Performance Evaluation of Classification Algorithm J48 on Heart attack and its diseases

A Survey Report

Shubham Sharma[1], Dr. Anjana Pandey[2], Dr. Mahesh Pawar[3]
[1] PG Scholar, [2] Assistant Professor, [3] Assistant Professor
[1] Department of Information Technology
[1] UIT- RGPV, Bhopal-462030, India
[1] Department of Information Technology, UIT- RGPV, Bhopal-462030, India

Abstract: With the forthcoming pattern of online exchange in all parts measure of recently made data builds each year. Information mining characterization calculations can take after three extraordinary learning approaches: semi-supervised learning, supervised learning and unsupervised learning. In this paper we apply and dissect the generally utilized grouping calculations on therapeutic informational collection that predicts coronary illness that records to be the essential driver of death around the world. Decision trees are appeared to be not as sufficient for course illness expectation as affiliation rules. Tests demonstrate decision trees have a tendency to find couple of clear standards, most principles have genuinely low unflinching quality, and most attribute parts are not exactly the same as restoratively general parts, and most guidelines insinuate little plans of patients. Conversely, association rules for the most part incorporate less complex association rules, they function admirably with client binned properties, run dependability is higher and rules by and large allude to bigger arrangements of patients.

IndexTerms: Algorithm, Classification, Diseases, Heart-Attack, J48

I. INTRODUCTION

The European Public Health Alliance uncovered that heart ambushes, fondle in addition to other circulatory contaminations speak to about 40% of all passing (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.

Apparatus which we have utilized is WEKA which is an open source's product and is extremely a great one for information mining and machine learning experimentations. Number of strategy like information mining, pre-handling, grouping, order, representation and highlight choice, extraction and so on are bolstered by WEKA. Information gave must be accessible in a solitary record which is in ARFF (Attribute Relation File Format), it ought to have traits which are connected somehow. The record can have expansions .arff/.csv. A few issues with WEKA are the failure to perform multi-social procedures and the absence of a union device for interrelated tables.

There is colossal measure of clinical information produced ordinary however in which crucial data is covered up

Coronary illness

Coronary disorder is a narrow of the minor veins that stream blood and oxygen to the heart. This is similarly named as Coronary Conduit sickness or by and large a Heart assault. Coronary thrombosis infirmity is normally cause by a state called atherosclerosis, which happens while sleek substance and a matter call plaque makes on the dividers of courses. This influences them to get confined. As the coronary veins constrain, circulatory system 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

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.

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

II. J48 DECISION TREE

Portrayal is the path toward construct a model of programme from a game plan of files that include class marks. Decision Tree Algorithm is to find the approach of the properties vector carries on for different cases. Furthermore on the bases of the readiness events the classes for the as of late made cases are being found. This estimation makes the precepts for the desire for the goal variable. With the help of tree gathering figuring the essential course of the data is viably sensible.

J48 is a development of ID3. The additional features of J48 are speaking to missing characteristics, decision trees pruning, predictable quality regard ranges, acceptance of measures, et cetera. In various computations the gathering is implemented recursively upto each end side is unmodified, to facilitate the request of the information should be as perfect as could be permitted. This figuring it delivers the rules from which particular character of that data is made. The objective is powerfully theory of a decision hierarchy in anticipation of the point that it grabs adjust of versatility with accuracy.

Fundamental Phases in the Algorithm:
(i) In case the events have a set through a comparable group of hierarchy addresses a sheet so the side is return with designation with a comparable class.
(ii) The possible information is figured for every attribute, known by a analysis on the quality. By then the get in information is assumed that would happen due to an analysis on the property.
(iii) After that the finest trait is establish on the present choice basis and that trait chose for spreading.

Highlights of the Algorithm:

(i) Both the discrete and consistent characteristics are dealt with by this calculation. A limit esteem is chosen by C4.5 for dealing with ceaseless qualities. This esteem separates the information list into the individuals who have their trait esteem beneath the limit and those having more than or equivalent to it.
(ii) This figuring also handles the missing characteristics in the planning data.
(iii) After the tree is completely built, this calculation plays out the pruning of the tree. C4.5 after its development drives back through the tree and difficulties to evacuate branches that are not helping in achieving the leaf hubs.

III. DATASET AND TOOLS

UCI SWITZERLAND DATABASE

The coronary ailment database which is from the Hungarian database, Irvine, Uci account is used. It contains four enlightening accumulations from the Cleveland focus foundation, Hungarian establishment of cardiology, v.a. therapeutic concentration and school specialist's office of Switzerland. It gives 920 records out and out. At first, the database had 76 unrefined properties. In any case, most of the appropriated examinations use only 13 of these. The Hungarian_csv database with 294 occasions and 14 properties cp, trestbps, chol, fbs, restecg, age, sex, talach, exang, oldpeak, slant, ca, thal and num were used here for the examination. The point by point info around the 14 characteristics or attributes has remained given underneath:

1. (age) age in years
2. (sex) sex (1 = male; 0 = female)
3. (chest_pain) chest_pain: chest torment compose
    Esteem 1: run of the mill angina
    Esteem 2: atypical angina
    Esteem 3: non-anginal agony
    Esteem 4: asymptomatic
4. (trestbps) resting pulse (in mm Hg on admission to the doctor's facility)
5. (chol) serum cholestoral in mg/dl
6. (fbs) (fasting glucose > 120 mg/dl) (1 = genuine; 0 = false)
7. (restecg) restecg: resting electrocardiographic outcomes
    Esteem 0: ordinary
    Esteem 1: having ST-T wave variation from the norm (T wave reversals and additionally ST rise or wretchedness of > 0.05 mV)
    Esteem 2: demonstrating plausible or positive left ventricular hypertrophy by Estes' criteria
8. (thalach) most extreme heart rate accomplished
9. (exang) practice incited angina (1 = yes; 0 = no)
10. (oldpeak) ST misery prompted by practice with respect to rest
11. (slant) the slant of the pinnacle practice ST fragment
    Esteem 1: upsloping
    Esteem 2: level
    Esteem 3: downsloping
12. (ca) number of real vessels (0-3) shaded by flourosopy
13. (thal) 3 = typical; 6 = settled deformity; 7 = reversible imperfection
14.(num) (the anticipated characteristic, determination of coronary illness (angiographic infection status)
    Esteem 0: < half distance across narrowing.

Esteem 1: > half distance across narrowing.

WEKA

WEKA is a tool/software created by the Academia of Waikato in New Zealand which is used for data-mining that concludes data filtering and figurings. Weka is a finest in class helps making machine training and learning systems and its application to genuine data-mining topics. It's an amassing of machine learning computations for data-mining structures. Weka is a mining gadget and data training. It contains various machine slanting estimations. It gives the workplace to portray our information over several computations. The estimations are associated direct to a dataset. Weka executes figuring for data pre-getting ready, course of action, backslide, gathering, association rules; it moreover joins an observation gadgets. The fresh machine learning designs can in like manner be made through this tool. Weka is open and free source programming distributed beneath the universal public allow.

IV. LITERATURE SURVEY

1. Strategy utilizing optimality criterion feature selection (OCFS) for proficient sickness finding and expectation. We expand the strategy for harsh feature selection in view of data entropy (RFS-IE). Initial phase during the time spent the OCFS sets is to change over the extricated features (i.e., information) into a choice table. The optimal attribute is a two-dimensional attribute that contains rows and segments. The line in the choice table relates to the distinct number of items (i.e., perception), while the section compares to the trait esteem and class name of the specific question.

2. In this examination work, Random Forest and J48 Classifiers are assessed for adroitness valuation of coronary illness forecast. Open source machine learning apparatus is utilized to explore the execution of Random Forest and J48 Classifiers. The execution is tried out utilizing the whole Training set and in addition utilizing diverse Cross Validation techniques. The class is anticipated by considering every one of the 13 qualities of the dataset. Random Forest Classifier gives 99.63% exactness for the preparation informational index. Different cross approval strategies are utilized to check its genuine execution. Random Forest gives a normal of 79.33% precision for coronary illness expectation. Classifier gives 91.4815% exactness for the preparation informational collection. Different cross approval techniques are utilized to approve its genuine execution. J48 gives a normal of 77.26% exactness for coronary illness expectation.

3. This paper portrays Association run mining and course of action are two essential functionalities of data mining. Affiliation control mining is used to find affiliations or associations among the thing sets. It is an unsupervised learning where no class characteristic is locked in with finding the affiliation run the show. Then again, arrangement is a regulated realizing where class characteristic is engaged with the development of the classifier and is utilized to group or anticipate the information obscure example. Rule positioning assumes an essential part in grouping and most of the acquainted classifiers select rules chiefly as far as their certainty levels. Without a doubt, even resulting to pruning intermittent things, the APRIORI affiliation run age strategy, makes an enormous no. of affiliation rules .If each one of the guidelines are used as a piece of the classifier then the exactness of the classifier would be high yet the working of portrayal will be direct. With a specific end goal to enhance the precision of affiliated grouping we propose an educational property entered lead age and speculation testing Z-measurements for coronary illness forecast. The class association rules are represented as chromosomes and Michigan approach is used to encode the rules.

4. This paper depicts distinguish utilization of Big Data examination in heart assault expectation and counteractive action, the utilization of advances appropriate to huge information, protection worries for the patient, and difficulties and future patterns and additionally recommendations for additionally utilization of these advances.

The national and worldwide databases were inspected to recognize contemplates directed about enormous information investigation in medicinal services, heart assault expectation and avoidance, innovations utilized as a part of huge information, and security concerns. A sum of 31 contemplates that fit these criteria were surveyed. This framework is fundamentally worried about two datasets—the first enormous dataset and the refreshed dataset. There are additionally two nodes: the Name Node which keeps a record of all documents in the document framework, and tracks the site of each document in a bunch; there is likewise the Data Node which distribution centers information in the Hadoop File System. Any proficient document framework is included in excess of one Data Node, with information replication. Hbase is utilized when a specialist needs irregular; continuous read or composes access to Big Data

5. The calculations are connected on the informational index utilizing stratified 10-fold cross-validation keeping in mind the end goal to evaluate the execution of characterization procedures for breaking down the patient informational index. The perplexity grid unmistakably orders the exactness of method. Assessment of perplexity network proves that REPTREE, J48 and SIMPLE CART demonstrate a forecast model containing of 89() characters with a peril dynamic optimistic as heart assaults. The prescient exactness controlled by REPTREE, J48 and BayesNet calculations recommend that limitations utilized are steady pointer to foresee the heart illnesses.

6. 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.

7. The model utilizing innocent bayes and Weighted Acquainted Classifier (WAC) to foresee the likelihood of sick-person getting heart assaults. Weighted Associative Classifier (WAC), diverse weights are allocated to various credits as indicated by their anticipating ability. Weighted Associative Classifier (WAC) is another thought that usages Weighted Association Rule for gathering. Weighted ARM uses Weighted Support and Confidence Framework to isolate Association regulate from data document. The WAC has been proposed as another Technique to get the immense run as opposed to overpowered with irrelevant association. Fresh DM (cross industry standard process for data mining) way to deal with create the mining or filtering methods. It covers of 6 noteworthy stages: commercial empathetic, data kind, data planning, demonstrating, assessment, and then arrangement. Business considerate stage centres around accepting the destinations and

necessities after a corporate opinion of interpretation, changing over this learning to a data-mining issue explanation, and planning a preparatory arrangement to accomplish the targets.

8. Utilizing neural systems. 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 center was bunched with the assistance of the K-implies grouping calculation, which will accept ready the data related incident from the warehouse.

9. An example of "Intelligent Heart Disease Prediction System (IHDPS)" created by authors of it. With the assistance of data mining structures, similar to: choice trees, innocent naïve Bayes and neuronal systems. Outcomes demonstrates 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 a 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

10. The shrouded designs that are found can be utilized to comprehend the issue emerge in the instructive field. This paper overviews the three components expected to make expectation on Students' Academic Performances which are parameters, strategies and apparatuses. This paper in like manner proposes a structure for foreseeing the execution of first year solitary wolf understudies in programming building course. Guiltless Bayes Classifier is used to expel outlines using the Data Mining Weka mechanical assembly. The structure can be used as an explanation behind the system execution and desire for Students' Academic Performance in Higher Learning Institutions. Waikato Environment for Knowledge Analysis (WEKA) has been utilized aimed at expectation because of the ability in finding, examination and anticipating designs.

11. In this paper, J48 Choice tree is the usage of calculation ID3 (Iterative Dichotomiser variant 3) actually prepared from WEKA undertaking group. J48 calculation is a clear C4.5 which is a decision making tree for gathering. The situation products 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 related for every entity in the datastore and yield in social affair for that tuple. It's given a productive way to deal with the extraction of noteworthy examples from the coronary illness information stockrooms for the proficient expectation of heart assault in view of the ascertained huge weightage. The continuous examples having esteem more prominent than a predefined limit were decided for the important forecast of heart assault. Three mining objectives are characterized in view of information investigation. Every one of these models could answer complex questions in anticipating heart assault. The rate of heart assault is expanding each year in coal mining areas, particularly in India, is a consistent terrorizing to the populace and a repeating issue for the wellbeing specialists.

V. SURVEY REPORT

Author	Topic	About	Parameter/Important
1.S. Prakash, K. Sangeetha, N. Ramkumar (Springer-2017)	An optimal criterion feature selection method for prediction and effective analysis of heart disease	optimality criterion feature selection (OCFS) and extend rough feature selection based on information entropy (RFS-IE)	**Qualitative Analysis-** 1.computational Time 2. Consistency 3. error rate 4. optimality criterion
2. Lakshmi Devasena C(IJCAR-2016)	Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction	Random Forest and J48 Classifiers are figured for capability estimation of coronary illness expectation	1. Cross Validation methods 2. confusion matrix for different test mode 3. Random Forest gives an average of 79.33% accuracy 4 J48 gives an average of 77.26% accuracy 5. Effectiveness comparison of both the classifiers
3.M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu (ICECIT, 2012)	Heart Disease Prediction System using Associative Classification and Genetic Algorithm	genetic algorithm and hypothesis testing Z Statistic proposed linking it with classification algorithms like j48 , naive bayes, neural networks and genetic network programming GNP	1.Heart disease data accuracy using j48 is 4% higher from naive Bayes and the accuracy has been enhanced with NN.J48 2.The systolic pressure is higher in Males (44% cases) 3. Average accuracy of J48 is 76.56, Naïve bayes is 74.65. 4. Accuracy with GNP using che-square method is 83.70.

4. Cheryl Ann Alexander and Lidong Wang(Journal of nursing and care,2017)	Big Data Analytics in Heart Attack Prediction	Look at current criteria, methodologies, and traditions for huge information to enhance a heart assault expectation framework to help suppliers in propelling higher standards of care.	1. General process of data analytics working. 2. Uses of IoT in disease prediction using sensors etc. 3. Tele cardiology in Heart attack prediction. 4. Process to develop wireless heart attack prediction system.
5. Ms. M.C.S.Geetha, Dr.I.Elizabeth Shanthi, Ms.N. Sanfia Sehnaz(IEEE,2017)	Analysing the Suitability Of Relevant Classification Techniques On Medical Data Set For Better Prediction	WEKA is used for the performance and comparison on accuracies of different prediction algorithms.	1. Classification Algorithms like J48, SIMPLE CART, REPTREE proves the un-surpassed practises. 2. Naïve Bayes algorithm outclassed by the Bayes Net algorithm 3. J48, SIMPLE CART and REPTREE deliver extra predictive exactness than further algorithms.
6. Miss. Chaitrali S. Dangare and Dr.Mrs.Sulabha S. Apte (IJCET),2012	Data Mining Approach for Prediction of Heart Disease Using Neural Networks	From the ANN, a multilayer perceptron neural network alongside back spread calculation is utilized to build up the framework	1. Multilayer Perceptron Neural Network (MLPNN) 2. multi-layered work done for predictions 3. Backpropagation network 4. Neural network accuracy with 99.25 with 13 attributes.
7.N. A. Sundar, P. P. Latha, and M. R. Chandra(IJESAT,2012)	PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE	2 data withdrawal classification sculpting approaches are used classification Matrix methods and DMX query language and purposes are used to form and access the models	1. DMX query language and functions are used to build and access the models. 2. Arrangement Matrix strategies are utilized to assess the adequacy of the models. 3. All parameters were set to the default setting with the exception of parameters "Least Support = 1" for Decision Tree and "Least Dependency Probability = 0.005" for Naïve Bayes 4. half of test dataset is anticipated effectively. The graph demonstrates that WAC gives (84%) trailed by Naïve bayes (78%)
8. Patil S.B., Kumaraswamy Y.S (EJSR 2009)	Intelligent and effective heart attack prediction system using data mining and artificial neural network	The pre-handled coronary illness information distribution center was grouped with the K-means clustering calculation so as to get information most relevant to heart disease/attack. The regular things were extracted effectively with the guide of MAFIA calculation.	1. K-means clustering used for data refining. 2. Frequent Item set Mining (FIM) using MAFIA (MAximal Frequent Itemset Algorithm) which syndicates various time worn and new algorithmic concepts to custom a applied algorithm 3. The proposed calculation is utilized for the abstraction of association directions from the assembled dataset other than performance productively when the database embraces of lengthy thing sets mainly. 4. Significance Weightage Calculation

9. Palaniappan S., Awang R. (IJCSNS,2008)	Intelligent heart disease prediction system using data mining techniques	Intelligent Heart Disease Prediction System (IHDPS) using three data-mining modelling methods, namely, Decision Trees, Neural Network and Naïve Bayes.	1. CRISP-DM methodology. 2. Data Mining Extension (DMX), a SQL-style query language for data mining. 3. Correct predictions (86.12%) surveyed through Neural Network (85.68%) and Decision Trees (80.4%).
10. AZIZ, N. ISMAIL, and F. AHMAD	MINING STUDENTS'ACADEMIC PERFORMANCE	Framework for predicting SAP based on the selected parameters and NBC is presented	1. Different data mining tools are explained. 2. prediction model presentations comparable a threatening system to sense possible fragile students
11. R. Rao (IJDKP),2011	SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES	Obtainable the prediction system in the basis of Neural network prediction using Time series analysis.	1.Time Series Data Mining (TSDM) 2.Autoregressive Integrated Moving Average (ARIMA) methodology involves finding solutions to the difference equation 3.Neural network (MLP) with accuracy rate 89.2%

REFERENCES

[1] S. Prakash, K. Sangeetha, N. Ramkumar, *An optimal criterion feature selection method for prediction and effective analysis of heart disease*, Springer-2017 https://doi.org/10.1007/s10586-017-1530-z

[2] Lakshmi Devasena C, *Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction*, ISSN 2305-9184, Volume 5, Number 1 (February 2016), pp.46-55

[3] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, *Heart Disease Prediction System using Associative Classification and Genetic Algorithm*, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

[4] Cheryl Ann Alexander and Lidong Wang, *Big Data Analytics in Heart Attack Prediction*, Journal of nursing and care, ISSN:2167-1168 Volume 6 • Issue 2, 1000393, 2017

[5] Ms. M.C.S.Geetha, Dr.I.Elizabeth Shanthi, Ms.N. Sanfia Sehnaz, *Analysing the Suitability of Relevant Classification Techniques On Medical Data Set For Better Prediction*, 978-1-5090-3243-3/17/$31.00 ©2017 IEEE

[6] 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

[7] 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

[8] 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

[9] 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

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

[11] 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.