Analysis of J48 Algorithm on Heart Attack and its Diseases

Shubham Sharma#1, Anjana Pandey*2, Mahesh Pawar#3

#Department of Information Technology, UIT- RGPV, Bhopal-402036, India

[1]shubham.ss@yahoo.com, [2]anjanapandey@rgtu.net, [3]mkpawar24@gmail.com

Abstract-The central objective of this research is to learn and reveal the reimbursements of J48 algorithm in Healthcare. The study is focused on deep study of this algorithm and will mainly reveal its benefits and accuracy in heart diseases mainly heart attack. Heart attack and its diseases becomes the cruellest thing now a days, an estimated 17.7 million **persons expired** after CVDs in 2015, demonstrating 31% of all **global** losses [1]. In this paper we try to predict heart diseases with machine learning algorithm on the basis of data provided on UCI repository. Although, the data isn't accurate and also it doesn't applicable in all the areas (due to some geographical conditions) at all but it provides an overview to achieve the Predicted results using machine learning and training algorithms. This research shows prediction accuracy at 99% and also includes comparison with several algorithms.

Keywords- J48, Heart attack, heart diseases, prediction algorithm, healthcare

I. INTRODUCTION

The European Public Health Alliance uncovered that heart ambushes, strokes and other circulatory contaminations speak to 41% of all passing's (European Public Health Alliance 2010) [2]. This examination work is expected to enhance determination exactness to enhance wellbeing results. A piece of the Decision Tree strategy engraves used like J4.8 and C4.5 Decision Trees rely upon Gain Ratio in the extraction of Decision Tree rules. There is colossal measure of clinical information produced ordinary however in which crucial data is covered up

Coronary illness

Coronary sickness is a narrowing of the tiny veins that stream blood and oxygen to the heart. This is equally termed as Coronary Conduit illness or generally a Heart attack. Coronary thrombosis ailment is typically caused by a condition called atherosclerosis, which happens when oily material and a substance called plaque creates on the dividers of courses. This makes them get restricted. As the coronary veins limit, blood stream to the heart can back off or quit, causing chest.

Agony compactness of inhalation, heart assault, and different side effects. Males in their 40's have greater danger of coronary thrombosis illness than ladies, yet as ladies gets more established, their hazard expands so this is relatively equivalent to a man's threat.

Significant hazard elements of coronary illness are
i) Diabetes
ii) Extraordinary pulse
iii) High LDL (terrible) cholesterol
iv) Low LDL (great) cholesterol

v) Not getting enough physical action
vi) Fatness
vii) Smoking

II. LITRATURE REVIEW

Coronary illness expectation

Various data mining techniques used as a piece of the finish of coronary ailment incredible precision. The identification of a coronary illness in light of a few variables or manifestations is a multi-layered .The powerful strategy is to abuse the learning and experience of a few masters in helping Diagnosis process [3].

Information mining methods as gullible bayes, neural networks, optimal tree and bolster vector machine for prediction and definition of heart infections.

The model utilizing innocent bayes and Weighted Acquainted Classifier (WAC) to foresee the likelihood of sick-person getting heart assaults been talked about in [4] N. Sundar et al.

Utilizing neural systems. also, Kumaraswamy Y. S. [5] planned a clever and viable heart assault forecast framework. Since the mining of essential examples after coronary illness, vaults on heart assault forecast, a profitable strategy has been proposed. At first, with a specific end goal to prove it reasonable for the information mining development, the information storehouse was pre-handled. When the pre-preparing gets halted, the coronary illness distribution centre was bunched with the assistance of the K-implies grouping calculation, which will accept ready the data related incident from the warehouse.

An example "Intelligent Heart Disease Prediction System (IHDPS)" created by Palaniappan S. furthermore, Awang R. [6] with the assistance of data mining structures, similar to: choice trees, innocent naïve Bayes and neuronal systems. Outcomes demonstrate that in understanding the point of the characterized mining objectives, every method has its supreme quality. IHDPS can counter composite "imagine a scenario in which" inquiries though customary choice enthusiastically helpful networks can't. It will anticipate the likelihood of sick-person gets coronary sickness, utilizing restorative information, for instance, age, sex, cardiovascular pressure and glucose. It gives surprising learning, e.g. designs, relations between restorative elements and coronary illness. IHDPS is easy to understand, Web-based, expandable, solid and adaptable.

Dataset portrayal

The informational index is involved from Data Mini Source of "University of California, Irvine (UCI)" [7]. At long last the framework is validated utilizing informational indexes from Hungarian, Cleveland and Switzerland. In those datasets, absolutely, fourteen properties, for example, age, sex, chest torment write, resting circulatory strain, serum fatty acid in mg/dl, fasting glucose, inactive electro-cardio realistic outcomes, and most extreme heart rate accomplished, practice prompted angina, ST wretchedness, and incline of the pinnacle practice ST fragment, number of real vessels, defrost and conclusion of coronary illness are possible.

Patient DATASET

The patient informational index is amassed from information gathered from restorative professionals in South Africa. Just 11 properties are considered for the expectations from the database are required for the coronary illness. The accompanying qualities with insignificant esteems are viewed as: Patient Identification Number (PID) (supplanted with sham esteems), Sex, graphical record i.e. Cardiogram, Age, Chest Ache, B.P. Level, Heart beats Rate, Fats or cholesterol, Smoke habit, Liquor utilization and Blood Sugar Level. Waikato Environment for Knowledge Analysis (WEKA) has been utilized aimed at expectation because of the ability in finding, examination and anticipating designs [8].

III. J48 Algorithm and DECISION TREE

J48 Choice tree is the usage of calculation ID3 (Iterative Dichotomise variant 3) actually prepared by the WEKA undertaking group. J48 calculation is a clear C4.5 decision tree for gathering. The situationproducts a double tree. The select tree technique is most steady in grouping issue. In this strategy, a tree is built to display the characterization procedure. Once a tree is made, it is associated for each tuple in the database and yield in gathering for that tuple [9] [10].

However developing a decision tree, J48 overlooks the lost esteems i.e. the incentive for that component can exist anticipated in light of what is perceived about the trait principles for the other record. The straightforward information is to the information into go focused on the estimations of the property for that thing that are start in the preparation display. J48 permits order by either choice trees or methodology produced from them [11] [12].

J48 is an expansion or expansion of ID3. The extra highlights of J48 are representing lost esteems, choice trees pruning, steady component esteem sorts, inference of principles, and so on. In the WEKA tool information mining process, J48 algorithm is an “open source Java” composed usage of the C4.5 calculation.

Essential Steps in the Algorithm: [13]

(I) In the event of the events have a place with the similar class the tree implies a leaf so the leaf is returned by plan with the planning class.

(ii) The potential information is computed for each quality, determined by a test on the trait. At that point the change in information is figured that would impact after a test on the property.

(iii) After that the best trait is begin based on the present choice measure and that property painstakingly decided for fanning.

Counting Gain

This procedure includes “Entropy” which tends to apportion of the figures dis-arrangements. The Entropy of $\vec{y}$ is measured by

$$\textbf{Entropy}\ (\vec{y}) = -\sum_{j=1}^{n} \frac{|yi|}{|\vec{y}|} \log\left(\frac{|yi|}{|\vec{y}|}\right)$$

$$\textbf{Entropy}\ (j\,|\,\vec{y}\,|\,) = \frac{|yj|}{|\vec{y}|} \log\left(\frac{yj}{\vec{y}}\right)$$

And Gain is

$$\textbf{Gain}\ (\vec{y}, j) = Entropy\ (\vec{y} - Entropy\ (j\,|\,\vec{y}\,|\,)$$

The goal is to take out the maximum of the Gain, divided by total entropy outstanding to splitting argument $\vec{y}$ by determine j.

Pruning

Due to the eccentric or original this is a key phase to the outcome. Some cases do exist in all data sets which are not 'well-defined' and vary from the other instances in its region.

The classification is executed on the cases of the training set and then tree is formed. The pruning is accomplished for decrement in classification errors which are being formed by focusing in the training set. Pruning is achieved for the generalisation of the tree.

IV. METHODOLOGY

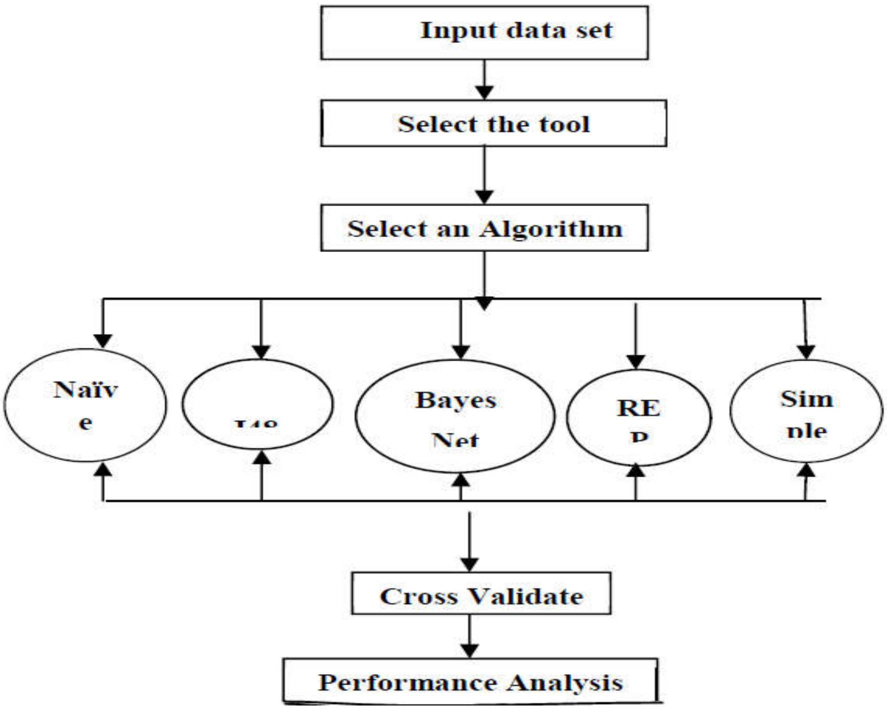


Figure 1. Working of WEKA

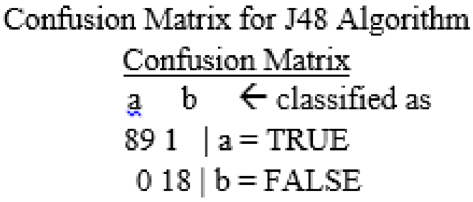
To calculate the processes of our method, patient's information set is uploaded in the tool WEKA. J48, Simple Cart, REPtree, Bayes Net and Naïve Bayes are other algorithm options available in this tool.

We select all of them. Data is then authenticated using, performance classifier calculator the outcomes and presentation of individually algois point in time equate to cross check individually. Figure 1 shows the functioning of WEKA tool.

The patient's data set is assembled from information collected together from multiple medical experts. Only 11 attributes are measured for the future predictions from the database prerequisite for the heart and its diseases. The given points with minimal standards are measured: Patient Identification Number (P.I.D) (swappedbyfalse values), Sex, Electro-Cardiogram, and Oldness, Chest ache, Blood Pressure (B.P) Level, Heart Rate, Fatty acids, Smoke habits, Alcohol ingestion and glucose Level.

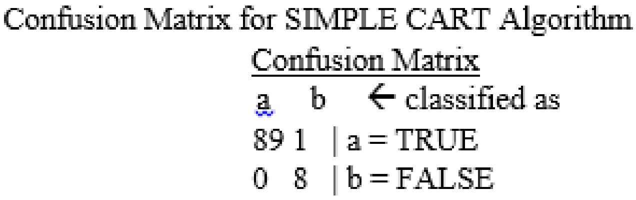
V. EXPERIMENTAL OUTCOMES

The set of rules are executed on the figures set with multiple 10-fold cross-validation in sequenced into evaluate the presentation of sorting methods for studying the sick-person's data set. The "Confusion matrix" of separately calculated algorithms are recorded and listed accordingly:



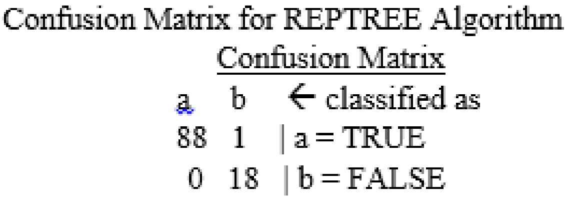
Confusion Matrix for J48 Algorithm

```
Confusion Matrix
a    b    ← classified as
89 1   | a = TRUE
 0 18 | b = FALSE
```



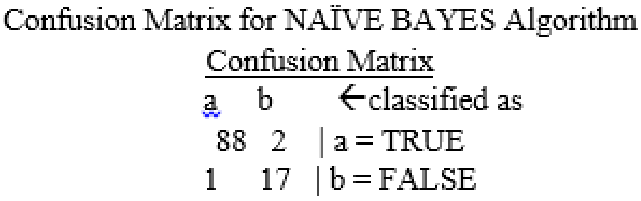
Confusion Matrix for SIMPLE CART Algorithm

```
Confusion Matrix
a    b    ← classified as
89 1   | a = TRUE
0   8   | b = FALSE
```



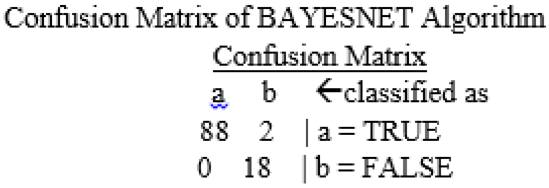
Confusion Matrix for REPTREE Algorithm

```
Confusion Matrix
a    b    ← classified as
88   1    | a = TRUE
 0   18   | b = FALSE
```



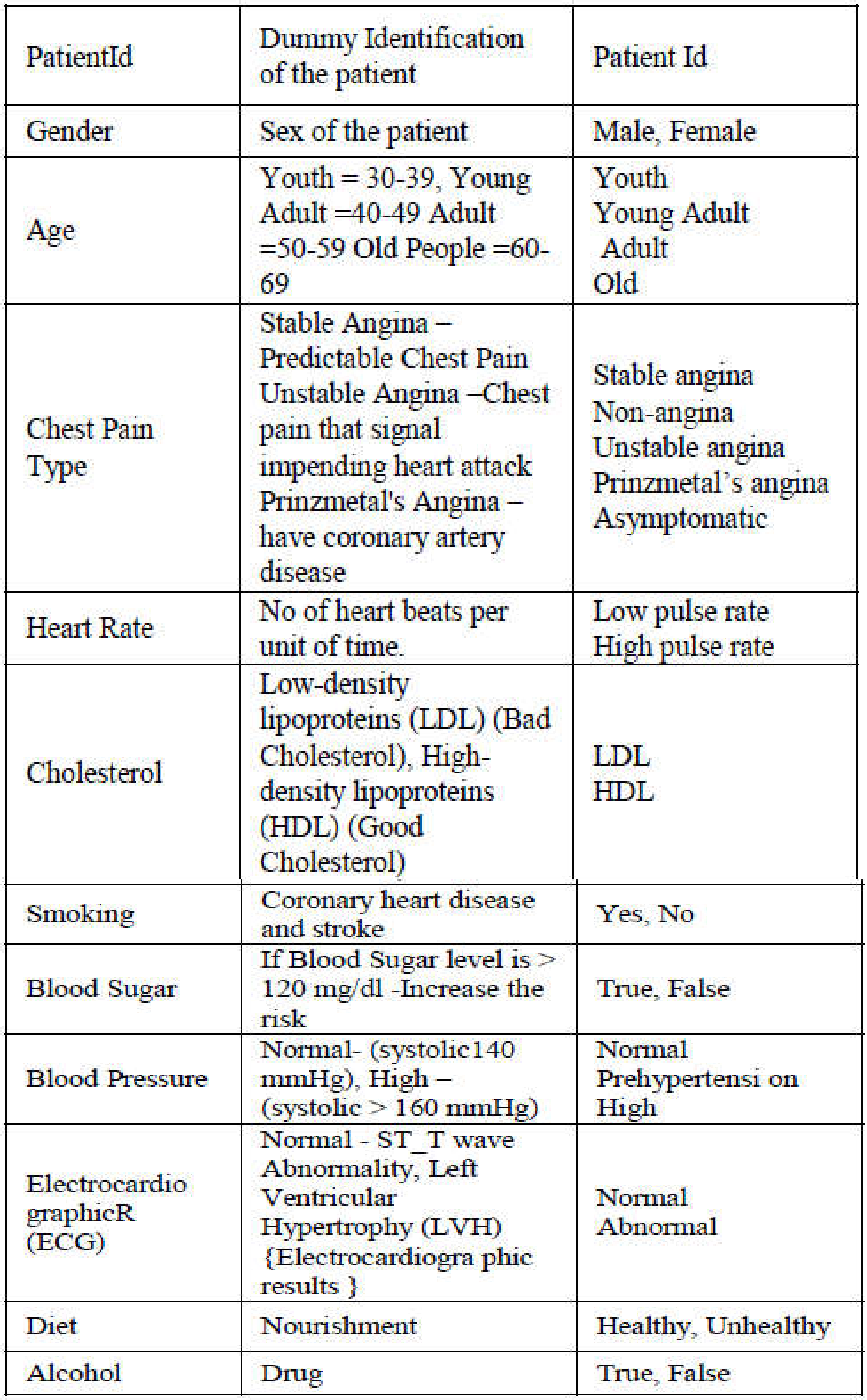
Confusion Matrix for NAÏVE BAYES Algorithm

```
Confusion Matrix
a    b      ←classified as
 88   2    | a = TRUE
1     17  | b = FALSE
```



Confusion Matrix of BAYESNET Algorithm

```
Confusion Matrix
a    b    ←classified as
88   2    | a = TRUE
0    18   | b = FALSE
```



PatientId	Dummy Identification of the patient	Patient Id
Gender	Sex of the patient	Male, Female
Age	Youth = 30-39, Young Adult =40-49 Adult =50-59 Old People =60-69	Youth Young Adult Adult Old
Chest Pain Type	Stable Angina – Predictable Chest Pain Unstable Angina –Chest pain that signal impending heart attack Prinzmetal's Angina – have coronary artery disease	Stable angina Non-angina Unstable angina Prinzmetal's angina Asymptomatic
Heart Rate	No of heart beats per unit of time.	Low pulse rate High pulse rate
Cholesterol	Low-density lipoproteins (LDL) (Bad Cholesterol), High-density lipoproteins (HDL) (Good Cholesterol)	LDL HDL
Smoking	Coronary heart disease and stroke	Yes, No
Blood Sugar	If Blood Sugar level is > 120 mg/dl -Increase the risk	True, False
Blood Pressure	Normal- (systolic140 mmHg), High – (systolic > 160 mmHg)	Normal Prehypertensi on High
Electrocardio graphicR (ECG)	Normal - ST_T wave Abnormality, Left Ventricular Hypertrophy (LVH) {Electrocardiogra phic results }	Normal Abnormal
Diet	Nourishment	Healthy, Unhealthy
Alcohol	Drug	True, False

Table 1.Dataset description

The confusion matrix [14] discovered uncovered restrictions, for example, rightness, affect-ability and correct measures and so forth. The grid speaks to models arrangements as evident and false. The lattice confirms the productivity of the model.

Plainly the disarray network arranges the suitability of model. Estimation of the perplexity lattice demonstrates that J48 forecast model of typical 89 examples with the entire riskcause positive for heart assaults.

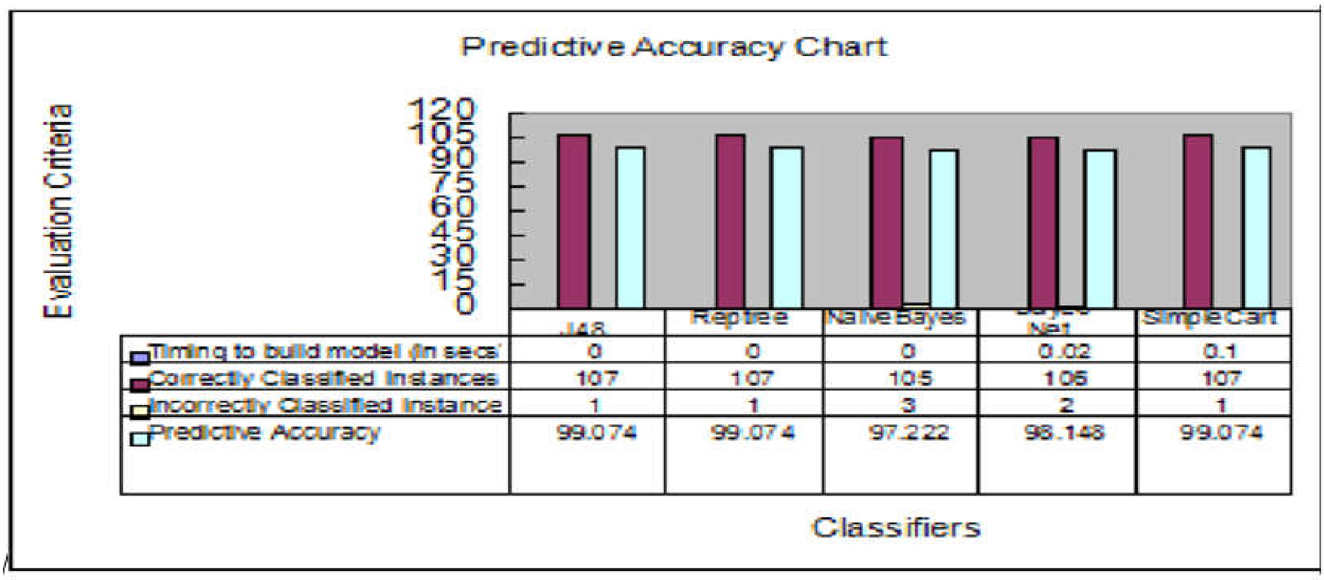
Table.2nd and Table.3rd shows the grouping rightness developed on different strategies connected, a neighbouring perception uncover that –J48 ALGORITHM demonstrates the finest order method, while Bayes Net calculation overwhelmed the Naïve Bayes calculation. Research coordinated demonstrates that J48, REPTREE and SIMPLE CART convey more prescient exactness than encourage calculations.

Evaluation criteria	**Classifiers**				
	J48	***Reptree***	***NaiveBayes***	***Bayes Net***	***Simple Cart***
Timing to build model (in secs)	0.0	0.0	0	0.02	0.1
Correctly Classified Instances	107	106	105	106	107
Incorrectly Classified Instances	1	1	3	2	1
Predictive Accuracy	99.074	99.073	97.222	98.148	99.074

Table II Predictive performance of the classifiers

Evaluation criteria	**Classifiers**				
	J48	***Reptree***	***Naïve Bayes***	***Bayes Net***	***Simple CART***
Kappa statistic	0.9674	0.9674	0.9022	0.9362	0.9674
Mean absolute error	0.018	0.018	0.071	0.053	0.018
Root mean squared error	0.099	0.099	0.165	0.140	0.099
Relative absolute error	6.547	6.547	25.280	18.952	6.547

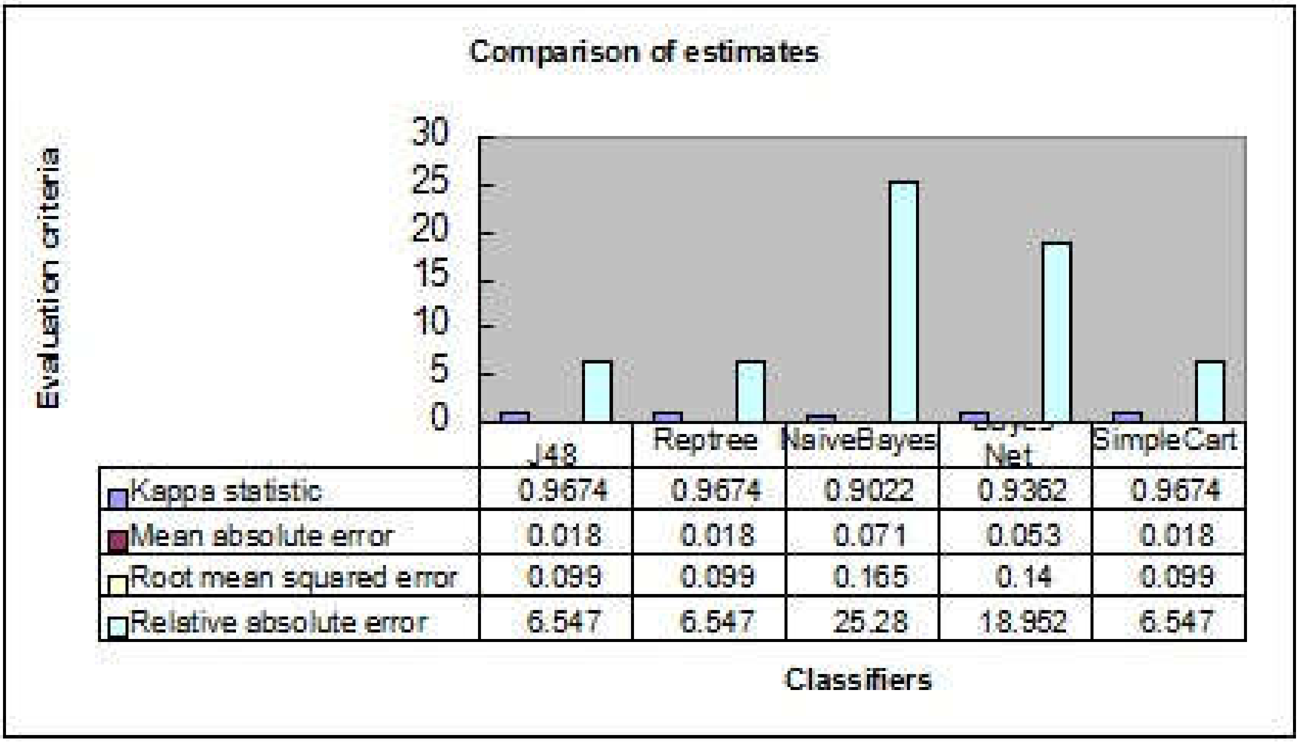
Table III Comparison of estimates



	J48	Reptree	NaiveBayes	Bayes Net	SimpleCart
□Timing to build model (in secs)	0	0	0	0.02	0.1
■Correctly Classified Instances	107	107	105	106	107
□Incorrectly Classified Instance	1	1	3	2	1
□Predictive Accuracy	99.074	99.074	97.222	98.148	99.074

Figure 2. Predictive Accuracy Chart

Figure 2displays the graph built on valuationconditions as Suitableform model in seconds, Properly Classified Occurrences, Inaccurately Classified Occurrences and Predictive correctness.

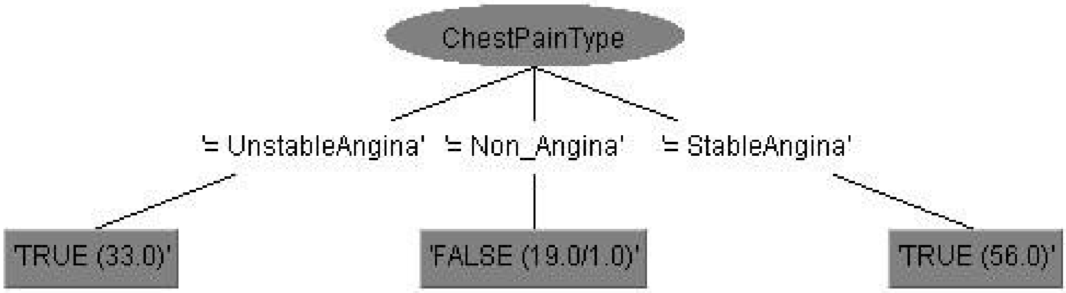


	J48	Reptree	NaiveBayes	Bayes Net	SimpleCart
□Kappa statistic	0.9674	0.9674	0.9022	0.9362	0.9674
■Mean absolute error	0.018	0.018	0.071	0.053	0.018
□Root mean squared error	0.099	0.099	0.165	0.14	0.099
□Relative absolute error	6.547	6.547	25.28	18.952	6.547

Figure 3. Comparison of estimates Chart

VI. DECISION TREE MODEL

The J48 prediction algorithm constructs personalised tree. Figure 4shows the tree illustration by means of the J48 algorithm.



VII. CONCLUSION

The goal of our work is to Research on investigation of various information mining strategies that can work proficiently in automated coronary illness forecast frameworks. Different techniques for information mining classifiers are characterized in this work with their outcomes separately that have been risen lately for productive and viable coronary illness determination. Choice tree accomplished well with precision by using properties .The applying data mining strategies to support social assurance professionals in the conclusion of coronary complaint is gigantic achievement, the utilization of information mining methods to order fitting treatment for Heart malady patients. The adjusted exactness endless supply of properties involved use for execution of the expanding coronary illness patients openness of enormous amounts of information scientist are utilizing information mining backgrounds in the exploration of coronary sickness.

REFERENCES

[1] Heart-Disease-and-Stroke-Statistics-2017-ucm_491265.pdf

[2] European Public Health Alliance. (July 2010-Febuary 2011). [Online]. Available: http://www.epha.org/a/2352.

[3] Miss. Chaitrali S. Dangare and Dr.Mrs.Sulabha S. Apte, Data Mining Approach for Prediction of Heart Disease Using Neural Networks, International Journal of Computer Engineering and Technology (IJCET), Vol- 3, Issue 3, October - December (2012), pp. 30-40

[4] N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.

[5] Patil S.B., Kumaraswamy Y.S., Intelligent and effective heart attack prediction system using data mining and artificial neural network, European Journal of Scientific Research, 31(4), 642-656, 2009

[6] Palaniappan S., Awang R., Intelligent heart disease prediction system using data mining techniques, International Journal of Computer Science and Network Security, 8(8), 108-115, 2008

[7] Newman D.J., Hettich S., Blake C.L., Merz C.J., UCI Repository of machine learning databases, University California Irvine, Department of Information and Computer Science, Irvine, CA, 1998

[8] A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.

[9] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006.

[10] R. Rao, "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 1, no. 3, pp. 14–34, 2011.

[11]http://www.jstor.org/discover/10.2307/40398417?uid=3738256&uid=2134&uid=368470121&uid=2&uid=70&uid=3&uid=368470111&uid=60&sid =21101751936641

[12] http://stackoverflow.com/questions/10317885/decision-treevs-naive-bayes-classifier.

[13] Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.

[14] Y. Xing, J. Wang, Z. Zhao, and A. Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in 2007 International Conference on Convergence Information Technology (ICCIT 2007), 2007, pp.868–872