Medicinal Decision Support System for Cardiovascular Disease using Data Mining Techniques

Poonam Rahul Hankare
Assistant Professor
SCOESCOE
Kharghar

Hemalata A.Gosavi
Assistant Professor
SCOESCOE
Kharghar

ABSTRACT
Restorative science industry has tremendous measure of information, however shockingly the vast majority of this information is not mined to discover out shrouded data in information. Propelled information mining systems can be utilized to find shrouded design in information. Models created from these systems will be valuable for medicinal professionals to take successful choice. In this examination paper, one of the information mining arrangement system Decision Tree calculation C4.5, ID3 and CART are dissected on cardiovascular illness dataset. Exhibitions of these calculations are thought about through affectability, specificity, exactness, blunder rate, True Positive Rate and False Positive Rate. In our studies 10-fold cross acceptance system was utilized to gauge the impartial evaluation of these expectation models. According to our outcomes, mistake rates for Decision Tree calculation C4.5, ID3 and CART are 02.756, 0.2755 and 0.2248 individually. Exactness of Decision Tree calculation C4.5, ID3 and CART are 80.06%, 81.08% and 84.12% individually. Our examination demonstrates that out of these three order method Decision Tree calculation CART predicts cardiovascular illness with minimum mistake rate and most astounding precision.

Keywords
Active learning, decision support system, data mining, medical engineering, C4.5, ID3 and CART.

1. INTRODUCTION
The heart is the organ that pumps blood, with its nurturing oxygen and supplements, to all
tissues of the body. In the event that the pumping activity of the heart gets to be wasteful, crucial organs like the cerebrum and kidneys endure and if the heart quits working through and through, death happens inside of minutes. Life itself is totally subject to the proficient operation of the heart. Cardiovascular infection is not infectious; you can't get it like you can this season's flu virus or a frosty. Rather, there are sure things that expand a man's possibilities of getting cardiovascular sickness. Cardiovascular malady (CVD) alludes to any condition that influences the heart. Numerous CVD patients have manifestations, for example, mid-section torment (angina) and exhaustion, which happen when the heart isn't accepting sufficient oxygen. According to a review almost 50 percent of patients,
then again, have no side effects until a heart assault happens. Various elements have been appeared to build the danger of creating CVD.
Some of these are:

- Family history of cardiovascular malady
- High levels of LDL (terrible) cholesterol
- Low level of HDL (great) cholesterol
- Hypertension
- High fat eating routine
- Lack of customary activity
- Obesity

With such a variety of components to examine for an analysis of cardiovascular sickness, doctors for the most part make an evaluating so as to find an understanding's present test outcomes. Past conclusions made on different patients with the same results are likewise analyzed by doctors. These unpredictable techniques are difficult. In this way, a doctor must be experienced and exceedingly talented to analyze cardiovascular ailment in a patient. Information mining has been intensely utilized as a part of the therapeutic field, to incorporate patient determination records to recognize best practices. The troubles postured by expectation issues have brought about an assortment of critical thinking systems.
It is troublesome, on the other hand, to think about the precision of the methods and decide the best one on the grounds that their execution is information subordinate. A couple studies have contrasted information mining and measurable methodologies with take care of expectation issues. The examination studies have for the most part considered a particular information set or the dissemination of the subordinate variable.

2. BACKGROUND
Up to now, a few studies have been accounted for that has concentrated on cardiovascular infection finding. These studies have connected distinctive ways to deal with the given issue and accomplished high characterization exactness's of 77% alternately higher.

Here are a few cases:

- Robert Detrano's test results demonstrated right grouping precision of around 77% with logistic relapse inferred discriminant capacity [3].
- Zheng Yao connected another model called R-C4.5 which depends on C4.5 and enhanced the effectiveness of attribution choice and parceling models. An analysis demonstrated that the standards made by R-C4.5s can give human services specialist's clear and valuable clarifications [4].
- Resul Das presented a technique that uses SAS base programming 9.13 for diagnosing coronary illness. A neural systems group strategy is at the focal point of this framework [5].
- Colombet et al. assessed execution and execution of CART and manufactured neural systems similarly with a LR model, keeping in mind the end goal to

foresee the danger of cardiovascular malady in a genuine database [6].

The troublesome of perceiving obliged affiliation rules for heart sickness forecast was examined via Carlos Ordonez. The information mining systems have been locked in by different works in progress to examine different illnesses, for example: Hepatitis, Cancer, Diabetes, Heart sicknesses. Continuous Item set Mining (FIM) is measured to be one of the essential information mining troubles that hopes to recognize accumulations of things or values or structures that co-happen routinely in a dataset. The term Heart sickness covers the different ailments that influence the heart. Coronary illness kills one in at regular intervals in the United States of America. This procedure is utilized while recommending the patient and this framework predicts which cure as solutions and restorative test suits best.

3. PROBLEM STATEMENT

Numerous healing centre data frameworks are intended to bolster patient charging, stock administration and era of straightforward measurements. A few healing centres use choice emotionally supportive networks, yet they are to a great extent constrained. They can answer straightforward inquiries like "What is the normal time of patients who have coronary illness?", "What number of surgeries had brought about healing centre stays longer than 10 days?" .However, they can't answer complex questions like "Distinguish the vital Preoperative indicators that expand the length of doctor's facility stay "and "Given patient records, foresee the likelihood of patients getting a coronary illness."

Clinical choices are regularly made in light of specialist's instinct and experience instead of on the information - rich information covered up in the database. This practice prompts undesirable inclinations, blunders and exorbitant therapeutic expenses which influence the nature of administration gave to patients. Wu, et al suggested that joining of clinical choice backing with PC based patient records could lessen therapeutic blunders, upgrade understanding security, diminish undesirable practice variety, and enhance persistent result. This proposal is promising as information displaying and examination apparatuses, e.g., information mining, can possibly create a learning rich environment which can help to fundamentally enhance the nature of cli.

4. DECISIONTREE ALGORITHM

Under this segment we will talk about after information mining Decision Tree calculations to anticipate cardiovascular sickness:

4.1 ID3 (Itemized Dichotomize 3)

Organized Dichotomize 3 calculation or otherwise called ID3 calculation was initially presented by J.R Quinlan in

the late 1970˝s. It is an avaricious calculation that chooses the following traits in light of the data addition connected with the properties [7]. All through the calculation, the choice tree is developed with each non-terminal hub speaking to the chose quality on which the information was part, and terminal hubs speaking to the class mark of the last subset of this branch.

Steps included in ID3 calculation:

- Calculate the entropy of each trait utilizing the information set,
- Split the set into subsets utilizing the property for which entropy is least (or, proportionately, data increase is most extreme)
- Make a choice tree hub containing that characteristic,
- Recurse on subsets utilizing remaining characteristics.

4.2 C4.5

C4.5 is a system utilized for creating scientific classification tenets utilizing choice trees from an arrangement of given information. C4.5 calculation is an expansion of the essential ID3 calculation and it was composed by Quinlan. C4.5 is one of generally utilized learning calculations. C4.5 calculation fabricates choice trees from an arrangement of preparing information like the ID3algorithm, utilizing the idea of data entropy. C4.5 is otherwise called a factual classifier.

Test calculation:

- Check for base cases.
- For every component x, find the standardized data pick up from part on x.
- Let x_best be the component with the most astounding standardized data pick up.
- Create a choice hub that breaks on a best.
- Repeats on the sub records got by part on x_best, and include those hubs as offspring of hub.

At every hub of the tree, C4.5 picks one characteristic of the information that parts its arrangement of tests into subsets advanced in one class or the other.

4.3 CART

The term Classification And Regression Tree (CART) examination is an umbrella term used to allude to both of the above techniques, initially presented by Breiman et al. Trees utilized for relapse and trees utilized for arrangement have a few likenesses - additionally a few contrasts, for example, the system used to figure out where to part. It utilizes Gini polluting influences and data increase to figure choice tree.

Truck advancements include:

- settling the "how enormous to develop the tree"-issue,
- utilizing entirely two-way (paired) part,
- joining programmed testing and tree formation.

DATASET Key attribute

- Patient_id – Patient’s identification number. Attribute value to be taken into the project for heart disease is as follows:

Heart disease dataset:

- Sex (value 1: Male; value 0 : Female)
- Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value 3:
- non-angina pain; value 4: asymptomatic)
- Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
- Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value2:showing probable or definite left ventricular hypertrophy)

- Exang – exercise induced angina (value 1: yes; value 0: no)
- Slope – the slope of the peak exercise ST segment (value1: unsloping; value 2: flat; value 3: down sloping)
- CA – number of major vessels colored by fluoroscopy (value 0 – 3)
- Thal (value 3: normal; value 6: fixed defect; value7:reversible defect)
- Trest Blood Pressure (mm Hg on admission to the hospital)
- Serum Cholesterol (mg/dl)
- Thalach – maximum heart rate achieved
- Oldpeak – ST depression induced by exercise relative to rest
- Age in Year
- Num Class (0 = healthy, 1 = have heart disease)

5. RESULTS

These data mining classification model were developed using data mining classification tool Wekaversion 3.6. Initially dataset had 14 attributes and 303 records. Algorithm for attribute selection was applied on dataset to preprocess the dataset. After attribute selection missing values records were identified and were deleted from dataset. After deleting records with missing values we were left with 296 records. On these 296 records data mining Decision Tree algorithms C4.5, ID3 and CART were applied. A distinguished confusion matrix was obtained tocalculate sensitivity, specificity and accuracy. The upper left cell denotes the number of samples classifies as true while they were true (i.e., TP), and the lower right cell denotes the number of samples classified as false while they were actually false (i.e., TN). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, conclusion the upper right cell denoting the number of samples classified as false while they actually were true (i.e., FN), and the lower left cell denoting the number of samples classified as true while they actually were false (i.e., FP).

Table 1 Confusion Matrix

	Classified as Healthy	**Classified as Unhealthy**
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

Below formulae were used to calculate sensitivity, specificity and accuracy:

- Sensitivity = TP / (TP + FN);
- Specificity = TN / (TN + FP);
- Accuracy = (TP + TN) / (TP + FP + TN + FN)

Table 2 Sensitivity, Specificity and Accuracy for Different Classification Algorithms

Decision Tree algorithms	**Sensitivity**	**Specificity**	**Accuracy**
C4.5	83.75 %	75.73 %	80.06%
ID3	86.25 %	75.82 %	81.08%
CART	90.0%	77.20 %	84.12%

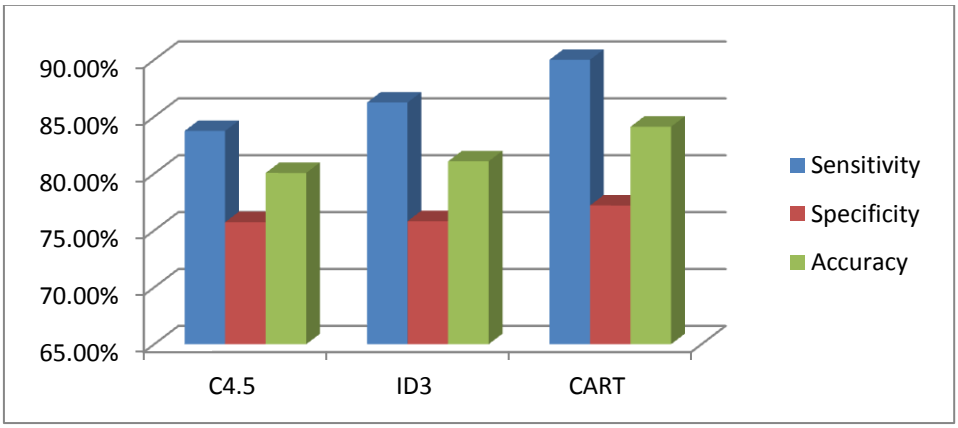

Fig1. Sensitivity, Specificity, Accuracy and Error Rate for Different Classification Techniques

As per results, error rates for Decision Tree algorithm C4.5, ID3 and CART are 02.756, 0.2755 and 0.2248 respectively . Accuracy of Decision Tree algorithm C4.5, ID3 and CART are 80.06%, 81.08% and 84.12% respectively.

- True Positive Rate = TP / (TP + FN);
- False Positive Rate = FP / (FP + TN)

Table 3.Shows True Positive Rate and False Positive Rate for Decision Tree algorithm C4.5, ID3 and CART. This will represent 100% True Positive Rate and no False Positive Rate which will be ideal case.

Table 3 True Positive Rate and False Positive Rate

Decision Tree Algorithm	True Positive Rate	False Positive Rate
C4.5	0.8375	0.2426
ID3	0.8625	0.2410
CART	0.9000	0.2279

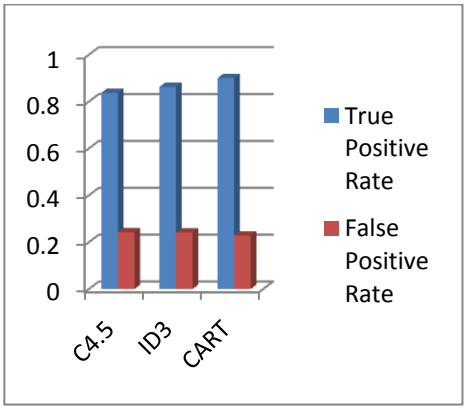

Fig2. Shows Bar Chart of TPR and FPR

6. CONCLUSION

There are diverse information mining procedures that can be utilized for the distinguishing proof and counteractive action of cardiovascular malady among patients. In this paper four grouping procedures ininformation mining to foresee cardiovascular sickness in patients are looked at: Decision Tree calculations C4.5, ID3 and CART. These methods are thought about on premise of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Our studies demonstrated that Classification and Regression choice tree calculation ended up being best classifier for cardiovascular infection expectation. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis.As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

7. REFERENCES

[1] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

[2] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth Int. Group, 1984.

[3] Detrano R, Steinbrunn W, Pfisterer M "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[4] Yao Z, Lei L, Yin J "R-C4.5 Decision tree model and its applications to health care dataset". Proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

[5] Das R, Abdulkadir S. (2008) "Effective diagnosis of heart disease through neural networks ensembles" , Elsevier,2008.

[6] Colombet I, Ruelland A, Chatellier G, Gueyffier F 2000) "Models to predict cardiovascular risk: comparison ofCART, multilayer perceptron and logistic regression", Proceedings of AMIA Symp 2000, pp. 156-160.

[7] Quinlan J R "Induction of Decision Trees," Machine Learning. Vol. 1. 1986. 81-106.

[8] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

[9] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

[10] Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/enus/virtuallabs/aa740409.aspx,2007.

[11] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[12] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005

[13] Tang, Z. H., MacLennan, J.: "Data Mining with SQL Server 2005", Indianapolis: Wiley, 2005.

[14] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[15] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.

[16] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare InformationManagement.16(4), 50-55, 2002.

[17] K.SrinivasB.Kavihta Rani Dr.A.Govrdhan Associate Professor, Dept. of CSE Principal andProfessor of CSE 'Applications of Data Mining Techniques in Healthcare and Prediction of HeartAttacks '-(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

[18] Giudici, P., Applied Data Mining Statistical Methods for Business and Industry. John Wiley & Sons Ltd, Chichester , England (2003).

[19] WHO., Fact Sheet: The Top Ten Causes of Death. World Health Organization. Geneva (2006).

[20] Han, J. and Kamber, M., Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco (2006).