PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS

[1]Dr. Ashok Kumar P S, [2]Riqa Rabiya, [3]Sowmya V, [4]Srishti Singh, [5]Srinidhi M
[1]Professor, HOD, Department of Information Science and Engineering, Don Bosco Institute of TechnologyStudent,
[2,3,4,5]Department of Information Science and engineering, Don Bosco Institute of Technology Mysore Road, Kumbalgudu, Bengaluru- 560 074.

Abst rac t : Machine learning techniques can be used to predict the medical datasets at an early stage for safe human life. A huge medical datasets are accessi ble in different data repositories which are used in real world applications. It is impractical for a common man to frequentl y undergo costl y tests like the ECG and various other tests. Hence there needs to be a sys t em in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus we propose t o devel op an application which can predict the vulnerability of a heart disease given basic sympt oms like age, sex, pulse rate etc. In this project we have worked on predicting potential heart disease in people using machine learning algorithms such as K Neighbors classi fier, Decision tree classifi er, Naïve Ba yes classi fier and Random forest classi fier

Inde x Terms: **Healthcare, cardiovascular disease, Prediction, Machine Learning Algorithms.**

I. INTRODUCTION

Nowadays, healthcare is increasing day by day due to lifestyle, hereditary. The modern world has cardiovascular disease as its deadliest enemy. This disease affects a person in such a way so that the patients can't be cured as easily as possible. So, diagnosing patients at the right time is the toughest work in medical field. Misunderstanding and wrong diagnosis made by the hospital leads to the bad reputation. India questions that the treatment for this disease is quite tough and can't is reachable by most of the patients. Everyone has different values for Blood pressure, cholesterol, and pulse rate. But per medically proven results the normal values of Blood pressure are *120/80*, cholesterol is less than 200 and pulse rate is 72.

Machine learning is an art of mastering system without being explicitly computed. They are used to analyze the analytical arrangement in high dimensional, diverse data sets like heart diseases. They are used in recognition of the arrangement that gives support for forecasting and controlling mechanism for analysis and medication. The world health organization reports suggest that greater than 12 million deaths are happening worldwide due to cardiovascular problems. The examination of the unhealthiness is a complex mechanism. It should be measured perfectly and precisely, Because lack of experts at some places is resulting in the patients in a hazardous position. Ordinarily, these are diagnosed by the cardiologists. It is extremely beneficial if these techniques are combined with the medical information system.

This paper suggests the different machine learning techniques that are used for forecasting the uncertainty levels of cardiovascular diseases based on the attributes present. The medical datasets used are taken from the research that had been fascinated throughout the world.

II. LITERATURE SURVEY

Machine learning algorithms such as naïve bayes, KNN etc. can be used for calculations for various types of heart-related issues [1]. In “Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients” , it is presumed that albeit most analysts are utilizing diverse classifier methods, for example, Neural system, SVM, KNN and twofold discretization with Gain Ratio Decision Tree in the conclusion of coronary illness [2].

Depiction of how these machine learning calculations are utilized to foresee the various diseases. In the present world, there are numerous logical innovations which help specialists in taking clinical choices however they won't be precise these can be done through various algorithms[3].

Specialists may some of the time neglect to take precise choices while diagnosing the coronary illness of a patient, in this manner coronary illness forecast frameworks which utilize machine learning calculations aid such cases to get exact outcomes [4].

There are many instruments accessible which utilize expectation calculations yet they have a few blemishes. A large portion of the instruments can't deal with huge information and most are not brought together, not conveyed on cloud and consequently not open on the web [5].

III. ARCHITECTURE

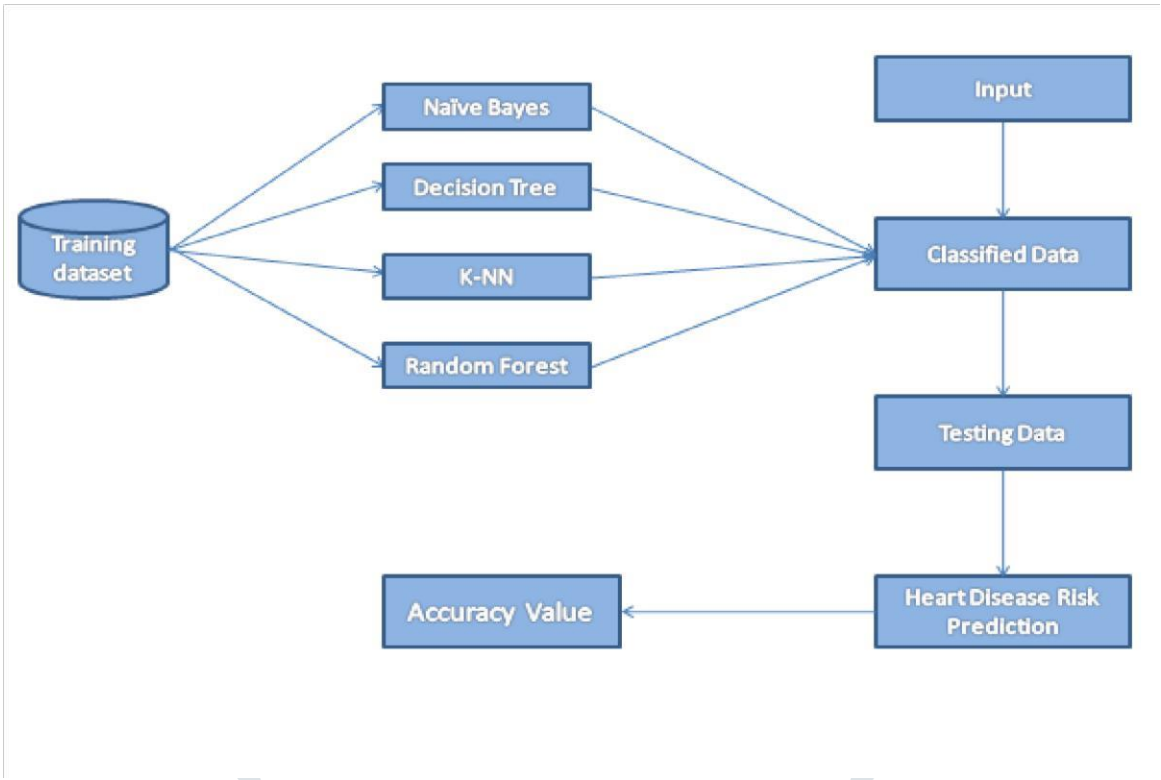


Figure 1. System Architecture

The above figure shows functioning of the system. It is described step by step

- Step 1. The Training dataset contains the details of the patients.
- Step 2. The dataset is classified using various algorithms such as Naïve Bayes classifier, Decision Tree, K-NN and Random Forest Algorithm.
- Step 3. The user input is given to classified data.
- Step 4. Testing is performed on the classified data to predict the heart disease risk.
- Step 5. It also finds the accuracy of the algorithms and compares the accuracy among all the algorithms.

IV. IMPLEMENTATION

1. Naïve Bayes

a) Definition

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

b) Formula

$$P(A|B)= \frac{P(B|A) * P(B)}{P(A)} \quad \text{……..(1.1)}$$

2. Decision Tree

a) Definition

Decision Tree algorithm belongs to the family of supervised learning algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

3. K-Nearest Neighbor

a) Definition

K-Nearest neighbor (KNN) is a simple, lazy and nonparametric classifier. KNN is preferred when all the features are continuous. KNN is also called as case-based reasoning and has been used in many applications like pattern recognition, statistical estimation. Classification is obtained by identifying the nearest neighbor to determine the class of an unknown sample.

b) Formula

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

……..(3.1)

4. Random Forest

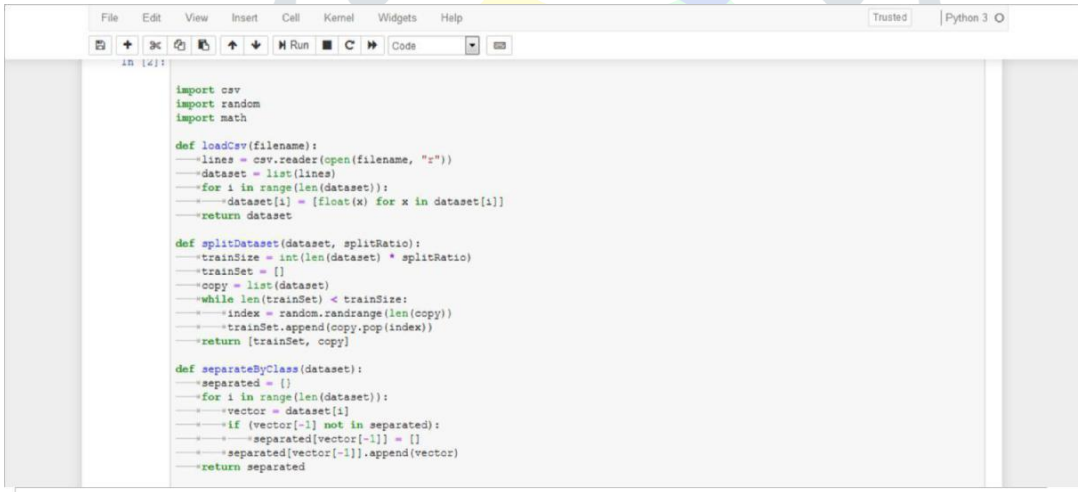
a) Definition

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

2. RESULT ANALYSIS

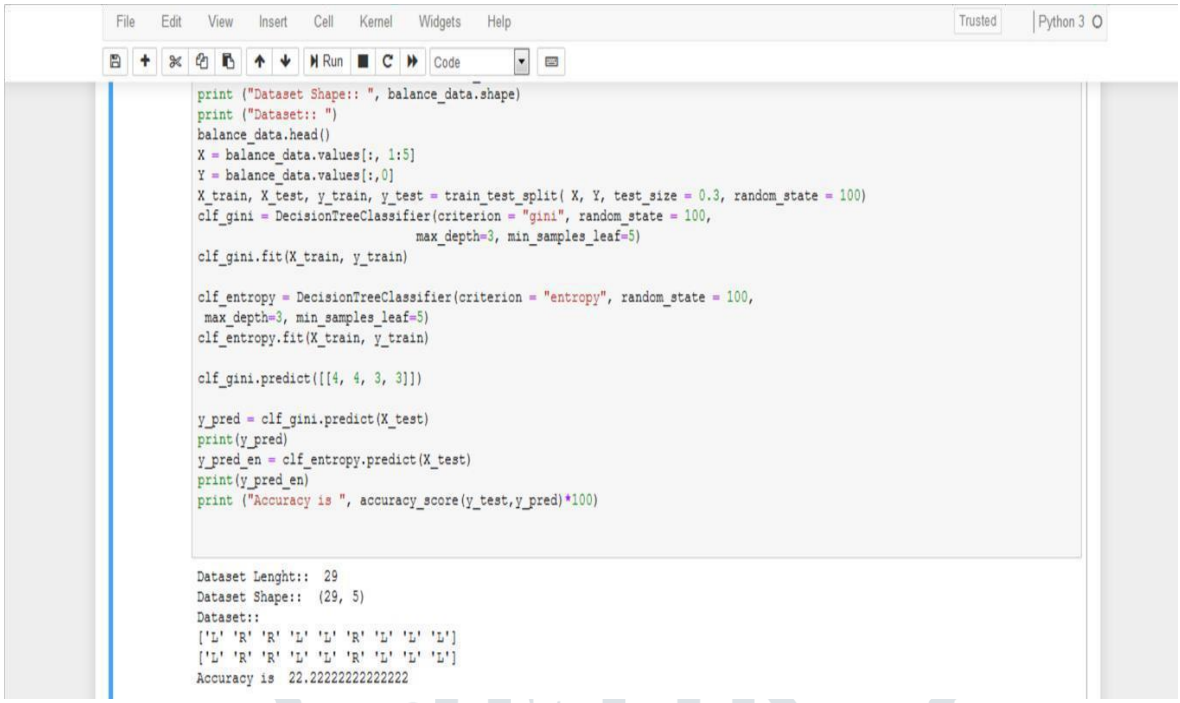
The system involved analysis of the heart disease patient dataset with proper data processing. Then, 4 models were trained and tested. The output of each algorithm with its snapshot is as follows:

1. Naïve Bayes




```
Split 768 rows into train=514 and test=254 rows
Accuracy for the presence of heart disease : 39.3700787401574748
```

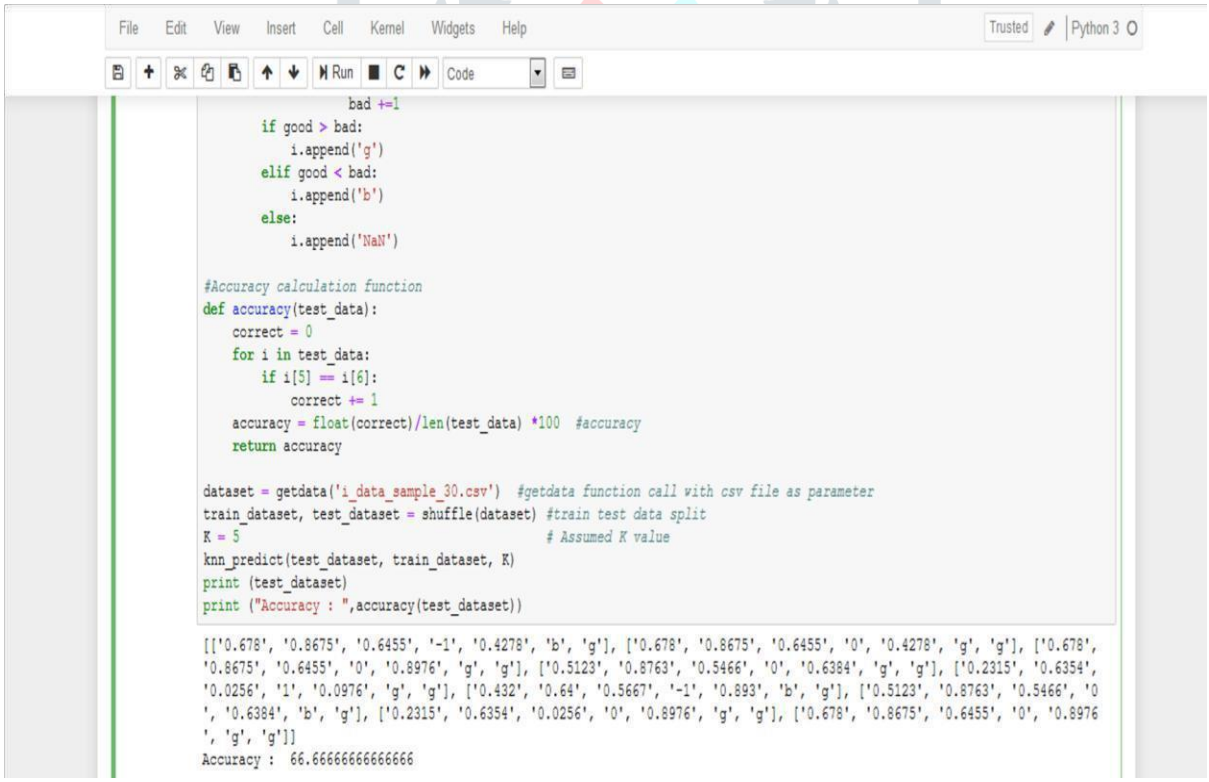
2. Decision Tree



```
print ("Dataset Shape:: ", balance_data.shape)
print ("Dataset:: ")
balance_data.head()
X = balance_data.values[:, 1:5]
Y = balance_data.values[:,0]
X_train, X_test, y_train, y_test = train_test_split( X, Y, test_size = 0.3, random_state = 100)
clf_gini = DecisionTreeClassifier(criterion = "gini", random_state = 100,
                                  max_depth=3, min_samples_leaf=5)
clf_gini.fit(X_train, y_train)

clf_entropy = DecisionTreeClassifier(criterion = "entropy", random_state = 100,
 max_depth=3, min_samples_leaf=5)
clf_entropy.fit(X_train, y_train)

clf_gini.predict([[4, 4, 3, 3]])

y_pred = clf_gini.predict(X_test)
print(y_pred)
y_pred_en = clf_entropy.predict(X_test)
print(y_pred_en)
print ("Accuracy is ", accuracy_score(y_test,y_pred)*100)
```

```
Dataset Lenght::  29
Dataset Shape::  (29, 5)
Dataset::
['L' 'R' 'R' 'L' 'L' 'R' 'L' 'L' 'L']
['L' 'R' 'R' 'L' 'L' 'R' 'L' 'L' 'L']
Accuracy is  22.22222222222222
```

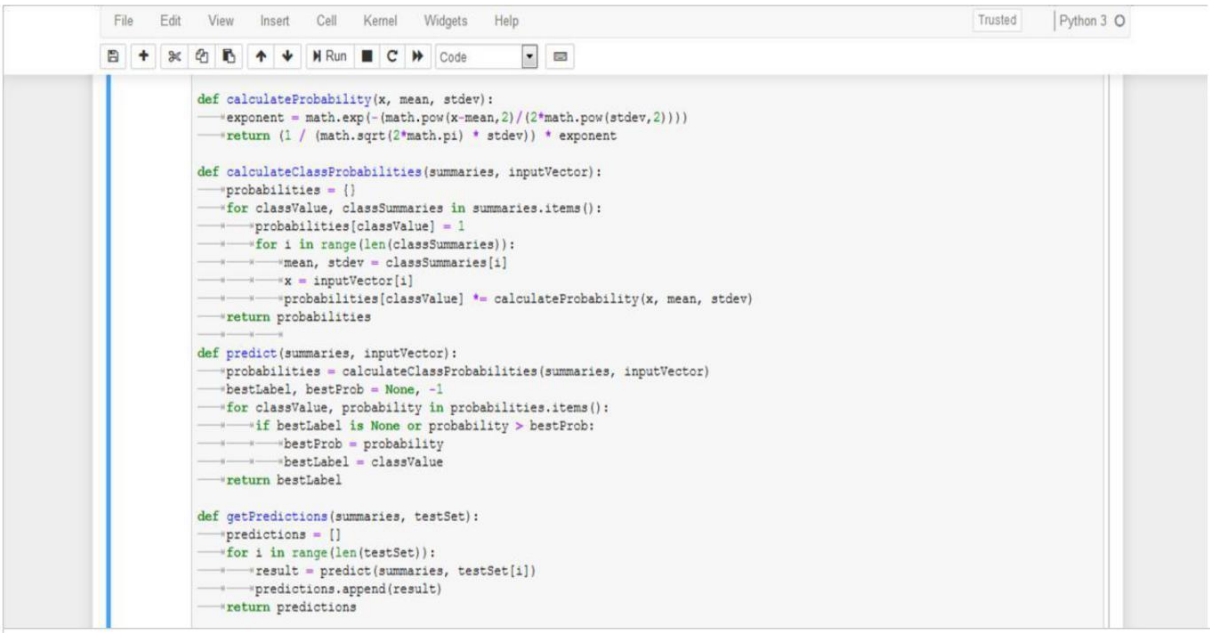
3. K-Nearest Neighbor



```
                bad +=1
        if good > bad:
            i.append('g')
        elif good < bad:
            i.append('b')
        else:
            i.append('NaN')

#Accuracy calculation function
def accuracy(test_data):
    correct = 0
    for i in test_data:
        if i[5] == i[6]:
            correct += 1
    accuracy = float(correct)/len(test_data) *100  #accuracy
    return accuracy

dataset = getdata('i_data_sample_30.csv')  #getdata function call with csv file as parameter
train_dataset, test_dataset = shuffle(dataset) #train test data split
K = 5                                  # Assumed K value
knn_predict(test_dataset, train_dataset, K)
print (test_dataset)
print ("Accuracy : ",accuracy(test_dataset))
```

```
[['0.678', '0.8675', '0.6455', '-1', '0.4278', 'b', 'g'], ['0.678', '0.8675', '0.6455', '0', '0.4278', 'g', 'g'], ['0.678',
'0.8675', '0.6455', '0', '0.8976', 'g', 'g'], ['0.5123', '0.8763', '0.5466', '0', '0.6384', 'g', 'g'], ['0.2315', '0.6354',
'0.0256', '1', '0.0976', 'g', 'g'], ['0.432', '0.64', '0.5667', '-1', '0.893', 'b', 'g'], ['0.5123', '0.8763', '0.5466', '0
', '0.6384', 'b', 'g'], ['0.2315', '0.6354', '0.0256', '0', '0.8976', 'g', 'g'], ['0.678', '0.8675', '0.6455', '0', '0.8976
', 'g', 'g']]
Accuracy :  66.66666666666666
```

4. Random Forest




```
Trees: 10
Scores: [75.60975609756098, 80.48780487804879, 92.6829268292683, 73.17073170731707, 70.73170731707317]
Mean Accuracy: 92.53746
```

➢ Algorithm Performance Analysis

ALGORITHMS	ACCURACY
Random Forest	92.53476%
Decision Tree	89.34284%
K-NN	66.66666%
Naïve Bayes	39.37008%

From the above performance analysis, it can be inferred that Random forest algorithm gives highest accuracy among the four algorithms.

3. CONCLUSION

It contributes the correlative application and analysis of distinct machine learning algorithms in the software which gives an immediate mechanism for the user to use the machine learning algorithms for forecasting the cardiovascular diseases. It is inferred that Random forest algorithm gives highest accuracy among the four algorithms. Future work includes different ensemble methods of these algorithms which can advance to better performance with more parameter setting for these algorithms.

4. REFERENCES

1. Jaymin Patel, Tejal Upadhyay, Samir Patel, Heart disease prediction using Machine learning and Data Mining Technique, vol. 7, no. 1, Sept 2017.
2. G. Parthiban, S. K. Srivasta, "Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients" in International Journal of Applied Information Systems (IJAIS)-, New York, USA:Foundation of Computer Science FCS,vol. 3, no. 7, August 2018, ISSN ISSN: 22490868.
3. K Thenmozhi, P Deepika, "Heart Disease Prediction using classification with different decision tree techniques", *International Journal of Engineering Research & General Science*, vol. 2, no. 6, pp. 6-11, Oct 2017.
4. Gregory F. Cooper, Constantin F. Aliferis, Richard Ambrosino, An evaluation of Machine learning methods for predicting pneumonia mortality, Elsevier, 2018.
5. Sanjay Kumar Sen, "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", *International Journal of Engineering And Computer Science*, vol. 6, no. 6, June 2017, ISSN ISSN: 2319-7242.
6. Abhishek Taneja, Heart Disease Prediction SystemUsing Data Mining Techniques, vol. 6, no. 4, December 2018.
7. Beant Kaur, Williamjeet Singh, "Review on Hear Disease Prediction System using Data Mining Techniques", *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 10, October 2017.
8. Younus Ahmad Malla, Mohammad Ubaidullah Bokari, "A Machine Learning Approach for Early Prediction of Breast Cancer", *International Journal of Engineering And Computer Science*, vol. 6, no. 5, May 2017.
9. M.A. Nishara Banul, B Gomathy, DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES, vol. 177, pp. 3799-3821, 2017.
10. Siva S. Sivanath, S. Geetha, A. Kannan, Decision tree based light weight intrusion detection using a wrapper approch, 2016.