TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

Comparative Analysis of Multiple Classification Algorithms on Heart Disease Prediction

[1]Harshitha M, [2]Sanju V
[1]M.Tech Part Time 6th semester Student, Computer Science & Engineering, School of Computing & Information Technology REVA University, Bangalore, India
[2]Associate Professor, School of Computing & Information Technology REVA University, Bangalore, India
[1]harshitham1991@gmail.com, [2]sanju.v@reva.edu.in

Article Info
Volume 83
Page Number: 4168-4175
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Heart disease is one of the main problem caused among the global group of people. It is one of the leading reasons of demise in the large group of middle aged population. It is essential to have a framework which could efficaciously recognize the coronary heart aliment in lot of samples at once. The proposed algorithms used in our work is (NB) Naive Bayesian, (DT) Decision Tree, (KNN) K-nearest neighbor, (ANN) Artificial Neural Networks, in predicting coronary heart disease. These algorithms can provide the likeliness of patients getting coronary heart problems. Few of the performance factors used in predicting heart disease are by using the factors Accuracy, Precision, Recall, F1-Score. In our work, we majorly focus on identifying the most efficient algorithm among the DT, NB, KNN and ANN.

Keywords**:** *Heart disease, multiple classification algorithms, Accuracy measure*

1. Introduction

The input information extracting technique involves gathering of very massive information from large data sources like databases. Usually healthcare region entails an abundance of information associated with patients, various diagnoses of the sickness and so on. Current the hospitals are adapting to the lifestyle IMS (Information Management Systems), to address their affected person's statistics systematically and efficaciously. A massive quantity of data is gathered by using such structures which might be represented the use of charts, numbers, text and snapshots, consolidated to a csv file. Such form of facts is rarely hired for making any scientific selections. The modern studies emphasizes on coronary heart disease diagnosis. Various strategies of system learning have been incorporated for diagnosing the disease thereby acquiring numerous probabilities. Concerning to the heart sickness predicting several systems are being endorsed that are being deployed by the way of various strategies and algorithms. Gaining a pleasant provider at a low cost rate says the top and tough situation for healthcare establishments. For imparting excellent offerings at par, there should be correct diagnosis of the sufferers alongside, a powerful dosage of drug treatments. Low exceptional medical analysis and remedy can result in inadequate effects. One solution for fee-slicing through Health Care institutions may be the utilization of PC-generated statistics or the use of DSS (Decision aid structures). Heath Care center includes an ambulance of information associated with patients, numerous diagnoses of the illness, aid control and so on.

Using an automated machine the affected person's remedy records can be stored and few strategies can be used to acquire various records and queries regarding the hospital. Supervised learning includes the utilization of training and studying the model parameters. No training set, is needed in unsupervised learning. Classification and prediction are the main primary approaches of facts mining. The Classification model techniques assist in segregating dis organized fact values, however prediction model anticipated non-stop values. The proposed work assures to be tremendous and powerful in handling

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

category, comparable to Machine Learning concerning to Deep Learning.

2. Literature Survey

Authors in [1] has focused on classification methods like KNN, Naive Bayes and Decision Tree. Data used is UCI heart disease dataset which has details of 303 patients. As per the accuarcies obtained by each algorithm, Naive Bayes is proved to be the most accurate algorithm.

Authors in [2] have presented that Naive Baye is most accurate, when compared to DT and Neural Network. The data used is the UCI heart disease data set.

Authors in [3] in their work have discussed KNN, Decision Tree (ID3), Gaussian Naïve Bayes, has justified Naive Bayes to be the highest accuracy obtained.

Authors in [4] has focused their work on data mining techniques like Decision Trees, Naive Bayes and Neural Network. Neural Network gives the highest accuracy, however their data set used is quite small.

Authors in [5], has worked on predicting heart disease using Artificial Neural Network

Our work focuses on using all the above classifiers combination like, Artificial Neural Network, Naive Bayes, KNN and Decision Tree in predicting Heart Disease, using different performance metrics like Accuracy, Precision, Recall, F1 score. Our Data has around 25000 patients records with 14 attributes. As per our work ANN(83.97%) is the most accurate algorithm, followed by Naive Bayes(83.06%), followed by KNN(82.79%) and DT(76.68%). Our work focuses, on not only accuracy, but also on other factors to measure the performance.

3. Methodology

Inspired by the way of the developing price of affected person's demise attributable to coronary heart disease every year, there's growing availability of affected person facts which can assist specialists. Acquisition in the machine getting to know includes two matters, statistics and version. When gathering the information it ought to have enough functions so that it may assist to predict the disorder and efficiently train the learning version. In a per-processing step, the information is to be cleaned and simplifies. By per-processing the records, we are able to extra effortlessly create significant features from records. After per-processing, deciding on the algorithms are applied to the device learning model, used to degree the accuracy of the predictions. Scoring is the process of producing values or rankings based on a skilled device mastering version. The values or rankings which are produced can represent sickness predictions of future values.

4. Dataset

The heart disease data set has been applied for training and trying out functions. However, most effective 14 of them were used because for closing attributes, the values were missing. We achieve a correct result with a reduced wide variety of features. Additionally, the processing is achieved by means of changing the lacking values of the characteristics (column) by using the columns mathematics suggests, in case of nominal information it's far changed with the mode. List the chosen attributes of the heart disorder data set. The overall performance of all of the classifiers is accessed and their outsource then analyzed based totally on accuracy. Some researchers, however, have used the Cleveland, Hungarian and lengthy-bench-through framingham data set together with 14 attributes, which alongside the values and their viable statistics sorts are described

5. Performance Metrics

Consequently, the performances of the models were calculated. The confusion matrix statistics are used for evaluation of model performances.

Table 1: Actual Class versus Predicted Class

Actual class\ Predicted class	Class 1: Predicted	Class 2: Predicted
Class 1: Actual	TN	FP
Class 2: Actual	FN	TP

Table 2: Definition of the Terms

Positive (P)	Observation is positive
Negative (N)	Observation is not positive
True Positive(TP)	Observation is positive, and is predicted to be positive.
False Negative(FN)	Observation is positive, but is predicted negative.
True Negative(TN)	Observation is negative, and is predicted to be negative.
False Positive(FP)	Observation is negative, but is predicted positive.

Table 3: Performance Measuring Factors

PERFORMANCE METRICS

Measure	Formula
Accuracy	(TP+TN) / (TP+FP+FN+TN)
Recall	TP / (TP+FN)
Precision	TP / (TP+FP)
F1 score	2TP / (2TP+FP+FN)

By using all these features we can able to calculate the most efficient model.

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

Precision	Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives. Said another way, "for all instances classified positive, what percent was correct?"
Recall	Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.
F1score	The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.
Accuracy	Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right

6. Classification Techniques

In our work, we are focusing on implementation of (NB) Naive Bayesian, (DT) Decision Tree, (KNN) K-nearest neighbor, (ANN) Artificial Neural Networkstechniques.

A. Decision Tree (DT) Algorithm

Decision Tree is articulated with the conditional grouping. It produces a versioning with the most relevant 14 attributes. Attribute with rank 1 is positioned as the root node, other different attributes form Rank-2 represent the intermittent nodes. A decision is made at every node and the leaf node gives us the final result.

Decision Tree Pseudo code:

Input an attribute-valued data set D

1. Tree = {}
2. if D is "pure" OR other stopping criteria met then
    a. terminate
4. end if
5. for all attribute a € D do
    a. Compute information-theoretic criteria if we split on a
7. end for
8. abest = Best attribute according to the above-computed criteria
9. Tree = Create a decision node that tests abest in the root
10. Dv = Induced sub-datasets from D based on abest
11. for all Dv do
    a. Treev C4.5(Dv)
    b. Attach Treev to the corresponding branch of Tree
14. end for
15. return Tree

B. Naive Bayes (NB):

Naive Bayes classifier is analytically dependent on total classifier Bayes Theory. It is assumed that the attributes are statistically and analytically independent. A Naive Bayes Classifier can be a term addressing a simple probabilistic type supported by using making use of Bayes theorem. In another clean phrases, a Naive Bayes classifier predicts that the absence (or presence) of a particular attributable of a category is unrelated to the absence (or presence) of the other characteristic. As an instance, the toy could also be thought of to be an ball, if it is crimson, round and has re-bounce capability. Even supporting those options rely upon the lifestyles of the alternative, a Naive Bayes classifier considers all of these to independently contribute to the likelihood that this toy is an ball.

Naive Bayes Pseudo code:

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

Given training data X, the posterior probability of a

hypothesis H, P(H|X), follows the Bayes theorem

P(H|X)=P(X|H)P(H)/P(X) (1.1)

1. Each data sample is represented by an n-dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

2. Suppose that there are m classes, C1, C2. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

P(Ci/X)>P(Cj/X) for all 1< = j< = m and j!= i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posterior hypothesis. By Bayes theorem,

3. As P(X) is constant for all classes, only P(X| Ci)P(Ci) needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally

likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would, therefore, maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by

P(Ci) = si/s, where Si is the number of training samples of class Ci, and s is the total number of training samples. on X. That is, the naive probability assigns an unknown sample X to the class Ci.

C. K-Nearest Neighbors (KNN) Algorithm:

The k-Nearest Neighbors algorithm is an uncomplicated algorithm to understand and to implement, and yet a powerful tool The model for KNN is the whole training data set. When a prediction is needed for an unseen statistics, the KNN set of rules will search through the education data set for the k-most similar times. The prediction attribute of the maximum comparable times is consolidated and lower back as the prediction for the unseen example.

K-Nearest Neighbor Pseudo code:

1. Calculate "d(x, xi)" i =1, 2, ..., n; where d denotes the Euclidean distance between the points.

2. Arrange the calculated n Euclidean distances in non-decreasing order.

3. Let k be a +ve integer, take the first k distances from this sorted list.

4. Find those k-points corresponding to these k distances.

5. Let ki denotes the number of points belonging to the ith class among k points i.e. k 0

6. If ki¿kj I j then put x in class i.

D. Artificial Neural Network Classifier (ANN):

For heart sickness predictions ANN technique is used with back-propagation. During the training phase, forward propagation is completed in an impartial network. After the forward pass output fee is generated at the output layer nodes. During the forward pass, to begin with, general entry to the node is calculated, and then the output of the node has calculated the usage of the activation function. In feed-forward ANN, the neurons receive several inputs; the neuron total input is calculated using formula:

Total Input = n1.w1+n2.w2+...+nm.wm+1wb

Where:

n1, n2,.....,nn - Input neurons

w1,w2,.....,wn – Weights associated with input neurons

wb – Weight associated with bias

Output of neuron is calculated using activation function Activation function=1/(1+e(−Total Input))

ANN Pseudocode:

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

```
input: (i) The original similarity matrix,M, between
           two ontologies/schemas;
       (ii) A set of training examples.
output: The learned weight vector
1  Initialization of w: wi <- 0.25;
2  for i <- 1 to a predefined iteration number do
3  Save training examples to a temporary variable;
4  AWi<--0;
5  while Training examples are not empty do
6  d<--GetCurrentTrainingExample();
7  r<--ObtainRowNumberinMatrix(d);
8  c<--ObtainColumnNumberinMatrix(d);
9  Od<--CalculateNetworkOutput(d);
10 tr<--FindMaximumSimilaritylnRow(r);
11 tc<--FindMaximumSimilaritylnColumn(c);
12 Awi<--Awi+n[(tr - od)+(tc-od)]sid;
13 RemoveCurrentTrainingExample();
14 end
15 Wi<--wi+Awi;
16 Restore training examples from the temporary
   variable;
17 end
18 output updated  wi;
```

7. Module Designed

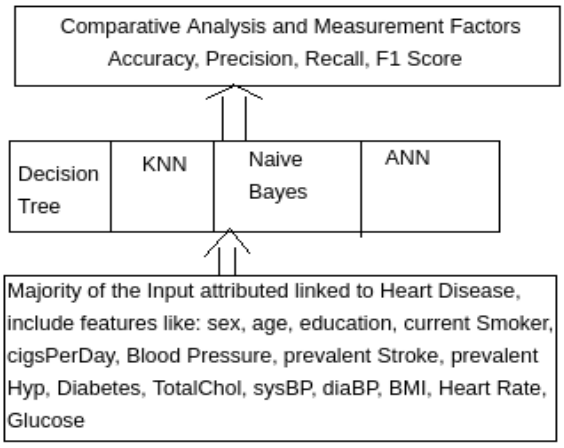

Figure 1: Proposed Module in Predicting Heart Disease

Classifiers namely DT, NB, KNN, ANN furthermore, mixes of these classifiers, utilizing outfit learning techniques, for example, stowing, boosting and stacking, are talked about. In every situation, the exhibition is determined to utilize the standard measurements, to be specific exactness, accuracy, precision, recall, particularity and measure.

Here, we collect the heart disease data set which has some of the features and we apply the algorithms for the data-set and find the accuracy of the data using precision and recall. We can refer to the above diagram.

8. Block Diagram

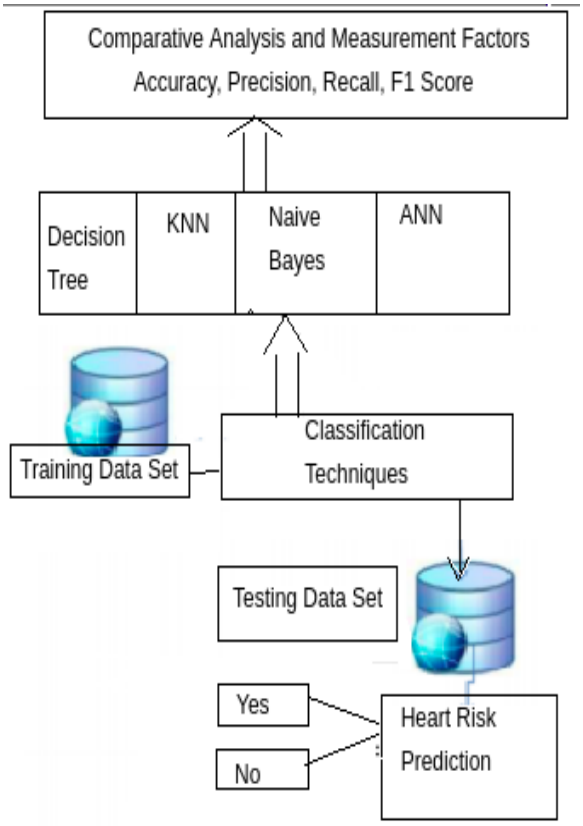

Figure 2: System Architecture using various Classification Techniques

To design and develop heart disease prediction using (ML) Machine Learning for assuming the occurrence of Heart Disease in the future and to check and compare the performances of algorithms.

- To collect large volumes of the medical data set and to reconstruct the missing data.
- To analyze the data using ML and deep learning algorithms.
- To predict if the person has heart disease.
- Performance Comparative of various algorithms (Naive Bayesian, KNN, Decision Tree, ANN).

Using the classification technique we input the data from the user and the user specifies one of the algorithms to predict the disease and the algorithm is applied and predicts that the human is having a heart problem has or not, we can refer to the below diagram.

9. Data Flow of Classification Techniques

A. Decision Tree:

It is a sort of managed learning, that is used for the most part utilized for grouping issues. Shockingly, it works for consistent attribute-valued factors. Right now, split the

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

given data into at least two comparable groups. This is done depends on the most critical properties/free factors to make as unmistakable gatherings as could be expected under the circumstances.

The input data-set is split into testing data and training data, the training data will be allocated for the learning algorithm and processing of the data ., however the testing data will assess the model and find the accuracy for the given model.

B. Naive Bayes:

The Naive Bayes algorithm is a basic prospect classifier that processes a large amount of probabilities by adding value regarded blends of the given input data-set. The utilization of the Naïve Bayes technique is required to have the option to foresee judge the heart sickness. 4241 records of information are used in testing data by the Naïve Bayes technique, at that point acquired a level of 83.6% for the exactness of expectation.

The necessary equations are given below:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\partial = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\partial}e^{-(x-\pi)^2}$$

Steps involved:

- Transform/ Create recurrence table from the informational index.
- Create a probabilistic table by finding each of their probability.'
- Now, utilize the Naive Bayesian condition to calculate the probability for each class. The class with the most noteworthy back likelihood is the result of the forecast

C. K-nearest Neighbor:

Another category based KNN technique is introduced right now. It pre-processes preparing information by utilizing grouping, at that point order with another KNN calculation, which receives a unique change in every emphasis for the local number K. This strategy would keep away from the asymmetrically characterization marvel and decrease the misconception of the limit testing tests. We have a test in and the outcome shows that it has great execution.

Euclidean distance measure is described below:

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

Steps Involved:

- Calculate the distance
- Find the closest neighbors
- Vote for the labels

D. Artificial Neural Network Classifier:

Counterfeit Neural system calculation appears to abstain from pruning issues and has higher effectiveness and exactness in preparing the datasets. Additionally the utilizing of the Decision tree is because they are anything but difficult to decipher, comprehend and have non-direct qualities between values. This holds well the presentation of the tree built which gives better yields. Adjacent to all the applications created utilizing AI in day to day life, the utilization of such learning calculations in such heart diseases will upgrade and advantage the clinical field.

Neurons are broken down to multiple layers: input, intermediate and output. The information layer is made not out of full neurons but instead comprises just of the record's qualities that go about as contributions to the following layer of neurons. The following layer is the concealed layer. A few shrouded layers can exist in one neural system. The last layer is the yield layer, where there is one hub for each class. A solitary clear forward through the system brings about the task of an incentive to each yield hub, and the record is allocated to the class hub with the most noteworthy worth. The Neural system is utilized to take care of issues in the manner that a human would. It has discovered application in a wide assortment of issues. These range from work portrayal to design acknowledgment.

10. Results

Various machine learning classification algorithms can be used for predicting Heart Problems. Five various different kind of machine learning algorithms are used in our work, all of them give different ac curacies. The algorithms used are KNN, decision tree, naive Bayes and ANN, below the detailed output of each algorithm are shown:

A. Decision Tree Output:

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

```
# Decison Tree: Accuracy for test data

acc_dt = accuracy_score(y_true=y_test, y_pred= y_pred)
print("Overall accuracy of DT model using test-set is : %f" %(acc_dt*100))

Overall accuracy of DT model using test-set is : 76.411658

# Confusion Matrix

print(confusion_matrix(y_test, y_pred))

[[787 144]
 [115  52]]

# Decision Tree: Precision, Recall, F1-score, Support

print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.87      0.85      0.86       931
           1       0.27      0.31      0.29       167
```

Figure 3: Decision Tree Output represent Confusion Matrix and its corresponding accuracies

B. Naive Bayes Output:

```
# Naive Bayes: Accuracy for test data

acc_nb = accuracy_score(y_true=y_test, y_pred= y_pred)
print("Overall Accuracy of NB model using test-set is : %f" %(acc_nb*100))

Overall Accuracy of NB model using test-set is : 83.060109

# Confusion Matrix

print(confusion_matrix(y_test, y_pred))

[[866  65]
 [121  46]]

# Naive Bayes: Precision, Recall, F1-score, Support

print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.88      0.93      0.90       931
           1       0.41      0.28      0.33       167
```

Figure 4: Naive Bayes Output represent Confusion Matrix and its corresponding accuracies

C. K-nearest Neighbor Output:

```
KNN: Accuracy for test data

acc_knn = accuracy_score(y_true=y_test, y_pred= y_pred)
print("Overall Accuracy of KNN model using test-set is : %f" %(acc_knn*100))

Overall Accuracy of KNN model using test-set is : 82.786885

# Confusion Matrix

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test, y_pred))

[[898  33]
 [156  11]]

# KNN: Precision, Recall, F1-score, Support

print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

           0       0.85      0.96      0.90       931
           1       0.25      0.07      0.10       167
```

Figure 5: KNN Output represent Confusion Matrix and its corresponding accuracies

D. Artificial Neural Network Classifier Output:

```
ANN: Accuracy for test data

accuracy = accuracy_score(Y_test, Y_pred)
print("Accuracy of ANN: %.2f%%" % (accuracy*100))

Accuracy of ANN: 83.97%

ANN: Precision, Recall, F1-score, Support

print(classification_report(Y_test, Y_pred))

              precision    recall  f1-score   support

         0.0       0.85      0.99      0.91       463
         1.0       0.33      0.02      0.04        86

Confusion Matrix

print(confusion_matrix(Y_test, Y_pred))

[[459   4]
 [ 84   2]]
```

Figure 6: ANN Output represent Confusion Matrix and its corresponding accuracies

11. Comparative Results

This paper focuses on comparing the execution of the best classification algorithms in the prediction of coronary heart disease, and identifying the most efficient algorithm. As part of Input data set, 14 attributes are used, As per our study, ANN provides more accurate results when compared to KNN, NB and DT Classifiers.
Table 4: Table showing accuracy (in %), precison, recall,

May-June 2020
ISSN: 0193-4120 Page No. 4168-4175

F1 score of various classifiers

	DT	NB	KNN	ANN
Accuracy	76.41	83.06	82.79	83.97
Precision	0.87	0.88	0.85	0.84
Recall	0.85	0.93	0.96	1
F1 score	0.86	0.9	0.9	0.91

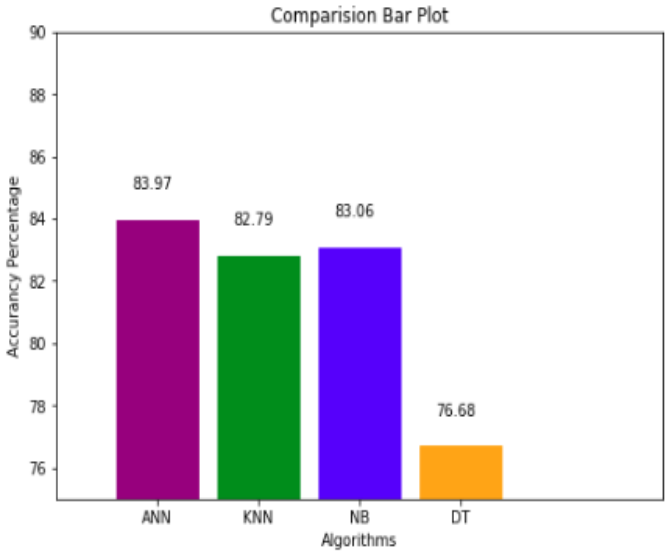

Figure 7: Bar plot comparison of Accuracy across various classifier techniques

12. Conclusion

The coronary heart disease has ended up as one of the primary issues. Major parts of the world are suffering from heart disease due to their life style habits. With a majority of these predominant factors prediction of cardiovascular diseases becomes are very important and the models that are expecting coronary heart might convey a major impact in lowering the mortality rate. Prevention is continually the primary step to stop any disease and subsequently that need to be considered a chief difficulty and worked upon. Studying machine learning algorithms affords a suitable software environment to work on prediction and predicts the ailment kind for that reason. Heart prediction works in first-class with ANN with the highest accuracy of 83.97% in comparison to the alternative three algorithms namely decision tree, Naive Bayes and KNN.

References

[1] Nilima Karankar, Pragya Shukla and Niyati Agrawal, Comparative Study of Various Machine Learning Classifiers on Medical Data, 2017 7th International Conference on Communication Systems and Network Technologies, 978-1-5386-1860-8/17/$31.00 ©2017 IEEE 267 DOI 10.1109/CSNT.2017.51.

[2] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez, A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017, 978-1-5386-1629-1/17/$31.00 ©2017 IEEE.

[3] S. M. M. Hasan, M. A. Mamun, M. P. Uddin and M. A. Hossain ,Comparative Analysis of Classification Approaches for Heart Disease Prediction, 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 10.1109/IC4ME2.2018.8465594.

[4] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in IEEE/ACS International Conference on Computer Systems and Applications. IEEE, 2008, pp. 108– 115.

[5] M. Raihan1, Parichay Kumar Mandal, “Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network”, 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019, 978-1-5386-9111-3/19/$31.00 ©2019 IEEE