Heart Disease Prediction Using Data Mining Techniques

Niraj Upadhayaya and Tejaswini Juluri

February 24, 2023

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

DR.NIRAJ UPADHAYAYA
Professor
Department of computer science& engineering
J.B.Institute of Engineering & Technology
nirajup@gmail.com

JULURI TEJASWINI
Department of computer science& engineering
J. B. Institute of Engineering & Technology
juluritejaswini.2000@gmail.com

Abstract

The heart disease prediction - HDP is an important task in health care domain now a day. Because for every minute, the number of people passing away with heart attack. It is difficult to HDP by physicians with huge health records. To overcome this complexity we need to implement the automatic heard disease prediction system to notify the patient and get to recovery from the disease. Here to gaining the automatic system we are using machine learning techniques to easily performing HDP with huge data. The machine learning techniques can be split into multiple types like unsupervised and supervised learning classifier. The unsupervised learning techniques used for prediction with unstructured data. But the supervised learning techniques working with structured data which is recommended to implement this classifiers. So, in this system we are using supervised machine learning techniques such as KNN, RF, NN, DT, NB, and SVM classifiers. For HDP, this system is using training dataset which is accessing from UCI machine learning repository. As well as this system is comparing accuracy performance between various ML - algorithms and shows the accuracy results with graphical presentation.

Keywords: - Dataset Collection, Classification, Prediction

1. INTRODUCTION

Day by data the huge of health records are raises in healthcare medical industry. So, there is a recommended to manage a huge data and make them as useful information for favorable decision making. Due to this problem, healthcare industry wants to implement an automatic system technique which will deliver productive decision from a huge dataset. So, the machine learning techniques are effective of resolving these kinds of issues very well. Because it can provide effectible methods to retrieve meaningful information without analyze the huge database. In the medical industry, the important data can be gathered from various patients' manifestations and clinical reports for analysis by physicians. These days at stage of lifetime lot of people are getting heart failure symptoms. But comparing between old people and young people, the senior citizens are facing this type of problems. However, the machine learning techniques can find correlations between different features for prediction of heart disease status from training dataset. By using this kind of training models, it can detect the heart disease patients without help of medical practitioners. Then it can pretend as an automatic system to categorize between positive heart disease patients and negative heart disease patients with accurately, so then it reduces the diagnosis time and cost of treatment.

In the health care domain, providing qualities services and predict the diagnosis status accurately is a main challenge task. According survey a lot of people passed away with heart disease even managed and controlled effectively by automatic system. Here any disease can be controlled by dependents of detection of that disease at right time. So in this system the proposed system can predict the heat disease status at advance stage to notify the patients and help them to recovery from that disease. The huge of medical records are generated by medical experts for analyze and retrieve the useful information form that database. The health care database contains mostly unattached information which is tedious task for prediction of heart disease. So, in the health care domain if we implement the machine learning techniques which is understand structured information to prediction of various diseases. Therefore, this system proposed a automatic system to physicians for prediction of heart disease at advance stage then they can provide treatment to patient and save them from rugged seriousness. So, the machine learning techniques have an important role in HDP with supervised classifiers at advance stage to diagnoses the patients.

2. RELATED WORK

The detection or prediction of heart disease is toughest task in the health care domain. The physicians can detect heart disease with some symptoms such as smoking, high in taking of fat and consuming alcohol etc. But with this symptoms the physicians can detect disease may or may not be accurately. Due to this reasons doctors cannot be do treatment to patients at early stages then there is a chance to face the harmful outcomes by patients. So, to detect or prediction heart disease we need to develop the tool for detection of any disease at early stages then the physicians will do the treatment to patients to preventing harmful consequences. Here the prediction tool can be implementing by the supervised machine learning techniques to HDP. Many clinical or hospitals generate huge medical records which is unstructured format. So by using this prediction tool can easily fetch the useful information to make the training dataset for further references. The machine leaning algorithms will take this training dataset as input and predict the heart disease status with current patient details as testing dataset.

In this system we are proposed HDP with six machine learning classification algorithms and shows the accuracy results between these six algorithms. Here the main task of this system is predict the accurately when patient was feel pain with heart disease. Regarding implementation with help of heart disease training dataset and machine learning classifiers we build the train model file and then the physicians can enter the input values which is get from patients health records and giving to the training model as input for HDP. Here the heart disease training dataset is downloading from UCI repository which is uploaded by the medical department experts. To build this system we had chosen python language which is has pre-trained libraries or packages to access the machine learning classifiers. Here all six algorithms providing best accuracies to prediction of heart disease.

3. IMPLEMENTATION

System Model

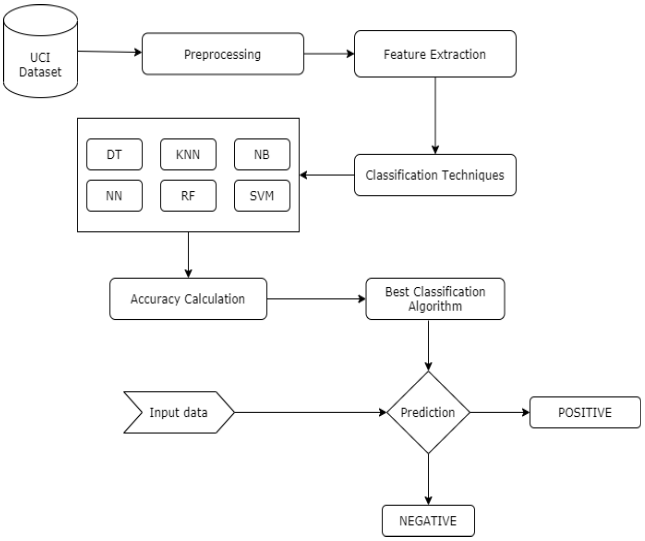
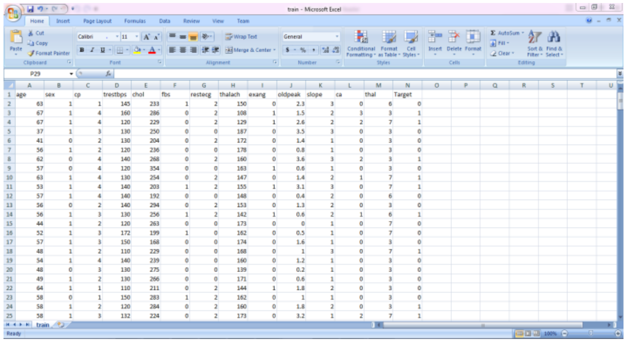


Figure.1 System Architecture

The figure.1 depicts about our proposed system model. In this system model they used heart disease training dataset which is downloaded from UCI repository. Later by preprocessing, it can read the training dataset and split the independent and dependent attributes by feature extraction and then build the training model with classification algorithm for HDP by giving input data, finally calculate the accuracy between six machine learning 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.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Target
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	1
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	1
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0
16	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
17	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
18	48	1	2	110	229	0	0	168	0	1	3	0	7	1
19	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
20	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
21	49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
22	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
23	58	0	1	150	283	1	2	162	0	1	1	0	3	0
24	58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
25	58	1	3	132	224	0	2	173	0	3.2	1	2	7	1

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.

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.

Classification Techniques

DT:

It is a tree-based classifier which is preparing the training model with tree format. This classifier will start with a single root and with multiple nodes and ended with a number of leaf nodes. After building the training model for the prediction process it can follow the IF and THEN methodology. Here first from the root it can check the next node satisfies the condition then it can move to another node like that it can reach up to leaf nodes where it determines the predictable of student performance with the admission of which university.

In this system, we are using *sklearn.tree* package to import the *DecisionTreeClassifier* to build training model for HDP. The below snippet will shows the building of DT classifier model.

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

RF:

This system will use *sklearn.ensemble* package to import the *RandomForestClassifier* to build training model for HDP. The below syntax will shows the preparation of build model.

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

NN:

It is a related deep learning technique where it can work with brain neurons. It most advance classifier compare with the remaining machine learning classifiers. The NN classifier will process with three layers such as input layer, middle layers like hidden layer, and final layer as output layer. After collecting the attribute values from feature extraction then these values will be feed to input layers. Here each input layer will be passing the values to each hidden layer to processing the network weigh values then later it can calculate all weight values and send them to the output layer. Finally, the output layer will be decided by the predicted result value by comparing the highest weight values with all hidden layers' weight values.

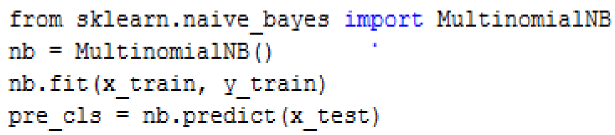
This classifier also can import *sklearn.neural_network* package *MLPClassifier* for HDP. Follow the below snippet code:

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

NB:

The NB classifier is a machine learning classifier where it can predict with probabilities methodology. This algorithm follows the Bayes rule to perform the prediction of student performance. It is the fastest and easily predictable classifier and it calculates posterior probability events with other events and this algorithm uses mostly for text classifications.

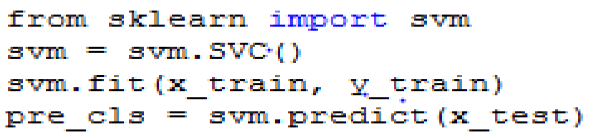
This classifier *MultinomialNB* is importing from *sklearn.naive_bayes* package. The classifier following the below snippet.



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

SVM:

In machine learning, it needs to know SVM classifier what kind of problem statements will be solved with the help of supervised learning. It is useful in solving both classification and regression problem statements. The five elements support vectors, Hyperplane, Marginal distance, linear separable, and Non-linear separable are involved in SVM classification. The main aim of this SVM, suppose if it considers a classification problem, it can easily separate the two classes points like positive and negative points. So that, the SVM can classify these points with the hyper plane which is the centerline of these two classes points. Here the SVM makes sure that when it creates the hyperplane then it also creates the two margin lines parallel and these two margin lines have computed some distance therefore it will be easily linearly separable for both the classification points. But the SVM makes sure that one of the margin lines passes through one of the nearest positive points, similarly, another margin line also passes towards negative points. These nearest points are called support vectors. The SVM not only focuses on generating the hyper lane it is also focused on generations of margin lines to get a better accuracy model, which is behind the intuition of SVM.
The below syntax is following the code of HDP.



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

KNN:

The K-NN algorithm is the most popular supervised machine learning algorithm. The K-NN classifier is also used for solving the regression and classification problems but regularly it is implemented for resolving the classification problems. Here *K* represents the integer value from 1 to n numbers. It easily classifies the non-linear data points which are distributed in a non-linear manner. Here there is no need to draw a straight line and classify those data points. The K-NN algorithm is majorly based on similarities, it can classify the new data points. In the process of classification problems, the *K* value should be selected where it will indicate that how many nearest values it considers in terms of distance. In the K-NN algorithm, the distance will be calculated by two parameters such as Euclidean distance and Manhattan distance. Most of the classification problems will be solved by the Euclidean distance formula in the K-NN algorithm.

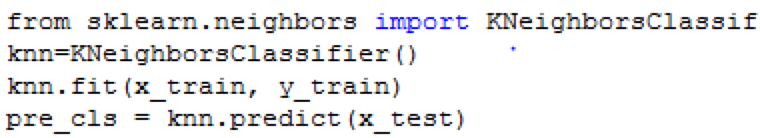
Euclidean:

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Where x, y are data points.

D indicates the distance

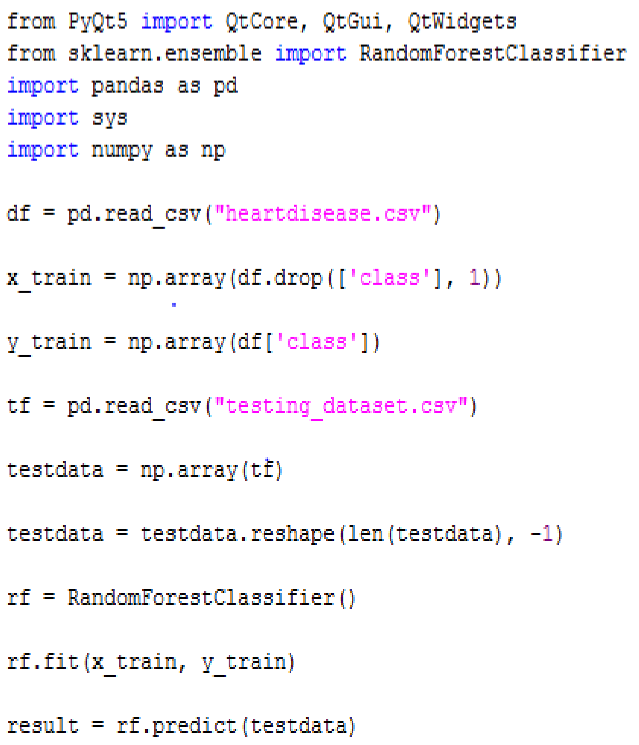
It is also import the *KNeighborsClassifier* module from this *sklearn.neighbors.* Here we took *K* value is 1 for pick the nearest distance value as output.



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

Prediction

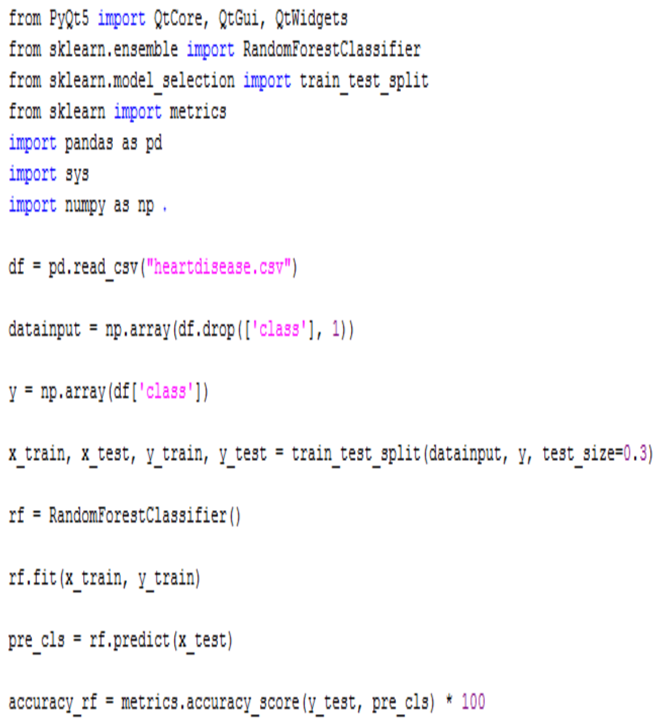
This module will be executing after build the training model with respective best classifier. For HDP we need to invoke *predict ()* method with testing dataset as input. This method will be available in every classifier. By calling this function it can start to comparing with training dataset with given testing dataset with respective classifier and it returns the target column as output which is matches to near with training dataset. The below syntax is used for prediction of heart disease as POSITIVE or NEGATIVE.



```
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 ()* with input *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

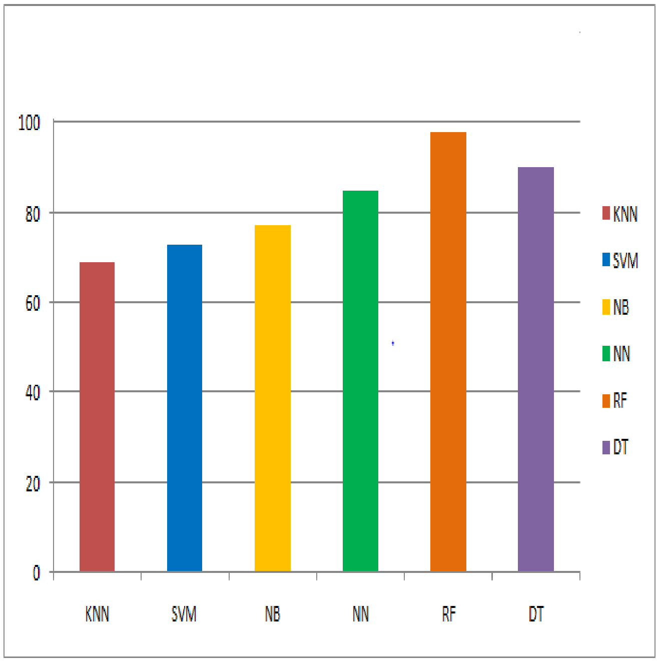


Figure.3 Accuracy Comparison between six Classifiers

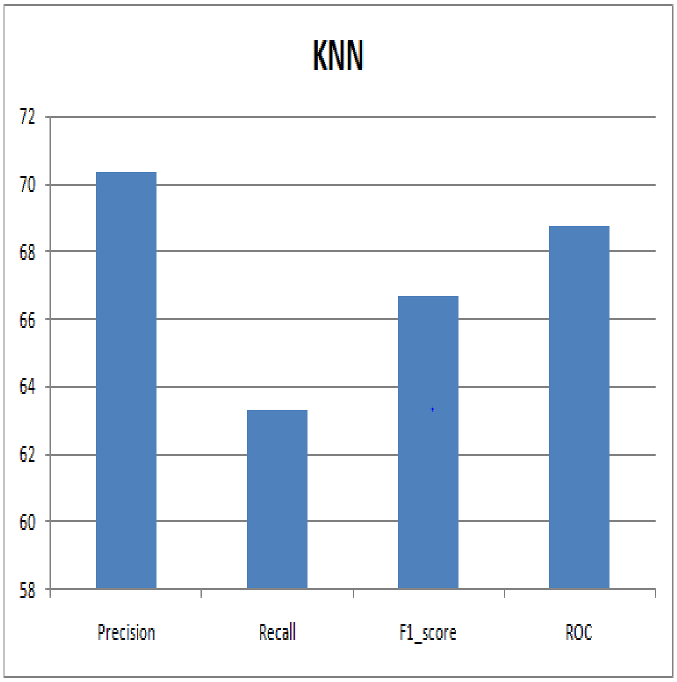


Figure.4 KNN Algorithm Metrics

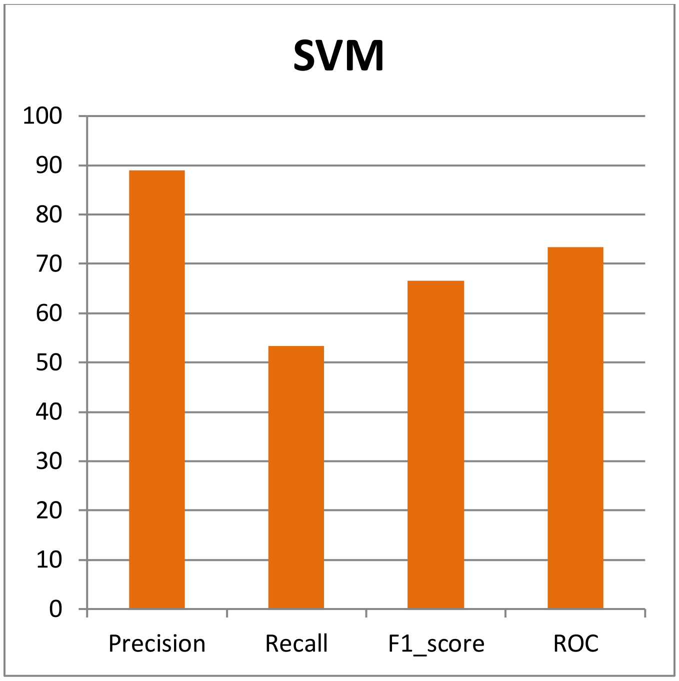


Figure.5 SVM Algorithm Metrics

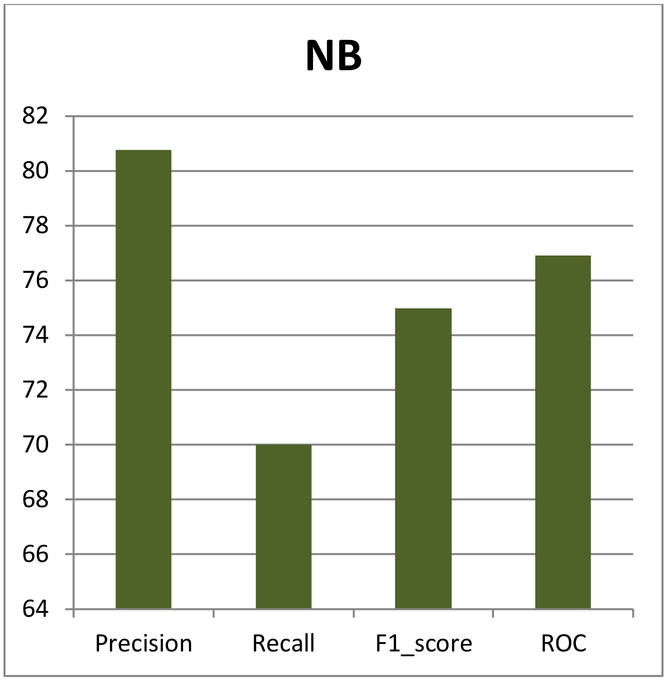


Figure.6 Naïve Bayes Algorithm Metrics

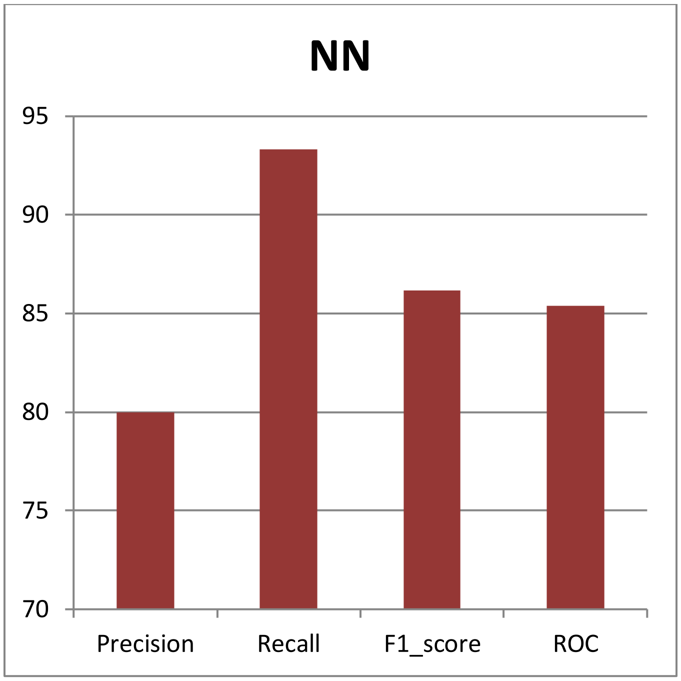


Figure.7 Neural Networks Algorithm Metrics

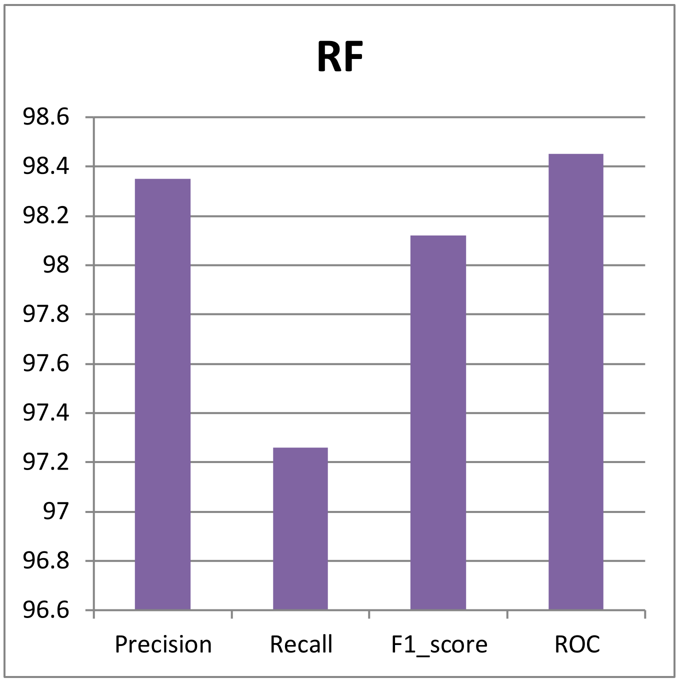


Figure.8 Random Forest Algorithm Metrics

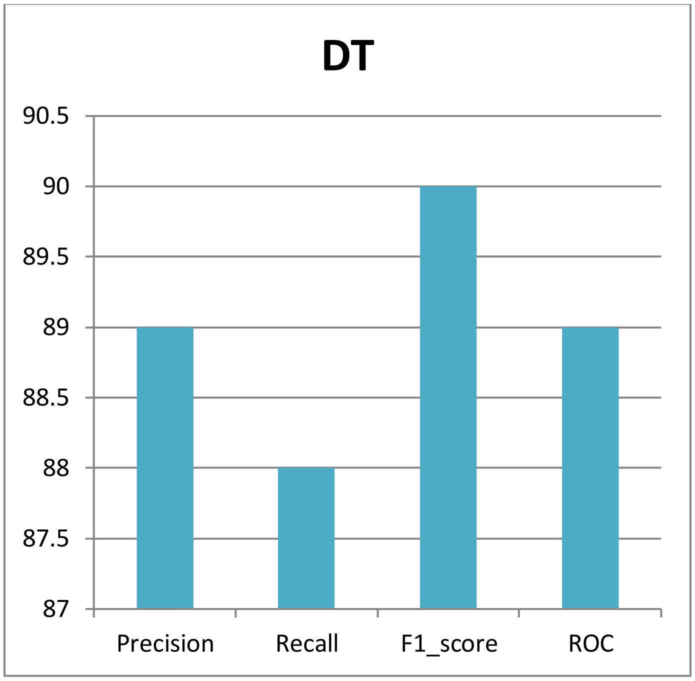


Figure.9 Decision Tree Algorithm Metrics

5. CONCLUSION

By taking the advantages of online world we have lot of medical history data is available. Extracting and analysis medical history data is become very necessary for the prediction of the diseases. Especially in heart diseases, the rate of deaths due to heart attacks is increasing day by day. This rate we can decrease by predicting disease by analyzing the heart patient's medical history data. In this project we propose a comparative analysis of HDP using popular classification algorithms. We classify and compare the results in terms of Accuracy calculation. Here we have used KNN, SVM, NB, NN, DT and Random Forest for classifying the heart attack medical data and calculate the accuracy score. In these algorithms we got 98% of accuracy for Random Forest algorithm. We deployed Random Forest algorithm for user HDPs.

6. REFERENCES

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

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

[3] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent 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 Naives Bayesian",International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[6] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

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

[8] Puneet Bansal and Ridhi Saini et al. "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier", 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 fuzzy approach: A review,'' Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51, 2016.

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