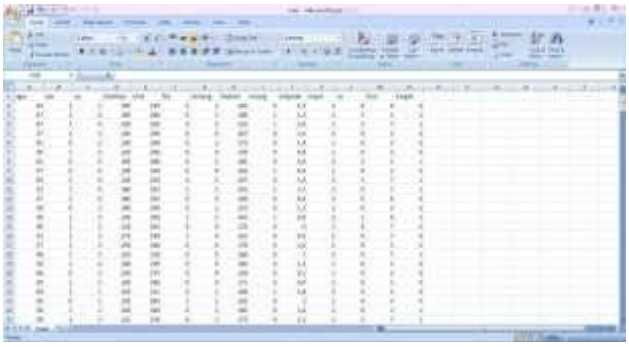

Figure.2 Heart Disease Dataset

Preprocessing

In the preprocessing we need to load or read the training dataset with help of pandas library and by importing the pandas library we can invoke *read_csv ()*method for read the entire dataset and store in a variable. The below snippet can show the training dataset loading processes.

```
from PyQt5 import QtCore, QtGui, QtWidgets
import pandas as pd
import sys
import numpy as np

df = pd.read_csv("heartdisease.csv")
```

Feature Extraction

After completion of preprocessing such as loading the training dataset, we need to get features of given dataset. By using feature extraction method this system can separate the input and output attributes and both are storing in *x_train* and *y_train* variable respectively. The below snippet will shows that syntax.

```
from PyQt5 import QtCore, QtGui, QtWidgets
import pandas as pd
import sys
import numpy as np

df = pd.read_csv("heartdisease.csv")

x_train = np.array(df.drop(['class'], 1))

y_train = np.array(df['class'])
```

By removing the target or class column using the drop() function, we can convert the input attribute values to array format using the numpy module, and then use the data frames to store only the values of the output variables or columns. The array conversion module from Numpy must be imported here.

Classification Techniques

DT:

One supervised machine learning approach is the DT classifier, which consists of a root node, many nodes, and finally, leaf nodes. The DT classifier may then arrange the dataset into a tree structure. It then applies the IF and THEN rules, which means it compares with every node until it reaches the leaf node, when the user enters the testing dataset for heart disease prediction. The desired column values, like POSITIVE or NEGATIVE, are stored in the leaf nodes. In the end, the values of the testing dataset and the final leaf node—which can represent the predicted values of the system—are identical. Here, we're training a model to predict the occurrence of heart disease by importing the DecisionTreeClassifier from the sklearn.tree package. The construction of a DT classifier model is shown in the following code sample.

```
from sklearn.tree import DecisionTreeClassif:
rf = DecisionTreeClassifier()
rf.fit(x_train, y_train)
pre_cls = rf.predict(x_test)
```

RF:

Combining decision trees is what makes up the RF classifier. Another algorithm that falls within the category of supervised machine learning. In this case, the RF classifier will randomly accumulate a number of decision trees before making a call. When making a disease prognosis, it requires all

decision trees' output values at random, and the system's predictable output status is determined by the class that received the most votes. The RF classifier outperforms the other algorithms in this system with an accuracy of 98%. The RandomForestClassifier will be imported into the system's training model for heart disease prediction using the sklearn.ensemble package. You can see the build model preparation syntax below.

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(x_train, y_train)
pre_cls = rf.predict(x_test)
```

NN:

Since it mimics how neurons in the brain function, the neural network classifier outperforms all others. There are three layers in a NN classifier: input, hidden, and output. In this case, the input layer gathers features from the dataset, passes them on to the hidden layers for feature categorization, and then shares the results with the output layer. When the classifier's output is consistent and the percentage it matches with the output layers is high. In order to forecast cardiac problems, this classifier may also use the sklearn.neural_network package MLPClassifier. Follow the below snippet code:

```
from sklearn.neural_network import MLPClassifie
rf = MLPClassifier()
rf.fit(x_train, y_train)
pre_cls = rf.predict(x_test)
```

NB:

In this system the naïve bayes algorithm can be used for prediction of heart disease. This algorithm follows the bayes rule to hear disease prediction. It

Journal of Management & Entrepreneurship
ISSN 2229-5348

UGC Care Group I Journal
Vol-11 Issue-01 June 2022

is fastest and easily predictable classifier and it calculates posterior probability events with other events and this algorithms uses mostly for text classifications. This classifier *MultinomialNB*isimporting from *sklearn.naive_bayes*package. The classifier following the below snippet.

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
nb.fit(x_train, y_train)
pre_cls = nb.predict(x_test)
```

SVM:

Due to its classification benefits, support vector machine classifiers are an essential kind of classifier. As an initial step in feature classification, support vector machines (SVMs) may create borders between classes and then use support vectors—the classes that are geographically closest to the hyper plane line—to divide them. Here, the system may develop a training model to forecast the state of heart disease by separating hyper planes with positive and negative characteristics, selecting the closest support vectors, and so on. The code for heart disease prediction is followed by the syntax below.

```
from sklearn import svm
svm = svm.SVC()
svm.fit(x_train, y_train)
pre_cls = svm.predict(x_test)
```

KNN:

In contrast to other machine learning classifiers, the K-nearest neighbour classifier uses the Euclidean distance formula to determine how far apart two points are. Along with making predictions, this classifier determines the distance It takes the distance between each record and stores it in a predictable way, so we can see which one has the shortest distance when we compare all of the data; this value becomes our heart disease prediction, which may be either positive or negative. Another module from sklearn.neighbors, KNeighborsClassifier, has to be imported. In this case, we selected the output value of the closest distance with a K value of 1.

```
from sklearn.neighbors import KNeighborsClassif
knn=KNeighborsClassifier()
knn.fit(x_train, y_train)
pre_cls = knn.predict(x_test)
```

Prediction

After the training model is built using the best classifier for each instance, this module will run. Invoking the predict() function with the testing dataset as input is necessary for heart disease prediction. All classifiers will have this approach. If you execute this method, it will begin comparing the training dataset with the provided testing dataset and classifier, and it will produce an output with the target column that is very close to the training dataset. You may use the syntax below to forecast whether heart disease will be positive or negative.

1.


```
from PyQt5 import QtCore, QtGui, QtWidgets
from sklearn.ensemble import RandomForestClassifier
import pandas as pd
import sys
import numpy as np

df = pd.read_csv("heartdisease.csv")

x_train = np.array(df.drop(['class'], 1))

y_train = np.array(df['class'])

tf = pd.read_csv("testing_dataset.csv")

testdata = np.array(tf)

testdata = testdata.reshape(len(testdata), -1)

rf = RandomForestClassifier()

rf.fit(x_train, y_train)

result = rf.predict(testdata)
```

Calculation of Accuracy

Here the system can calculate accuracy between six supervised machine learning algorithms. For this we need to split the heart disease dataset with 70% as training dataset and remaining 30% as testing dataset can be done with help of *train_test_split()* method which is importing from *sklearn.model_selection*package. Therefore it can return the *x_train, x_test, y_train, y_test* parameters then by taking the inputs as *x_train, y_train*and build the training model with respective classifiers and get the predicted classes *pre_cls* by invoke the prediction function by taking input as *x_test*then finally calling *metrics.accuracy_score ()*withinput*y_test*and*pre_cls* then it returns the accuracy of each respective classifier. The below snippet will show the process of accuracy calculation.

```
from PyQt5 import QtCore, QtGui, QtWidgets
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
import pandas as pd
import sys
import numpy as np

df = pd.read_csv("heartdisease.csv")

datainput = np.array(df.drop(['class'], 1))

y = np.array(df['class'])

x_train, x_test, y_train, y_test = train_test_split(datainput, y, test_size=0.3)

rf = RandomForestClassifier()

rf.fit(x_train, y_train)

pre_cls = rf.predict(x_test)

accuracy_rf = metrics.accuracy_score(y_test, pre_cls) * 100
```

4. EXPERIMENTSL RESULTS

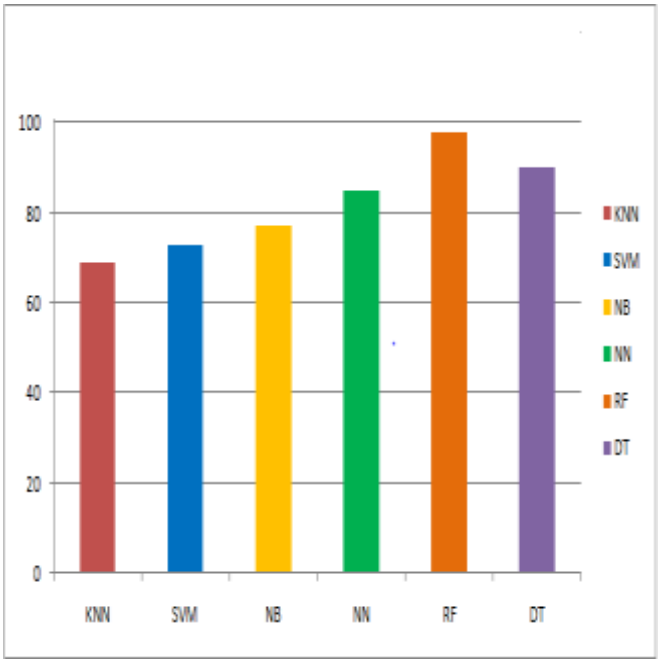

Figure.3 Accuracy Comparison between six Classifiers

2.

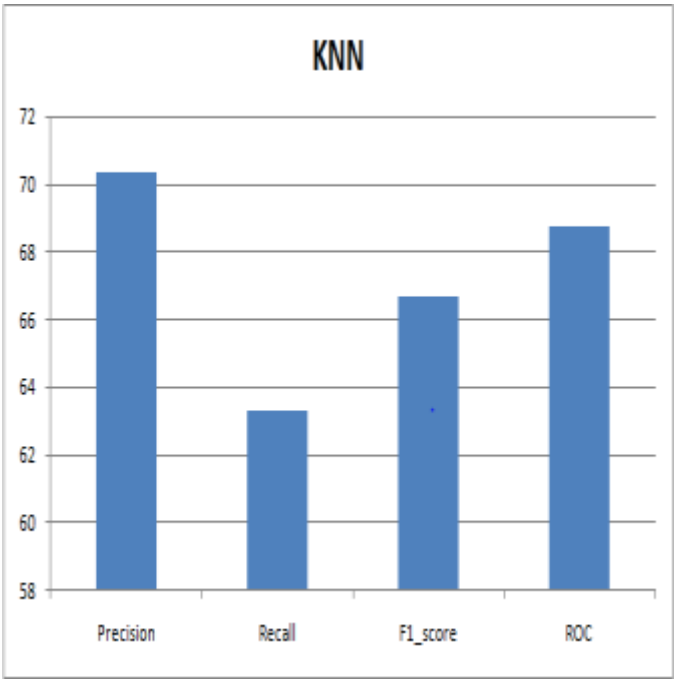

Figure.4 KNN Algorithm Metrics

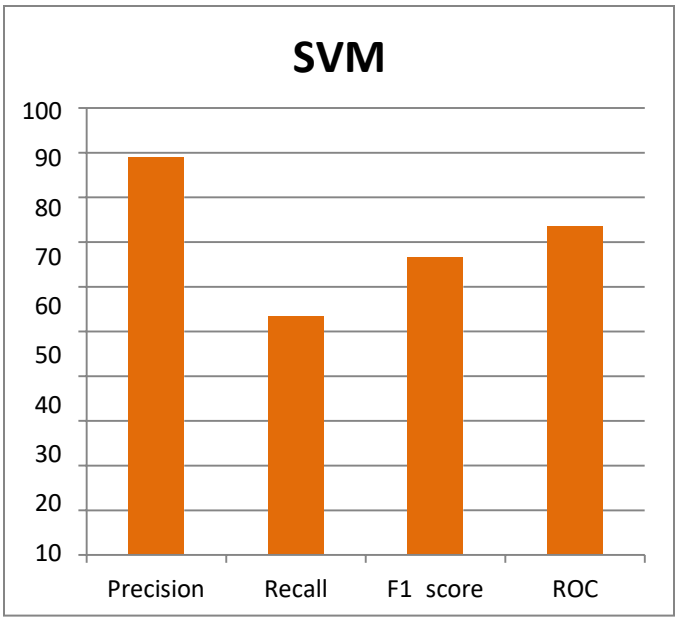

Figure.5 SVM Algorithm Metrics

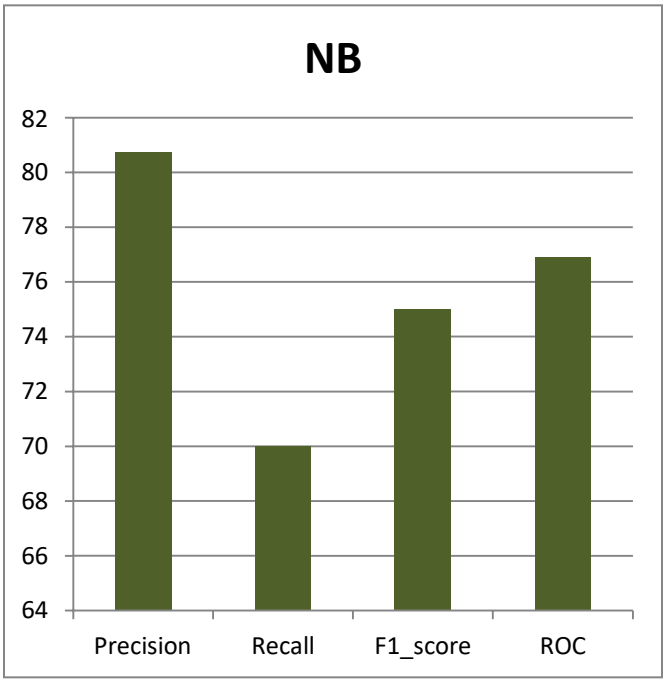

Figure.6 Naïve Bayes Algorithm Metrics

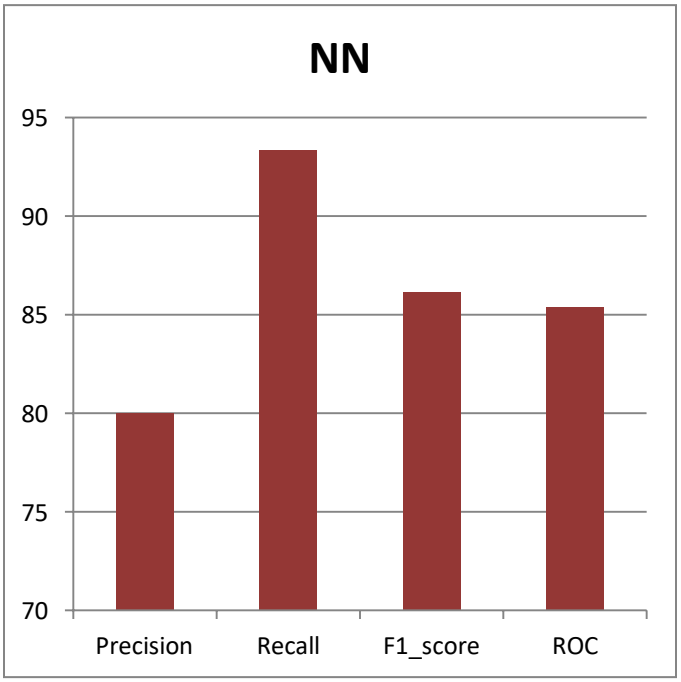

Figure.7 Neural Networks Algorithm Metrics

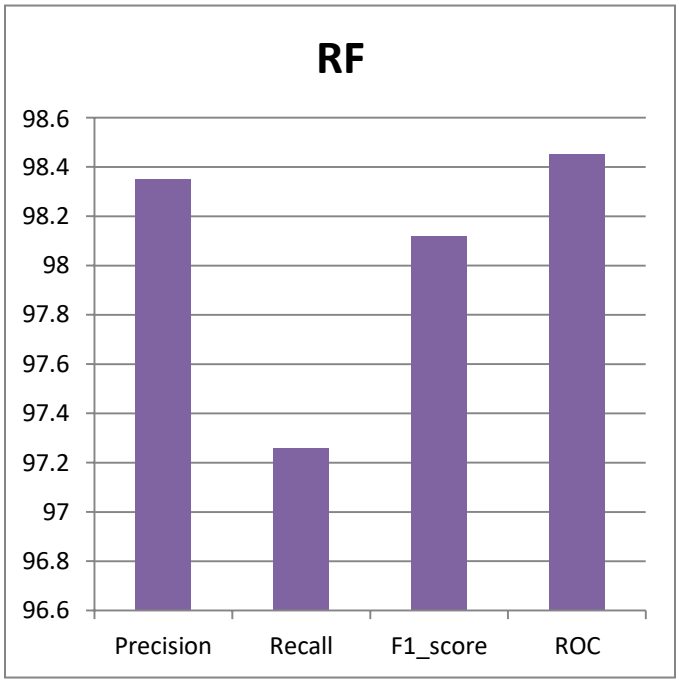

Figure.8 Random Forest Algorithm Metrics

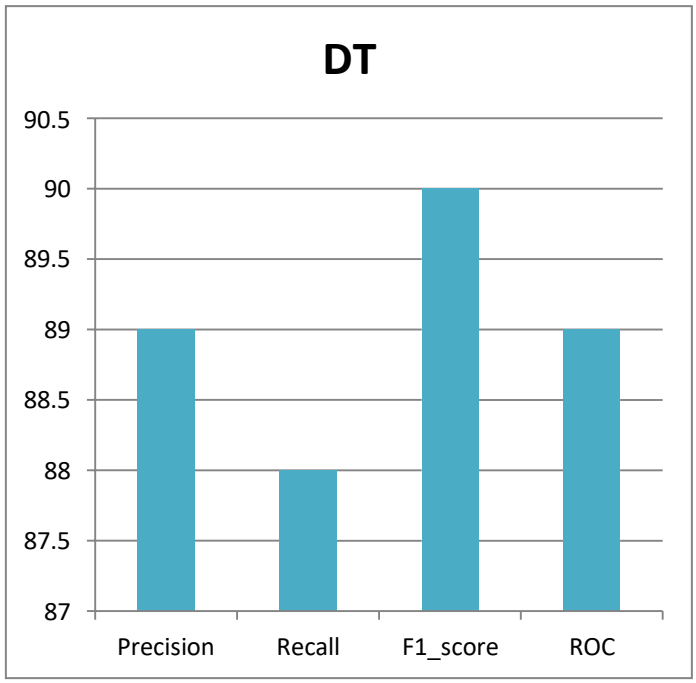

Figure.9 Decision Tree Algorithm Metrics

5. CONCLUSION

We have access to a wealth of medical history data because to the internet. Disease prognosis now relies heavily on the extraction and interpretation of medical history data. The

mortality toll from cardiovascular disorders, in particular, is rising daily. It's eThNiso:r2a2t0e.

We can lessen the likelihood of sickness by examining the medical history data of heart patients in order to make predictions. Our proposal for this research is to compare and contrast many well-known classification algorithms for the purpose of predicting the occurrence of heart disease. We sort the data according to the accuracy calculations and compare the outcomes. In this case, we classified medical data from heart attacks and computed an accuracy score using KNN, SVM, NB, NN, DT, and Random Forest. We achieved an accuracy rate of 98% using the Random Forest algorithm in these methods. We used the Random Forest method to forecast the likelihood of heart disease in users.

6. REFERENCES

[1] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Predictionusing Data Mining with Mapreduce Algorithm", International Journalof Innovative Technology and Exploring Engineering (IJITEE) ISSN:2278-3075, Volume-8 Issue-3,January 2019.

[2] Fahd Saleh Alotaibi," Implementation of Machine Learning Model toPredict Heart Failure Disease", (IJACSA) International Journal ofAdvanced Computer Science and Applications,Vol. 10, No. 6, 2019.

[3] AvinashGolande, Pavan Kumar T, "Heart Disease Prediction UsingEffective Machine Learning Techniques", International Journal ofRecent Technology and Engineering, Vol 8, pp.944-950,2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti,Ramya G Franklin,"Design And Implementation Heart Disease Prediction Using NaivesBayesian",International Conference on Trends in Electronics andInformation(ICOEI 2019).

[5] Nagaraj M Lutimath,ChethanC,Basavaraj S Pol.,"Prediction Of HeartDisease using Machine Learning", International journal Of RecentTechnology and Engineering,8,(2S10), pp 474-477, 2019.

[6] Theresa Princy R,J. Thomas,"Human heart Disease Prediction Systemusing Data Mining Techniques", International Conference on CircuitPower and ComputingTechnologies,Bangalore,2016.

[7] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―MedicalData Mining for Heart Diseases and the Future of Sequential Mining inMedical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[8]Puneet Bansal and Ridhi Saini et al. "Classification of heartdiseases from ECG signals using wavelet transform and kNNclassifier", International Conference on Computing,Communication and Automation (ICCCA2015).

[9] V. Krishnaiah, G. Narsimha, and N. Subhash, „„„Heart disease prediction system using data mining techniques and intelligent fuzzyapproach: A review,"" Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51,2016.

[10] S. Radhimeenakshi, „„„Classification and prediction of heart disease riskusing data mining techniques of support vector machine and artificialneural network,"" in Proc. 3rd Int. Conf. Comput. Sustain. Global Develop.(INDIACom), New Delhi, India, Mar. 2016, pp. 3107–3111.