International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Prediction of Heart Disease using Naïve Bayes Technique of Data Mining

Arshdeep kaur, Anil kumar

Abstract: ***Coronary illness is responsible for deaths in all age groups and is common among males and females. An excellent answer for this issue is to have the option to predict what a patient's health status will in future so the specialists can begin treatment much sooner which will yield better outcomes. Data mining plays most significant role in area of investigation by means of the objective to finding essential data from massive amount of information. Currently, data mining strategies and tools are utilized by researchers in the field of healthcare, especially for prediction of sickness. Data mining methodology affords improvement approach to interchange huge data into beneficial information for attaining selection. In utilising data mining patterns they desires considerably fewer amount of funding intended for the forecasting the ailment alongside better accurate and precision. Moreover, analysis of study paper depicts the estimation of coronary illness in clinical field by utilizing data mining. Various popular data mining algorithm on the dataset of 13 attributes is applied to forecast the coronary ailment at initial stage. The dataset is collected from UCI machine learning repository and analysed with various parameters like Accuracy, Recall, Precision, F-measure, ROC area and Kappa statistics. Experimental results show that the Naïve bayes algorithm is always becomes the best-performing data mining method which accomplishes an accuracy of 86.716% in coronary illness prediction.***

Keywords. : Weka, Heart disease, Data mining, Naïve bayes

I. INTRODUCTION

The important tough-running body part is heart. Most importantly the circulatory system, in which blood is moved or circulates through veins in heart. The heart plays an essential function as per it transports blood, oxygen, and different resources all over the human body. If the heart is no longer response properly that means it cause critical well-being situations comprising death. The outcomes in numerous infection, incapacity or dying. Adjustable chance factors contain more weight, smoking, lack of exercise and so on. If the heart doesn't feature properly, this could distress the alternative components of the human system which include brain, kidney and so forth. Heart ailment is a form of ailment which results the functioning of the heart. In these day's technology coronary heart ailment is the primary reason for deaths. WHO has predicted that 12 million people die each year due to heart sicknesses. Some heart illnesses are cardiac, coronary illness, and knock. Knock is like a sort in coronary disorder that takes place due to strengthening, blocking off or lessening of blood vessels which force via the brain or it could also be initiated by means of high blood strain.Illness analysis plays a main function in medical field. Intellectual data mining techniques address trouble in medical dataset forecast related to numerous ideas.

Different person body can show extraordinary symptoms of coronary heart ailment which may range accordingly. Though, they regularly encompass back pain, jaw pain, Even though heart disorder is recounted as the ultimate chronic kind of disease in the arena, it can be maximum avoidable one also on the same time. A wholesome manner of existence (important prevention) and well timed evaluation (inferior prevention) are the two predominant origins of coronary heart disease director. Conducting consistent take a look at-ups (inferior prevention) shows outstanding function within the judgment and early prevention of coronary heart disease problems. Several checks comprising of angiography, chest X-rays, echocardiography and exercising tolerance take a look at help to this large problem. Nevertheless, those tests are high priced and involve availability of correct clinical system, neck pain, stomach problems and tininess of breath, chest pain, arms and shoulders pains. There are a variety of different coronary heart diseases which incorporates coronary heart failure and stroke and coronary artery ailment. As This paper forecast several heart disease prediction via using process of data mining recommended in current years.

II. PROBLEM STATEMENT

The difficulty towards envisage heart ailment training taken away massive extent about information beyond several biasness amongst classes. With the help of classification algorithm it has finished by performing comparable observation. In multiclass classifier method, naïve bayes taken as base classifier. The heart ailment dataset comprising 13 attributes and 270 instances.

III. RELATED WORK

There have numerous investigations in the field of executing machine learning procedures in medical management Machine learning in health maintenance has been one of the top needs for the specialists.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
Arshdeepkaur, Computer Science and Engineering ,Guru Nanak Dev University,gurdaspur, india. Email:arshdeepkaur977@gmail.com
Anil kumar*, Computer Science and Engineering ,Guru Nanak Dev University,amritsar, india. Email: anil.gndu@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Utilizing different data mining methods and learning about the hidden examples, bits of knowledge can be separated. These experiences can additionally be utilized for sickness and disease expectation. Datasets from the effectively available repositories can be utilized for preparing the machine. Data pre-handling or data cleaning is one the significant perspective to be completed before executing machine learning methods forming purposes. We have to systematize or standardize the information for a superior knowledge. For this, lots of papers are readout in relevant to data mining in health field.

The author [1] survey various data mining tools for coronary disorder forecasting or analysis in health field or domain. A heart contamination expectancy version makes use of algorithms of data mining to guide docs in forecasting heart ailment on the idea of medical records. In medical field the various data mining methods are used like Naïve Bayes, decision tree and so forth, allows in improved result. Algorithms amassing support toward making choices extra correct or faster.

Data mining strategies along with clustering, class, regression, association rule mining, CART (Classification and Regression Tree) are broadly utilized in medical field. The fundamental emphasis [2] is to investigate data mining approaches essential for scientific facts , mainly find out regionally common ailments such as heart illnesses, lung tumor, breast most cancers and so forth. For this reason data mining methods are used.

A web based totally software has added in [3] using Naïve Bayes algorithm that take signs as of consumer and provided prognosis outcome to person and affected person.

For predicting diabetes disorder on weka tool, author's research perspective [4] had offered a evaluation between Naïve bayes algorithm and decision tree algorithm and finished machine accuracy of about 79.56% and 76.96% respectively. The author [5] compare different classification model in predicting coronary heart sickness using decision tree, neural network, and Naive Bayes. Through the comparable check out, amore advanced hybrid version can be design in predicting coronary ailment.

In this paper author[6] proposed the enhanced records mining set of rules for healthcare utility. This proposed method is used to expect the coronary heart sickness prognosis of person disorder patients because of the opportunity of spreading to excessive-danger signs and symptoms in medical fields. The author confirmed a language version-like approach for predicting excessive-hazard diagnosis from analysis histories of sufferers the use of recurrent neural networks (RNNs), i.e., analysis estimate the usage of RNN .The suggested PP-RNN practices more than one RNNs for gaining knowledge of from judgment code arrangements of sick person so that you can are expecting occurrences of excessive-threat illnesses.

The author discussed [7]. Heart Ailment analysis Using Data Mining Method In this paper, using only 14 attributes example age , gender, bmi,sugar , down sloping and fat. Less accuracy.

In this paper author [8] have been used hybrid attribute choice technique for forecast of coronary sickness.

The author [9] has studied forecast structures of coronary ailment analysis in the usage of more range of enter attributes. The system uses extraordinary attributes which includes intercourse, blood pressure, cholesterol stage, diabetes, pulse rate, etc like attributes for the prediction of the possibility of sick person getting a Heart disease. The strategies used for information mining category, particularly , Naïve Bayes, Decision tree and Neural Networks had been studied on database of coronary sickness.

The success applications of data mining in numerous fields like e-commerce, retail and advertising and marketing has caused its software in diverse other sectors. Among those sectors a developing area is healthcare. The healthcare surroundings is wealthy in records but having a lack of knowledge. There is a wealth of statistics to be had in the healthcare structures. But due to absence of powerful evaluation tools to find out unknown associations in facts the records is not correctly utilized. The author [10] aims to offer a analysis on green techniques at gift of understanding finding in databases the usage of data mining that may be efficiently utilized in these days scientific study primarily in guess of heart disease. Various experiments have been performed to associate the overall achievement of analytical data mining approach on the equal datasets attributes and the consequences of those experiments display that Decision Tree achieves well erstwhile Bayes algorithm also shows with an exactness as of decision tree but dissimilar analytical techniques like Neural Networks, entirely based on clustering and KNN do now not perform nicely as of decision tree. Another thing which may be concluded from the effects is that the accurateness of the Bayes algorithm and Decision Tree might be better afterwards applying genetic algorithm to it in order that the real records length get decreased to get the best subclass of attributes which is sufficient for coronary disorder forecast. The major objective in this paper [11] to use the data mining method namely, Naive Bayes in order to develop intellectual system. A consumer friendly wed primarily based machine. In this system a consumer responses to already defined queries. It repossesses the secreted statistics commencing the previously kept database after which matches consumer answers by educated facts sets to analyze the end result. They can reply to complicated questions for the prediction of heart illness and as a consequence facilitates medical specialists to make intellectual choices which cannot be finished via conventional selection support structures. Other gain of this machine is that by means of offering effective treatments at early stage, it also helps to reduce remedy expenses.

He conferred data mining methods, plus also various special examines which might be on going and useful to medical analysis of ailment. The have a look at found out that relying at the kind of dataset that for every model be different in their execution. Dataset includes unlabeled instructions , at that time the clustering perform better for pattern reputation the various procedures, are followed via studies owed to its ease[12]Data mining performs a dynamic role to forecasting infections inside fitness precaution enterprise [13]. He go through revisions papers that specifically targeting guessing coronary disorder, Breast most cancers and Diabetes.

Here changed into have a look at of Naïve Bayes, Decision tree algorithm and k-nn algorithm for analysis the coronary ailment dataset. In order to classify the dataset, the Tanagra data mining tool is utilized. The classified dataset is calculated the use of 10 fold cross validation and the outcomes are in comparison. The dataset is divided into testing and training i.e. 70% of information is used for training and 30 % is used for testing.

In this paper author [14] studied several category methodology for forecast of coronary sickness with reduced range of attributes.

The author [15] defined an advanced decision support device the usage of two data mining classifiers Jelinek-Mercer smoothing and naïve bayes for coronary ailment prediction.

In [16] Naive Bayes, Decision tree algorithms particularly used for liver ailment detection with 10 functions. The end result evaluation with appreciate to accuracy NB Tree set of rules has the highest accuracy while with appreciate to computational time Naive Bayes set of rules performs higher.

The author[17] proposed approach in which he withdraw the features as of the dataset. In view of the highlights the decision table is built. Unimportant characteristics are expelled by applying features selection algorithm. Further, the reliance among the attribute towards recognizing the illness is controlled by utilizing optimality model function. Subsequently the time taken to forecast the coronary illness is decreased contrasted with different algorithms. The dataset is gathered from UCI and investigated utilizing the Optimality Criterion Feature selection algorithm.

In This paper (18) author illustrate the probability of heart sickness in health field by utilizing data science. The identical analysis about done study recognized with that issue however the accuracy of expectation be supposed to have been better. beside these lines, this study centers around include feature selection methods and techniques where various heart sickness datasets are utilized for study and to display the exactness enhancement. By utilizing the Rapid miner as tool; Logistic Regression, Decision Tree, Random Forest, Naïve Bayes algorithms are used as feature selection method and enhancement is occur in the result by determine the accuracy.

In This paper author (19) examines a technique named ensemble classification, that utilized for correcting the precision of feeble algorithm for consolidating different classifiers.. A relative logical methodology was applied for correcting the expectation exactness of coronary illness using ensemble technique. The focal point of this paper isn't just expanding the accurateness of feeble classification algorithms, yet additionally execute the algorithm using heath or clinical dataset, to demonstrate its utility to predict sickness at a beginning time. As outcomes of this investigation demonstrate that ensemble methods, for example, and boosting and bagging, are viable for enhancing the forecast precision of feeble classifiers, and show acceptable execution for distinguishing danger in coronary illness.

IV. HEART DISEASE PREDICTIONS RESULTS AND ANALYSIS

As various research paper are available for mining of the datasets which applies several algorithm regarding exclusive information. Hence, for the innovative and improved evaluation methods, at that time here a requirement for an expert to understand whichever set of rules implements nice designed for a specific kind of dataset

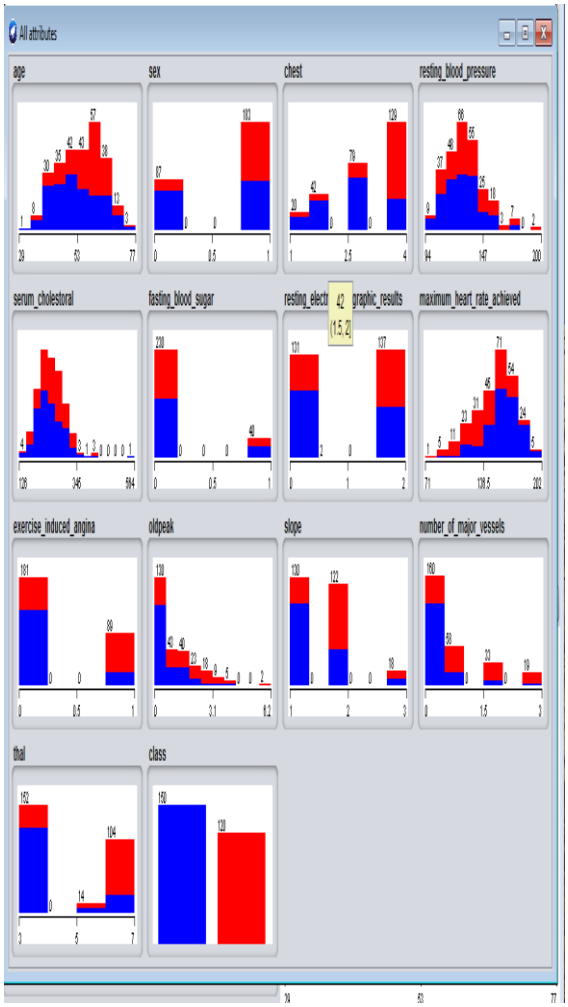
- Dataset Collected:

The database intended for the study work pick up from the StatLog dataset in UCI repository. It incorporates 13 attributes. The coronary illness dataset remembered for the study task contains of 270 occurrences without any lost attributes. The dataset is usually utilized for unrelated sorts of coronary illness, for example, normal angina, non-angina abnormal anginal. The study work is likely to predict the coronary ailment nonessential of the ailment types. The attribute is always taken as numeric data type that signifies age of sick person and extents from 28 to 67 years. The attribute cp for deciding the pain, signified in the range1-4. The trestbpd is a resting blood pressure which is somewhere in the range of 9oand 110; The fasting sugar level whichever a 0or 1 signifying to Boolean values false or true.The thalach is the highest heart rate which is ranging from 81 to 184. The attribute exang is the activity provoked angina which is Boolean value. The target class of the dataset is illness that indicates the coronary sickness nearness no or yes.

Attribute information:

@attribute age{<29,29-64,>=65}
@attribute sex {male, female}
@attribute chest pain { average,abnormal angina non-angina pain asymptomatic}
@attribute resting blood pressure { mm /Hg }
@attribute cholestoral mg/dl
@attribute fasting blood sugar {1 = true; 0 = false}
@attributerestingelectrocardiographicresult{normal, abnormal}
@attribute maximal heart rate attained
@attribute exercise induced angina { 0 = no;,1=yes}
@attribute ST despair induced by work out comparative to have a rest
@attribute slope of the highest exercise{flat,upslope}
@attribute number of major vessel
@attribute thal:{4 = normal; 7 = fixed defect; 9 = reversable defect}
diagnosis of heart disease {angiographic disease status}

- Investigation through Weka programming: The data values that entered to the Weka software as training and afterward testing. At that point we pick our proposed classifiers individually and note down classifier outcome for each.
- Comparative investigation of classifiers: Based on execution factors, we have correlated the performance of all classifiers.



V. DATA MINING TECHNIQUES:

i. Bayes net:

Bayesian Network or essentially Bayes algorithm is extremely incredible for illness predict system since it can make results in instance of lost input data. This classifier utilizes a directed chart model to tell associations among different attributes. It utilizes likelihood functions to anticipate the presence or absence of different illnesses.

ii. PART:

PART is the abbreviation for Projective Adaptive Resonance Theory. It is a neural network created by Cao. The PART algorithm is appropriate for extraordinary structural datasets. Most importantly the PART grid based network lies in the closeness of a unknown layer of neurons, that figure the dissimilarities amongst yield and input neurons, and work on lessening the likeness contrasts.

iii. J48:

J48 is a Decision tree that is a usage of ID3 (Iterative Dichtomiser 3) created by the WEKA project group. R language additionally has a package to execute this. J48 doesn't require discretization of numeric attributes.

iv. Decision tree:

A Decision tree is to facilitate implementation device that usages a graph or model of choices and their potential results including chance occurrence results and utility. It is one of the approaches to show an algorithm. Decision trees are usually utilized in operation examination, explicitly in decision analysis to help and recognize a technique that will doubtlessly arrive at the objective. A Decision tree can without much of a stretch be changed to a lot of rules by mapping as of the root node to the leaf node individually. At long last by adhering to these rules, suitable conclusions can be reach.

v. Random forest

Random Forest is an individual from decision tree algorithm family. It alters the past work of decision trees in building the characterization trees. Random Forest splits nodes utilizing the best among a subset of predictors which picked at random. The algorithm steps are as per the following. To start with, draw the ntree bootstrap tests. At that point, for each example, grow an unpruned grouping tree: at every node, randomly test m attempt of the predictor and pick the best split from those factors. At that point, the prediction can be made by gathering the expectation of n tree trees with larger votes . From some examination, Random Forest has a few points of interest like: can be utilized in multi-class issues, great predictive execution.In any case, some way or another, Random Forest isn't excessively oftentimes utilized in medical microarray examines.

vi. Simple logistic

Simple logistic is a predictive algorithm we can utilize when the variable we need to predict is categorial, which means having two classifications, and they can be either numerical or categorial. It approximates likewise the probability that an occurrence will happens for an arbitrary chose observations versus the likelihood that an occurrence doesn't happens.

vii. Zero R

ZeroR is the easiest classification technique which depends on the target and disregards all predictors. ZeroR classifier essentially predicts the bulk part category (class). In spite of the fact that there is no consistency power in ZeroR, it is helpful for deciding a pattern execution as a benchmark for other classification techniques.

viii. K-star

K-star is only the K-Mean based clustering. Moreover it is a method to discover attractive examples with regards to a particular dataset. k-means algorithm is a progressive algorithm that picks up its title as of technique of operation. The clusters mean is then more analyzed and the procedure starts once more. In k-means clustering beginning centroid selection utilizing inliers technique is applied to discover the outcome.

VI. PROPOSED WORK

In our proposed work, naïve bayes is particularly used for prediction of heart disease. The whole prediction is based on the patient heart disease database in which we use 13 attributes. Moreover, the Naïve Bayes classifier , is based on the Bayes theorem.

It is an exceptional instance of the Bayesian system, and it is a likelihood based classifier. In the Naïve Bayes network, all features are conditionally independent. The adjustments in a single component subsequently doesn't influence another component. The Naïve Bayes algorithm is reasonable for grouping structural datasets. The classifier algorithm use conditional independence. Conditional independence accept that values of an attribute which is not depend upon the other attributes in a class. The objective of arrangement is to accurately predict the estimation of an assigned discrete class variable given a vector of indicators or attributes. As in our paper we use WEKA tool for performance of data mining algorithms in order to predict the heart patients and results are discussed below.

VII. RESULT

i. Predict Accuracy

Consequent to performing WEKA investigation for nine chosen data mining algorithms, the after effect of every calculation's accuracy is outlined in Table 1

Serial.no	Algorithm	Accuracy
1	Bayes net	82.71%
2	Decision Tree	70.37%
3	Zero R	55.55%
4	J48	77.77%
5	Simple logistic	82.716%
6	PART	76.54%
7	Random forest	79.0123%
8	K-Star	71.604%
9	**Naïve bayes**	**86.716%**

From the outcome, Naïve bayes accompanies the best accuracy of (86.716%) followed by, bayes net (82.71%), Simplelogistic(82.716.%),Randomforest(79.0123%),J48(77 77%),PART(76.54%),kStar(71.60%),Decisiontree(70.37%), and end comes ZeroR(55.55%).

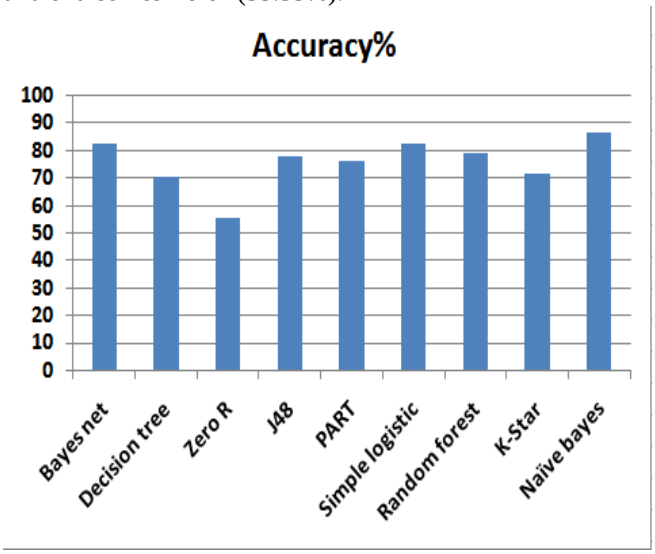


Figure 1: Accuracy of classifiers

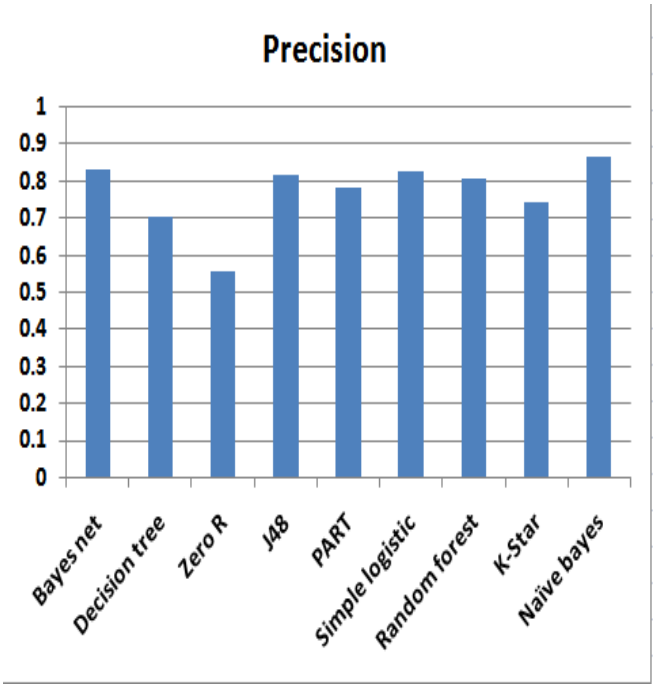


Figure 2: Precision of classifier

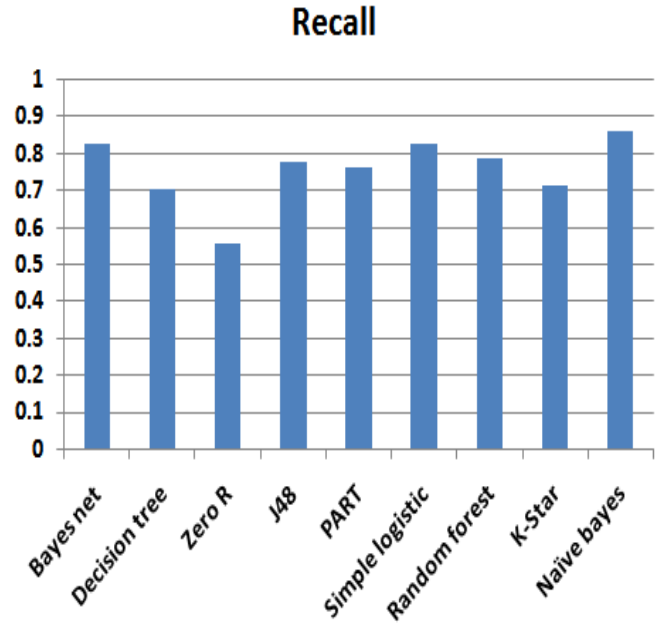


Figure 3: Recall of classifiers

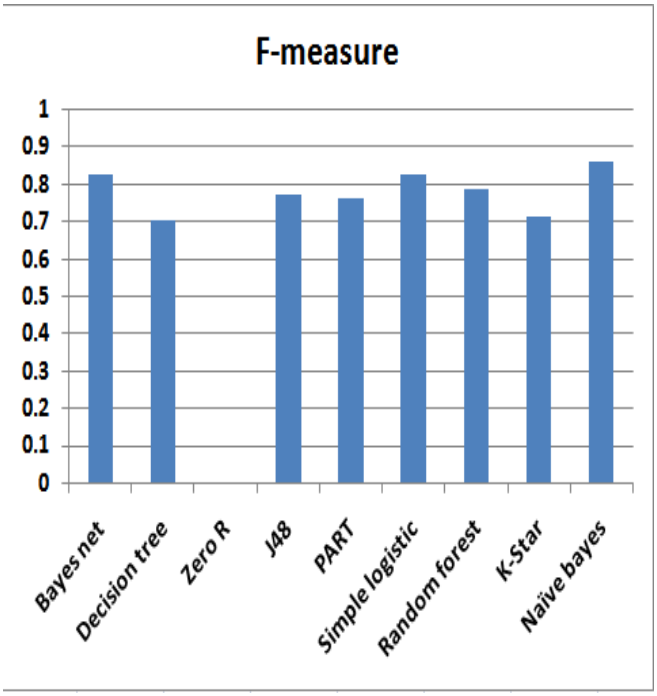


Figure 4: F-measure of classifiers

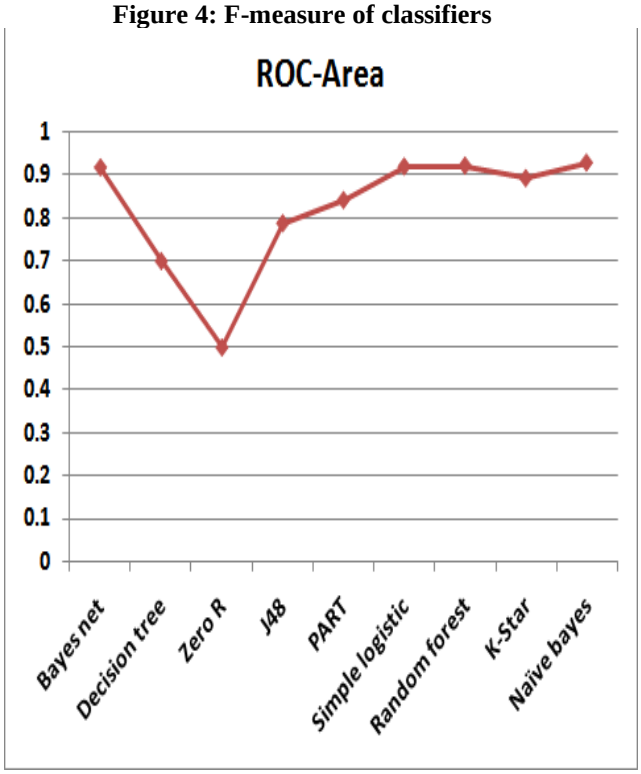


Figure 5: ROC area of classifiers

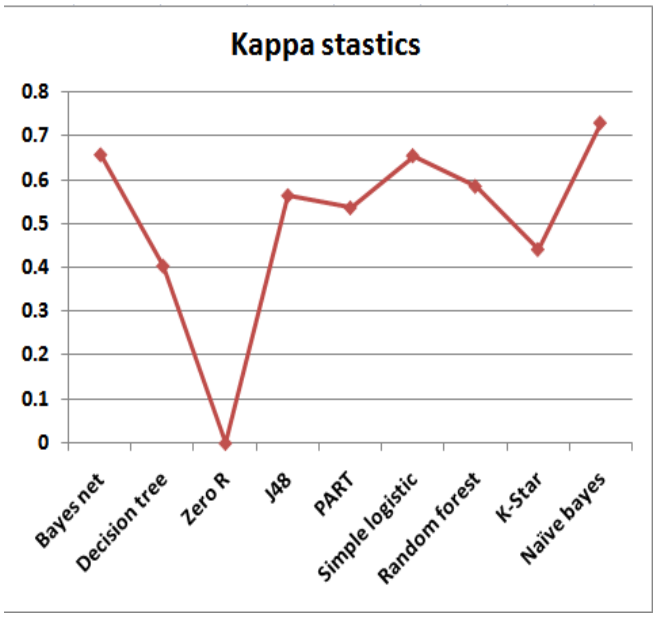


Figure 6: Kappa statistics of classifier

VIII. CONCLUSION

The Naïve bayes algorithm is used to predict more precisely the heart ailment which automatically means to lessen the tests that are required by patients. Heart diseases are one of the predominant medicinal issues in world. This investigation gave knowledge of various data mining methods that can be utilized in computerized coronary illness estimation framework. Coronary illness is one of the main source of death in world and the initial or primary forecast of coronary illness is significant. The models that were utilized to fit the preparation set were analyzed and assessed their exhibition as far as accuracy,
recall, precision and ROC Area. In this paper the best performing classification technique that improve the accuracy of coronary illness prediction were chosen. It is found that Naive bayes (NB) classification based algorithm executed better with most noteworthy accuracy of 86.716%. This work can be stretched out to improve the prediction accuracy utilizing gathering machine learning methods.

REFERENCES

1. Devi Kiruthika .S, S. Krishnapriya and D. Kalita, "Prediction of Heart Disease using Data Mining Techniques,"Indian Journal of Science and Technology, vol. 9, no. 39, pp. 1-5, 2016
2. Khaleel, Mohammed Abdul, Sateesh Kumar Pradham, and G. N. Dash. "A survey of data mining techniques on medical data for finding locally frequent diseases." International Journal of Advanced Research in Computer Science and Software Engineering 3, no. 8 (2013).
3. Bhandari, P., S. Yadav, S. Mote, D. Rankhambe, U. Scholar, and Pune APCOER. "Predictive system for medical diagnosis with expertise analysis." Int. J. of Eng. Sci. and Comput., IJESC 6 (2016): 4652-4656.
4. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint arXiv:1502.03774 (2015).
5. Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques of data mining." In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 520-525. IEEE, 2015.

6. N.Sowri Raja Pillai ,K.Kamurunnissa Bee, J.Kiruthika.:Predictions of Heart Disease Using RNN Algorithm.IRJETVolume: 06 Issue: 03 | Mar 2019
7. Babu, Sarath, E. M. Vivek, K. P. Famina, K. Fida, P. Aswathi, M. Shanid, and M. Hena. "Heart disease diagnosis using data mining technique." In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 1, pp. 750-753. IEEE, 2017.
8. Peter, T.J. and Somasundaram, K., 2012. Study and development of novel feature selection framework for heart disease prediction. International Journal of Scientific and Research Publications, 2(10), pp.1-7.
9. Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47, no. 10 (2012): 44-48.
10. Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17, no. 8 (2011): 43-48.
11. Pattekari, Shadab Adam, and Asma Parveen. "Prediction system for heart disease using Naïve Bayes." International Journal of Advanced Computer and Mathematical Sciences 3, no. 3 (2012): 290-294.
12. Lokanayaki, K., and A. Malathi. "Exploring on various prediction model in data mining techniques for disease Diagnosis." International Journal of Computer Applications 77, no. 5 (2013).
13. Vijiyarani, S., and S. Sudha. "Disease prediction in data mining technique–a survey." International Journal of Computer Applications & Information Technology 2, no. 1 (2013): 17-21.
14. Patel, Shamsher Bahadur, Pramod Kumar Yadav, and D. P. Shukla. "Predict the diagnosis of heart disease patients using classification mining techniques." IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) 4, no. 2 (2013): 61-64.
15. Patil, Rupali R. "Heart disease prediction system using Naive Bayes and Jelinek-mercer smoothing." International Journal of Advanced Research in Computer and Communication Engineering 3, no. 5 (2014): 2278-1021.
16. Alfisahrin, Sadiyah Noor Novita, and Teddy Mantoro. "Data mining techniques for optimization of liver disease classification." In 2013 International Conference on Advanced Computer Science Applications and Technologies, pp. 379-384. IEEE, 2013.
17. Prakash, S., K. Sangeetha, and N. Ramkumar. "An optimal criterion feature selection method for prediction and effective analysis of heart disease." Cluster Computing 22, no. 5 (2019): 11957-11963.
18. Bashir, Saba, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, and Khurram Bashir. "Improving Heart Disease Prediction Using Feature Selection Approaches." In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 619-623. IEEE, 2019.
19. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019): 100203.

AUTHORS PROFILE

Arshdeep kaur, dept- computer science and engineering. I had completed my btech in CSE and currently pursuing mtech in cse in Guru Nanak Dev University Amritsar.

Anil Kumar, dept- computer science engineering, he is assistant professor in Guru Nanak Dev University Amritsar. he had completed many publications in reputed journal and his research areas are data mining ,parallel computing.