International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

THE PREDICTION OF HEART DISEASE USING NAIVE BAYES CLASSIFIER

S. Spino[1], Dr. M. Mohamed Sathik[2], Dr. S. Shajun Nisha[3]

[1]M.phil Research Scholar, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
[2]Principal, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
[3]Assistant Professor & Head, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India

---

Abstract - *Data Mining is a fascinating field of research its significant goal is to discover intriguing and useful patterns from huge data sets. Nowadays, health diseases are increasing day by day due to life style and hereditary. Especially, heart diseases has become prevalent lately. Heart disease is ranked as the major cause of death in the world accounting to about 17.3 million deaths per year. The heart disease records to be the main source of death around the world. Data mining Classification technique predicts the heart disease risk level of each person based on attributes such as age, gender, Blood pressure, cholesterol, pulse rate. This Paper focuses around the prediction of heart disease accuracy value using the Naive Bayes classification technique. Naive Bayes classifier is a statistical based classifier which is based on Bayes Theory.*

***Key Words*: Data mining, Heart Disease, Heart Disease Dataset, Classification, Naive Bayes.**

1. INTRODUCTION

The heart is an imperative organ of our body. The heart functions as a pump in the circulatory system to provide unceasing flow of blood throughout the body. This movement consists of the system circulation to and from the body and the pulmonary circulation to and from the lungs. Blood in the pulmonary circulation serve carbon dioxide for oxygen in the lungs through the process of respiration. If the operation of a heart is not proper, it will influence other parts of a human. Healthy life depends on effective working of the heart. The term Heart disease alludes to disease of heart and the blood vessel system within it. Cardiovascular diseases are one of the most noteworthy flying infections of the Modern world. There are number of elements which increment the Heart disease such as Hyper pressure , Physical inertia , Corpulence, Hypertension, Cholesterol, Family ancestry of heart disease, Smoking. In Classification Algorithm the main objective is to predict the target class by analysing the training dataset [4].Data mining tools have been created for compelling investigation of medicinal data, so as to help clinicians in improving their conclusion for the treatment purposes. In heart disease research, data mining strategy have played out a huge role. The Heart Disease contains the screening clinical information of heart patients [5].From the distinctive elucidation between the healthy people and the heart ailing people. The effectively existing medical data is an obvious and extraordinary methodology in the investigation of heart related infection characterization to discover the disguised medicinal data. Building accurate and efficient classifiers for Medical databases is one of the essential tasks of data mining and machine learning research [7]. The aim of this paper is to employ and analyzed the data mining Classification technique to predict the heart disease risk level in a patient through extraction of interesting patterns from the dataset using vital parameters. Our work present the implementation of Naive Bayes to predict the presence of heart disease in a person. The efficiency of the model is based on accuracy and time complexity.

1.1 Literature Review

Ritika Chandha [1] have defined a specific method that precisely identifies the contingency of heart disease in a patient. For this we have used three methodical algorithms namely Artificial Neural Network, Decision tree, Naive Bayes .It takes Cleveland dataset as its input, these dataset contains 8 attributes. ANN was advocated as an optimistic tool for making decisions with regard to medical applications. The data set is partitioned into testing and training datasets. ANN produced an accuracy of 100% with the provided Data set. Naive Bayes produces a veracious percent of 85.86 08% in foreboding the heart disease. ID3 is one of the effective Decision tree algorithm, the accuracy produced by the ID3 algorithm of Decision tree is 88.0253%. The enumerated results of each algorithms is analyzed to put forth that the ANN renders the highest accuracy. Sujata Joshi[2] have employed data mining technique in diagnosis of heart disease, this research concentrates on using three classification techniques called Decision Tree ,Naïve Bayes and KNN, the datasets are acquired from the UCI Learning Repository ,it consist of 14 attributes . The dataset analyzed using the KNN algorithm produces a precision of 100%.The decision tree technique of the Naive Bayes evaluates to 92.0792% of correctly classified instance .The Analogy of these classification algorithms, resolves to bring forth the KNN algorithm as a prominent method of predicting the Heart diseases. K.Gomathi [3] have proposed a method in which he interprets that an immense amount of heart disease that are deadly to humans beings can be envisaged through their initial symptoms, prior to its manifestation. It would divulge the chance of prevention and early treatments .It is carried out using the Data Mining classification technique. The Dataset played a critical role in diagnosis of the heart disease, the provided dataset are evaluated by three types of Classification techniques namely ANN, Decision tree and Naive Bayes. The estimation done through ANN produces an accuracy of 79.9043% whereas the Naive Bayes generates a precision of 76.555%, and the accuracy

procured, through the estimation done by Decision tree is 97.033%. The enumerated results of each algorithms is analyzed to put forth that the ANN renders the highest accuracy.

1.2 Motivation and Justifications

Naive Bayes is an efficient eager learning classifier, which classification techniques is particularly used for real time applications since it is capable of producing instantaneous results in a modest amount of time, it does not requires large set of data for analogy. The estimation of test data can be carried out with a minimal amount of training data.

1.3 Outline of the Proposed Work:

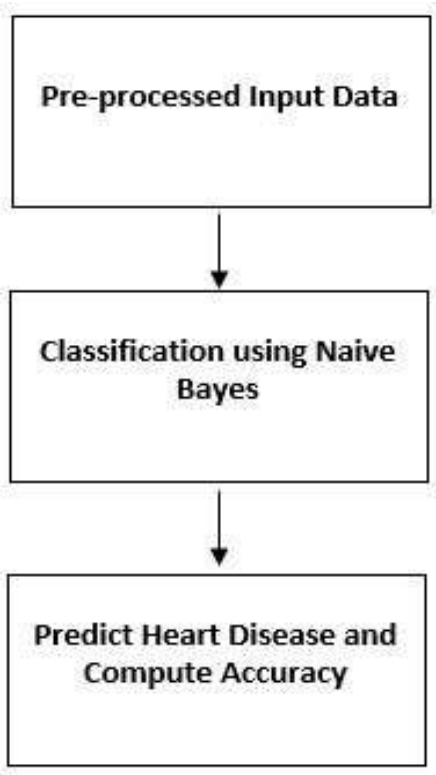

Fig -1: Outline of the Proposed Work

2. METHODOLOGY:

2.1 Dataset:

The Heart Disease data set from the UCI Learning Repository is used for this study. The Heart disease data set is divided into Training data and Testing Data .Training Data consists of 457 records and 13 attributes. Testing Data Consists of 88 records and 13 attributes. The Heart Disease contains the screening clinical information of heart patients.

2.2 Naive Bayes Classifier

Naive Bayes classifier is a statistical based classifier which is based on Bayes Theory. It assumes that attributes are statistically independent. This classifier is based on probabilities. Given two events A and B, P (A) is prior probability and P (A|B) is posterior probability, then according to Bayes theorem.

P (A|B) = P (B/A) P (A)/P (B) and P (B|A) is computed as P (A ∩ B) =P (A)

These Bayesian probabilities are used to determine the most likely next event for the given instance given all the training data. Conditional probabilities are determined from the training data. The Naive Bayes model is based on the conditional independence model of each predictor give the target class [6]. This classifier yields optimal prediction (given the assumptions). It can also handle discrete or numeric attribute values.

3. EXPERIMENTAL RESULT

3.1 Performance Metrics

3.1.1 True Positive Rate

The true positive rate also mention to sensitivity or recall, is used to evaluate the rate of actual positives which are precisely identified.

$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

3.1.2 False Positive Rate:

The false positive rate is used to evaluate the rate of actual negative which are precisely identified.

$$\text{False Positive Rate} = \frac{TP}{FP+TN}$$

3.1.3Precision

Precision refers to how close rate from dissimilar samples are to each other. For example the excellence error is a measure of Precision. When the standard error is small, estimates from dissimilar samples will be close in value.

$$\text{Precision} = \frac{TP}{TP+FP}$$

3.1.4 Recall

Recall is called sensitivity in binary classification. It can be viewed as the probability that an applicable document is retrieved by the query.

$$\text{Recall} = \frac{TP}{TP+FN}$$

3.1.5 F-Measure

The F-Measure (or F-score) examine both Precision and recall of the test to compute the measure .The Precision P is

the number of true positive outcome and the Recall R is the number of true positive outcome divided by the true positive outcomes should have been returned. The F-Measure is the Harmonic Mean of Precision and Recall.

$$\text{F-Measure} = 2\left(\frac{Precision*Recall}{Precision+Recall}\right)$$

3.1.6 ROC Area

Receiver Operating Characteristic curve (ROC) is a plot of the correct positive rate opposed to the false positive rate for the dissimilar feasible cut points of a diagnostic experiment.

3.1.7 Precision Recall Curve

PRC are a metric used to assess a classifier's status, distinctly when class are extremely variance.

3.1.8 Matthews Correlation Coefficient

The Mathews Correlation Coefficient (MCC) is used to evaluate of the quality of binary classifications .It returns a point between -1 to +1. A Coefficient of +1 indicates a refine prediction, 0 indicates no perfect than random forecast and -1 represents total dissent between prediction and examination.

$$MCC\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

3.1.9Kappa statistic:

Kappa Statistics is a standard benchmark, mainly used to compare the observed accuracy values with the excepted accuracy values, this statistical evaluation can be performed on both single classifiers and multiple classifiers among themselves.

$$\mathbf{k} = \frac{P_O - P_e}{1-P_e}$$

3.1.10 Mean absolute error (MAE):

The MAE is primarily used to ascertain the average error magnitude of the forecast sets, excluding its direction. The exactness of the continuous variable is measured using an equation provided in library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

3.1.11 Root mean squared error (RMSE):

RMSE is termed as a quadratic scoring rule, it is utilized to measure the errors average magnitude, the associated equation for root mean squared error is give in a couple of reference, either it is to express the formula in words or to express the unassociativity between the forecast and observed values, that are produced by averaging the sample values. Ultimately the square root of the average is provided.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{n}}$$

3.1.12Relative Absolute Error:

The relative absolute error is alike the relative squared error in the sense that it is also relatively to an easy predictor, which is just the average of the actual values. In this case, though, the error is just the overall absolute error instead of the total squared error.

$$\text{RAE=} \quad \frac{\sum_{j-1}^{n}|P_{(i)}-a_i|}{\sum_{i=1}^{n}|\bar{a}-a_i|}$$

3.2 PERFORMANCE EVALUATION:

Table -1: Detailed Accuracy by Class (Training Data)

	Hungary	Switzerland	VA	Weighted Average
TP Rate	0.993	1.000	0.869	0.963
FP Rate	0.012	0.035	0.003	0.013
Precision	0.993	0.800	0.989	0.969
Recall	0.0993	1.000	0.860	0.963
F-Measure	0.993	0.889	0.920	0.969
MCC	0.981	0.879	0.901	0.950
ROC Area	1.000	1.000	0.998	1.000
PRC Area	1.0000	0.998	0.995	0.999

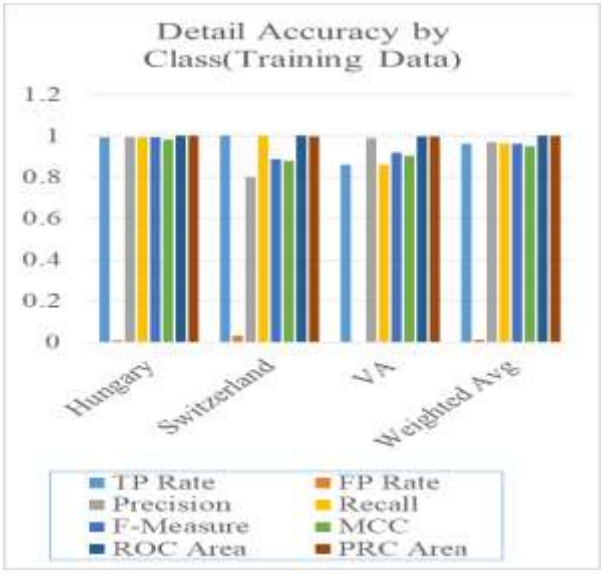

Chart -1: Detailed Accuracy by Class (Training Data)

Table -2: Detailed Accuracy by Class (Testing Data)

	Hungary	Switzerland	VA	Weighted Average
TP Rate	1.000	0.963	1.000	0.989
FP Rate	0.000	0.000	0.015	0.004
Precision	1.000	1.000	0.955	0.989
Recall	1.000	0.963	1.000	0.989
F-Measure	1.000	0.981	0.977	0.989
MCC	1.000	0.973	0.970	0.989
ROC Area	1.000	1.000	0.999	1.000
PRC Area	1.0000	1.000	0.989	0.999

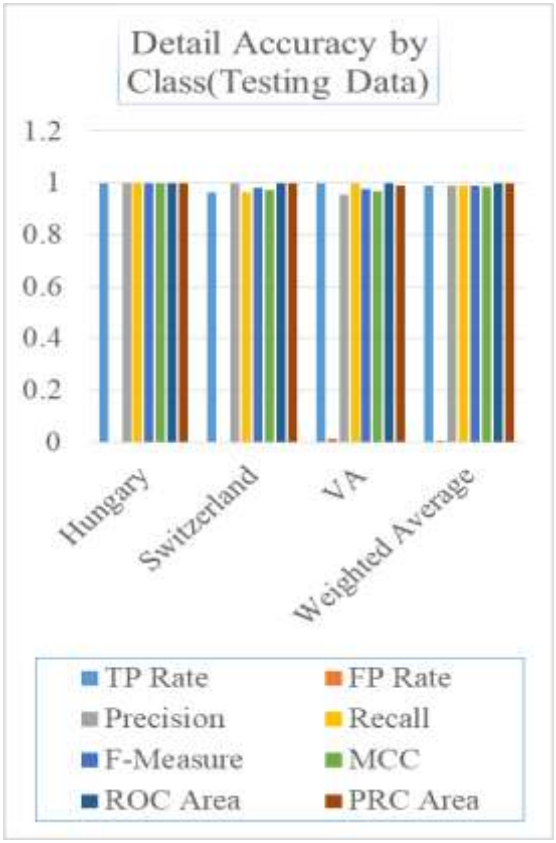

Chart -2: Detailed Accuracy by Class (Testing Data)

Table -3: Dataset Classification Based On Its Properties

	Training Data	Testing Data
Correctly Classified Instances	17	1
Incorrectly Classified Instance	0.9284	0.9823
Kappa Statistics	0.0251	0.0162
Mean absolute error	0.1325	0.0957
Root absolute error	7.274%	3.783%
Relative absolute error	31.9496%	20.6715%
Root squared relative error	3.7199%	1.1364%
Loss	3.7199%	1.1364%
Accuracy	96.2801%	98.8636%

IV.CONCLUSION:

In this paper, Naive Bayes classification of Data Mining has been discussed that can be used for predict the accuracy of Cardiovascular disease. The accuracy or prediction rate of Naive Bayes is 98.8636%.The inquiries are conducted with WEKA tool and the Naive Bayes algorithm applied on the heart dataset. An important challenge in Data Mining is to build precise and computationally efficient classifiers for Medical application.

References

[1] Ritika Chandha and Shubhankar Mayank, "Prediction of heart disease using datamining techniques," Springer, 2016.

[2] Sujata Joshi and Mydhili k.Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques," Springer, vol. 2, 2015.

[3] K.Gomathi and Dr.Shanmugapriyaa, "Heart Disease Prediction Using Data Mining Classification," International Journal for Research in Applied Science & Engineering Technology(IJRASET), vol. 4, no. II, February 2016.

[4] Rishabha Saxena, Aakriti Johri, Vikas Deep and Purushottam Sharma, "Heart Disease Presdiction System Using CHC-TSS Evolutionary,KNN,and Decision Tree Classification Algorithm," Springer, 2019.

[5] Purushottam, Pro.(Dr)Kanak Saxena and Richa Sharma, "Efficiant Heart Disease PredictionSystem," Elsevier, 2016.

[6] Ajad Patel, Sonali Gandhi, Swetha Shetty and Prof.Bhanu Tekwani, "Heart Disease Prediction Using Data Mining," International Research Journal of Engineering and Technology(IRJET), vol. 04, no. 01, January 2017.

[7] A.Kathija, S.Shajun Nisha and Dr.Mohamed Sathik, CLASSIFICATION OF BREAST CANCER DATA USING C4.5 CLASSIFIER ALGORITHM," International Journal of Recent Engineering Research and Development (IJRERD), vol. 2, no. 2.

BIOGRAPHIES

S.Spino, M.Phil Research scholar, currently pursiung at sadakathullah appa college, I had completed my PG at St.Mary's College(Autonomous), Manonmaniam Sundaranar University Tirunelveli in Computer Science and completed my B.Sc., Computer Science at St.Mary's College(Autonomous), Manonmaniam Sundaranar University, Tirunelveli in Computer Science. I had a certification of NPTEL courses. My research area is in Data Mining.

Dr.M.Mohamed Sathik, Principal Sadakathullah Appa college,Tirunelveli.He has completed Ph.D(Computer science &

engineering) Ph.D(Computer science), M.Phil. (Computer Science),M.Tech(Computer Science and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli. He has so far guided more than 40 research scholars. He has published more than 100 papers in International Journals and also two books. He is a member of curriculum development committee of various universities and autonomous colleges of Tamil Nadu. He is a syndicate member Manonmaniam Sundaranar University, Tirunelveli. His specializations are VRML, Image Processing and Sensor Networks.

Dr.S.Shajun Nisha, Assistant Professor and Head of the PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli. She has completed M.Phil. (Computer Science) M.Tech (Computer and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli and She had completed Ph.D (Computer Science) in Bharathiar university, Coimbatore. She has involved in various academic activities. She has attended so many national and international seminars, conferences and presented numerous research papers. She is a member of ISTE and IEANG and her specialization is Medical Image Processing and Neural Networks