International Journal of Engineering Research and General Science Volume 2, Issue 2, Feb-Mar 2014
ISSN 2091-2730

Assessment of AP, STEMI, NSTEMI and therapy Prescription based on vascular age- A Decision tree approach

C. Premalatha[1],B. Suganyadevi[2], S.Chitra[1]

[1]PG Scholar, Department of Computer Science and Engineering, Ranganathan Engineering College, Coimbatore, TamilNadu

[2] Assistant Professor, Department of Computer Science and Engineering, Ranganathan Engineering College, Coimbatore, TamilNadu

E-Mail- premalathajck@gmail.com

ABSTRACT - AP, STEMI and NSTEMI are the main categories of acute coronary syndrome which causes damage to the coronaries and make the patients prone to high risk of death. Several studies with different technologies have been made in diagnosis and treatment of the events, which includes association rules, logistic regression, fuzzy modeling, and neural network, CART, ID3. The existing techniques are confined to small datasets that are specific to one particular disease and this knowledge mined is not indispensible for classification of risk factors for the events. The implemented methodology uses C4.5 and C5.0 decision tree algorithm for identification of related risk factors by constructing two different decision trees for the events that includes Angina Pectoris, St-elevation Myocardial Infarction and Non-St-Elevation Myocardial Infarction based on attribute selection measure that includes Information Gain, Gain Ratio. Using performance measures, correctly classified values have been found for both the algorithms and accuracy is calculated. The implemented methodologies, C4.5 and C5.0 decision tree algorithm gives high classification accuracy of 86 % and 89.3% compared to the aforementioned existing techniques. Rule based classification technique provides a therapy selection for the events diagnosed, based on the vascular age, which aids the patients in reducing their risk levels and doctors to treat the patient with required therapy instead of angioplasty.

Keywords—Classification, Attribute selection measures, Information gain, Gain ratio, C4.5 and C5.0 decision tree algorithm, risk factors, Rule based classification.

INTRODUCTION

The objective of the implemented system was to develop a data mining system based on decision trees for the assessment of acute coronary syndrome related risk factors targeting in the reduction of the events. Decision-tree-based algorithms give reliable and effective results that provide high-classification accuracy with a simple representation of gathered knowledge, support decision-making processes in medicine. Data-mining analysis was carried out using the C4.5 and C5.0 decision tree algorithms extracting rules based on the risk factors (age, sex, FH, SMBEF, SMAFT, TC, TG, HDLM, HDLW, GLU, HXHTN, HXDM, SBP, DBP and LDL)

The C4.5 algorithm, which uses the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses a selected criterion to build the tree. It works top–down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split. The C5.0 algorithm boosts the constructed decision tree such that the misclassification error over the classified data is found and removed which results in higher accuracy over classified risk factors identified for the events AP, STEMI, NSTEMI.

In the implemented system, the following attribute selection measures were used: Information Gain, Gain Ratio. Based on these attribute selection measures, different decision trees are constructed. Using performance measures, training and testing datasets are compared and accuracy is calculated. Rule based classification technique provides a therapy selection for the events diagnosed, based on the vascular age, which aids the patients in reducing their risk levels and doctors to treat the patient with required therapy instead of angioplasty.

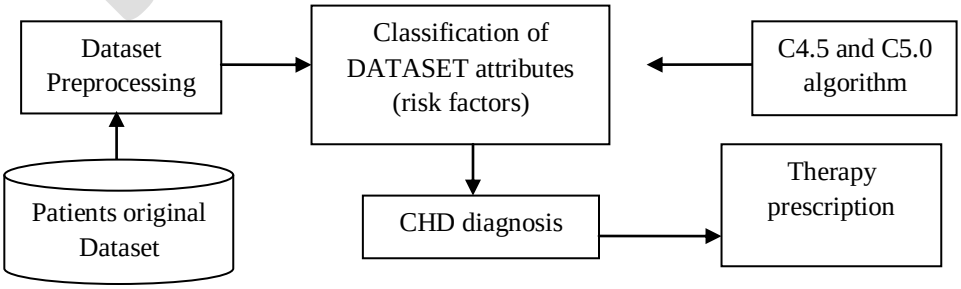


Fig.1. Block Diagram of the Acute coronary syndrome diagnosis system

DATASET PREPROCESSING

The data preprocessing is the first processing module that analyze data that has not been carefully screened, unscreened data can produce misleading results. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult.Thus, the representation and quality of data is first and foremost before any process. Steps involved in dataset preprocessing are as follows,

- Missing values are filled using K-Nearest Neighbor algorithm
- Duplications were removed
- Data were coded

The Steps involved in **filling up the missing values** are:

1. Determine parameter K = number of nearest neighbors

2. Calculate the distance between the query-instance and all the training samples

3. Sort the distance and determine nearest neighbors based on the K-th minimum distance

4. Gather the values of 'y' of the nearest neighbors

5. Use average of nearest neighbors as the prediction value of the query instance and replace the missing field with the Predicted value.

If both the row has same value that is, the values **duplicated**, then any one of the row is removed from the dataset. None of the row is removed if at least one value differs in any column of the tuple. It is done after filling up the missing values in the dataset.

if (Rown==Rowm&&Missing values==Nil) **then**
Delete (Rown||Rowm)
Else if (Rown==Rowm &&Missing values==found)
Apply K-NN
Return (Missing value: K-NN value)
Repeat until Missing values==Nil
If (Rown==Rowm) **then**
Delete (Rown||Rowm)
Else
Checkout next record
Else
Return (no duplication found)

Data coding is the process of assigning the dataset attribute values to a specified categorical or numerical value. It is done in order to make the representation of risk factors precise and classification to be done efficiently with that simpler representation.

Risk factors	***Coded values***			
Age	30-40: 1	41-50: 2	51-60: 3	60+: 4
Sex	Female: F	Male: M		
Family History	Yes: Y	No: N		
Diabetes	Yes: Y	No: N		
Hypertension	Yes: Y	No: N		
Smoking (B/A)	Yes: Y	No: N		
Systolic blood pressure	N: 120	H>140	L<100	

Diastolic blood pressure	N: 80	H>100	L<70
Total Cholesterol	N: 180	H>200	
High Density Lipoprotein	N: 50-70	H>70	L<40
Low Density Lipoprotein	N: 130	H>130	L<130
Triglyceride	N: 160	H>160	
Glucose	N: 100-140	H>145	L<60
Class	AP: 1	STEMI: 2	NSTEMI: 3

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
65	2	1	1	2	1	2	80	90	200	50	30	80	67	112	1
31	1	1	1	2	1	1	100	80	45	60	50	100	56	110	2
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
80	2	2	1	1	1	1	150	?	190	80	60	23	150	150	3

TABLE I. ORIGINAL DATASET

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
65	2	1	1	2	1	2	80	90	200	50	30	80	67	112	1
31	1	1	1	2	1	1	100	80	45	60	50	100	56	110	2
45	1	2	2	2	1	2	149	60	80	70	40	120	100	90	3
80	2	2	1	1	1	1	150	**70**	190	80	60	23	150	150	3

TABLE II. PREPROCESSED DATASET

Age	Sex	FH	SMBEF	HXHTN	HXDM	SMAFT	SBP	DBP	TC	HDLW	HDLM	LDL	TG	GLU	CL
3	N	Y	Y	N	Y	N	L	H	H	N	L	N	N	N	1
1	Y	Y	Y	N	Y	Y	N	N	N	N	N	H	N	H	2
1	Y	N	N	N	Y	N	H	N	N	H	N	H	N	N	3
4	N	N	Y	Y	Y	Y	H	N	H	H	H	N	H	H	3

TABLE III. CODED DATASET

CLASSIFICATION OF RISK FACTORS AND CHD DIAGNOSIS

The C4.5 algorithm employs a divide-and-conquer approach to construct decision tree. The algorithm uses a selected criterion to build the tree using attribute selection measures that includes Information Gain and Gain Ratio. The attribute producing highest measure thrive to be the root node based on which further splits occur. Finally, it works top–down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split.

Input:
1) Training dataset *D*, which is a set of training observations and their associated class value.
2) Attribute list *A*, the set of candidate attributes.
3) Selected splitting criteria method.
Output: A decision tree.

C4.5 decision tree construction module having the following attribute selection measures are to be investigated for training the dataset.

1. Information Gain (IG)

Information gain is based on Claude Shannon's work on information theory. InfoGain of an attribute *A* is used to select the best splitting criterion attribute. The highest InfoGain is selected to build the decision tree

$$InfoGain(A) = Info(D) - InfoA(D) \quad \text{..Eq. 1}$$

Where,

$$Info(D) = -\sum_{i=1}^{m} pi\ log2(pi) \quad \text{..Eq. 2}$$

$$InfoA(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} info(Dj) \quad \text{..Eq. 3}$$

2. Gain Ratio (GR)

Gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain

$$Gain\ Ratio(A) = \frac{Info\ Gain(A)}{SplitinfoA(D)} \quad \text{..Eq. 4}$$

$$SplitinfoA(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \, log2\left(\frac{|Dj|}{|D|}\right) \qquad ..\text{Eq. 5}$$

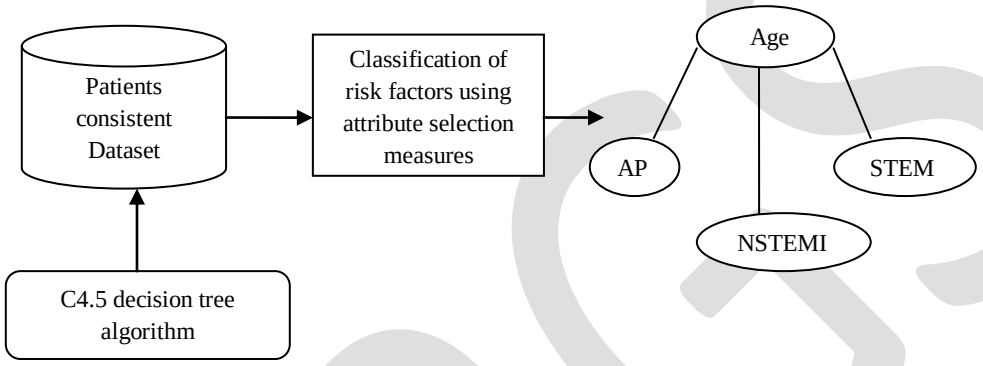


Fig.2. Classification of risk factors and CHD Diagnosis

Classification of Risk factors using Attribute selection measures for the coded dataset after Preprocessing

1. ***Information Gain(IG) - Calculated for Age***

Info Gain (A) = Info (D) − InfoA (D)

$$Info(D) = -\frac{1}{4} \, log2\left[\frac{1}{4}\right] - \frac{1}{4} log2\left[\frac{1}{4}\right] - \frac{2}{4} log2\left[\frac{2}{4}\right] = 0.4515$$

$$InfoA(D) = 2/4\left(-\frac{1}{4} \, log2\left[\frac{1}{4}\right] - \frac{1}{4} log2\left[\frac{1}{4}\right] - 0\right) + 2/4\left(0 - 0 - \frac{2}{4} log2\left[\frac{2}{4}\right]\right) = 0.2257$$

$$Info\ Gain(A) = 0.4515 - 0.2257 = 0.2258$$

2. ***Gain Ratio(GR) - Calculated for Family history***

$$\frac{Info\ Gain(A)}{SplitinfoA(D)} = \frac{0.2258}{-\frac{2}{4} log2\left[\frac{2}{4}\right] - \frac{2}{4} log2\left[\frac{2}{4}\right]}$$

$$= \frac{0.2258}{0.3010} = 0.7501$$

Attribute having highest Gain Ratio is considered to be the root node based on which further classification of risk factors proceeds. The heart disease dataset obtained from UCI Repository contains 250 records in which 150 are considered as training dataset and 100 as testing dataset.

Deploying of C4.5 algorithm over the training dataset results in a decision tree construction, for which attribute Age produces highest measure over the other splitting criterion such that attribute age is assigned to be the root node, based on the latter further classification of risk factors occurs. The accuracy of risk factors classification obtained using C4.5 algorithm is 86% which is higher than its predecessors ID3, CHI-SQUARED STASTICS and CART.

.

C5.0 algorithm offers boosting to generate several classifiers on the training data. When an unseen example is encountered to be classified, the predicted class of the example is a weighted count of votes from individually trained classifiers. C5.0 creates a number of classifiers by first constructing a single classifier. A second classifier is then constructed by re-training on the examples used to create the first classifier, but paying more attention to the cases of the training set in which the first classifier, classified incorrectly. As a result the second classifier is generally different than the first.

- Choose K examples from the training set of N examples each being assigned a probability of 1/N of being chosen to train a classifier.
- Classify the chosen examples with the trained classifier.
- Replace the examples by multiplying the probability of the misclassified examples by a weight B.
- Repeat the previous three steps X times with the generated probabilities.
- Combine the X classifiers giving a weight log (BX) to each trained classifier.

BOOSTING PROCESS:

STEP1: Take N = 250; k = 250
Probability of P (N) = 1/N = 1/250 = 0.004

STEP2: Classification using attributes selection measures

STEP3: Weight (B) = 1, 2, 3(class label)
Probability of misclassified records, Pm (K) = 1/10
Probability of records,
P (K1) = Pm (K)*B1 = 1/10*1 = 0.1 (class 1)
P (K2) = Pm (K)*B2 = 1/10*2 = 0.2 (class 2)
P (K3) = Pm (K)*B3 = 1/10*3 = 0.3 (class 3)

STEP4: Repeat for all misclassified records (k=10) such that total time of execution, X=3

STEP5: Combine the classifier by assigning Weight to each classified record
Weight = log (B3*X) = log (3*10) = 1.4

STEP6: Total misclassification error = Total probability * weight assigned for misclassified record.
Error = 0.004 *1.4
= 0.0056% for 10 misclassified records

Rule Sets

C5.0 can also convert decision trees into rule sets. This is due to the fact that rule sets are easier to understand than decision trees and can easily be described in terms of complexity. That is, rules sets can be looked at in terms of the average size of the rules and the number of rules in the set.

Rules can be represented as follows.

Rule No. : (Records Manipulated/Records with positive result, decision branch (lift))
Attribute-1 ……Attribute - n
Class label [accuracy]

DECISION TREE CONSTRUCTION:

Read 250 cases (16 attributes) from heart disease. Data
Age = 1:
:....Diabetes-Y
: : :..DBP-H: 3(60/10)
: : :..DBP-N: 1(15/7)
: : :..DBP-L:
: : :..HDL-L: 2(10/3)

: :..Diabetes-N: 1(70/5)
Age = 2
:....Family History-Y:
: :..TCL-H: 2(8/2)
: :..TCL-N: 1(20/6)
:.Family History-N
:..SBP-H: 3(18/12)
:..SBP-N: 1(5/2)
:..SBP-L: 2 (13)

RULE SET GENERATION:

Rule 1: (60/10, lift 1.2)
DBP-H
Diabetes-Y
Age=1
--> Class 3 [0.889]

Rule 2: (8/2, lift 1.2)
TCL-H
Family History-Y
Age=2
--> Class 2 [0.905]

Rule 3: (5/2, lift 1.6)
SBP-N
Family History -N
Age=2
--> Class 1 [0.872]

Rule 4: (10/3, lift 1.5)
HDL-L
DBP-L
Diabetes-Y
Age=1
--> Class 2 [0.883]

Evaluation on training data (100 cases):
Rules

No	Errors
4	100 (10.7%)

(1)	(2)	(3)	<-classified as
----	----	----	(1) Class 1
23	27	50	(2) Class 2
			(3) Class 3

C5.0 algorithm provides high classification accuracy of 89.3% by employing boosting over misclassified records and generation of rule set for decision making process to be more precise, understandable and efficient. Compared to C4.5 algorithm, C5.0 is less time consumption, reduce error rate, simple to interpret and produces more accurate result.

VASCULAR AGE (THERAPY PRESCRIPTIONS)

It is based on Rule-based classification in which rules are set for risk factors and those that satisfy the rules is considered for further process of vascular age determination and therapy prescription**.** Risk Factors used for prescribing therapy for CHD patients are Age, HDL, Smoking, Diabetes, SBP and TC. Each factor has its own score such that summation of all the factors gives a total scores

value that determines vascular age. A therapy is prescribed for a patient based on his/her vascular age that includes nitrate, statin, aspirin, Ace inhibitor, Beta blocker, etc.,

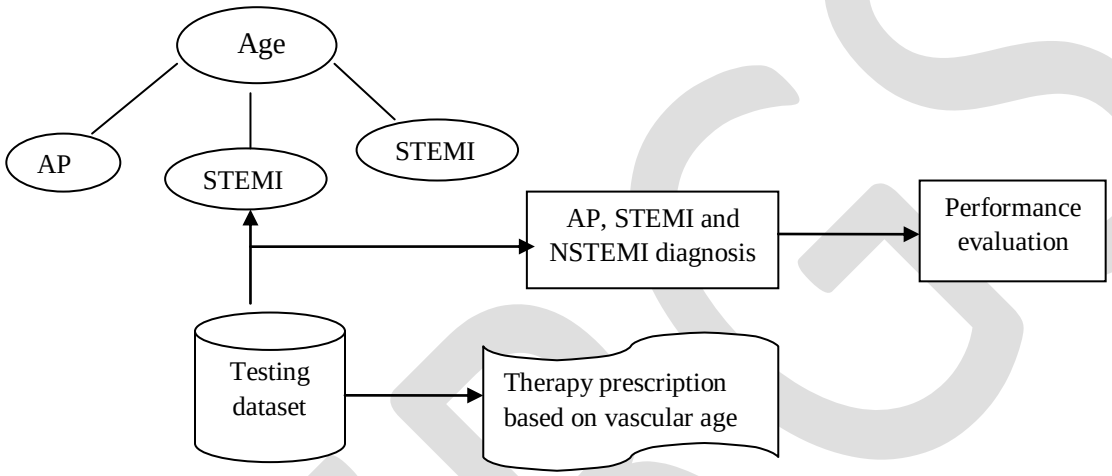


Fig.3. Vascular age determination and Therapy Prescription

TABLE IV AGE SCORE

Age	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75+
Men	0	2	5	6	8	10	11	12	14	15
Women	0	2	4	5	7	8	9	10	11	12

TABLE V TOTAL CHOLESTEROL SCORE

TC	<160	160-199	200-239	240-279	280+
Men	0	1	2	3	4
Women	0	1	3	4	5

TABLE VI HIGH DENSITY LIPOPROTEIN SCORE

HDL	<35	35-44	45-49	50-59	60+
Men	2	1	0	-1	-2
Women	2	1	0	-1	-2

TABLE VII SYSTOLIC BLOOD PRESSURE SCORE

SBP	<120	120-129	130-139	140-149	150-159	160+

Men	0	2	3	4	4	5
Women	-1	2	3	5	6	7

TABLE VIII SMOKING AND DIABETES SCORE

Smoking	No	Yes	Diabetes	No	Yes
Men	0	4	Men	0	3
Women	0	3	Women	0	4

TABLE IX TOTAL POINTS FOR VASCULAR AGE DETERMINATION

Total Points	< -1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
Vascular Age: Men	<30	30	32	34	36	38	40	42	45	48	51	54	57	60	64	68	72	76	80+
Vascular Age: Women	<30	<30	31	34	36	39	42	45	48	51	55	59	64	68	73	79	80	80	80+

In case, if the training dataset contains a patient record having above specified risk factors as in Table IX, then all their corresponding risk factor scores are retrieved from the stored score tables and finally summation of all the retrieved scores for the specific risk factors gives a total points based on which vascular age is determined for the patient diagnosed with CHD. Here, his Total score points are 9, for which his vascular age is determined to be 51. For such cases, the therapy or medication to be prescribed are Nitrate, Statin, Ace-inhibitor and Beta-blocker. If the prescribed medicine has no effect or curing or reducing the risk factor levels then angioplasty is preferred for the patient.

TABLE X VASCULAR AGE CALCULATION

No.	**Risk Factor**	**Risk Score**
1	Age	2
2	Sex	M
3	TC	1
4	HDL	1
5	SBP	2
6	Smoking	0
7	Diabetes	3
Total Points		9
Vascular age		51

PERFORMANCE EVALUATION

In order to evaluate the performance of C4.5 and C5.0 algorithms, the following factors are to be investigated.

1) Correct classifications (%CC): is the percentage of the correctly classified records.

$$\%CC = (TP + TN)/N$$

2) True positive rate (%TP): corresponds to the number of positive examples correctly predicted by the classification model.
3) False positive rate (%FP): corresponds to the number of negative examples wrongly predicted as positive by the classification model.
4) True negative rate (%TN): corresponds to the number of negative examples correctly predicted by the classification model.
5) False negative rate (%FN): corresponds to the number of positive examples wrongly predicted as negative by the classification model.
6) Sensitivity: is defined as the fraction of positive examples predicted correctly by the model.

Sensitivity = TP/ (TP +FN)

7) Specificity: is defined as the fraction of negative examples predicted correctly by the model.

Specificity = TN/ (TN+FP)

8) Support: is the number of cases for which the rule applies (or predicts correctly); that is, if we have the rule X → Z, Support is the probability that a transaction contains {X, Z}.

$$\text{Support} = P(XZ) = \text{no of cases that satisfy X and Z}/|D|$$

9) Confidence: is the number of cases for which the rule applies (or predicts correctly), expressed as a percentage of all instances to which it applies, that is, if we have the rule X → Z, Confidence is the conditional probability that a transaction having X also contains Z

$$\text{Confidence} = P(Z|X) = P(XZ)/P(X)$$

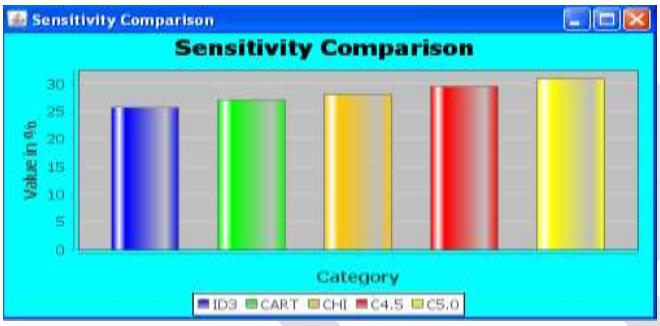


Fig.4. Sensitivity comparison of Decision tree algorithms

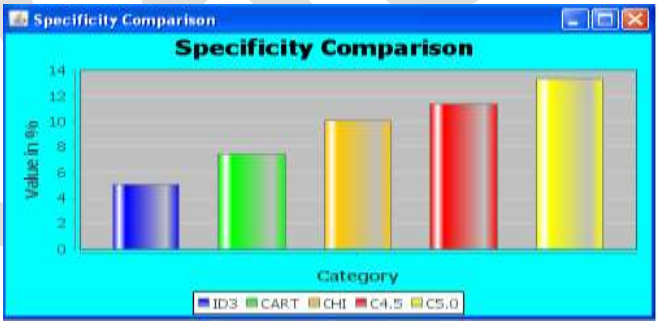
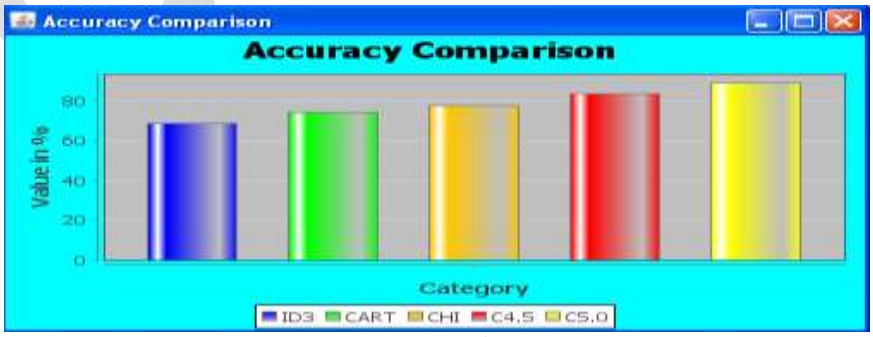


Fig.5. Specificity comparison of Decision tree algorithms

RESULT ANALYSIS

C4.5 and C5.0 algorithm used attribute selection measures for constructing two different decision trees. The training and testing datasets were compared after decision tree construction for finding out correctly classified values. Using Performance measures, the dataset's attribute value has been correctly classified and accuracy is calculated. C4.5 and C5.0 decision tree algorithm gives high classification accuracy of 86 % and 89.3%.Accuracy comparison graph proves the accuracy of classification of risk factors for the events AP, STEMI and NSTEMI such that high classification accuracy of 89.3% is obtained by deploying C5.0 decision tree algorithm over the datasets.



\

Fig.6. Accuracy of Decision tree algorithms

Finally, testing dataset value is used for determination of vascular age, based on which a specific therapy is prescribed for a patient diagnosed with CHD.

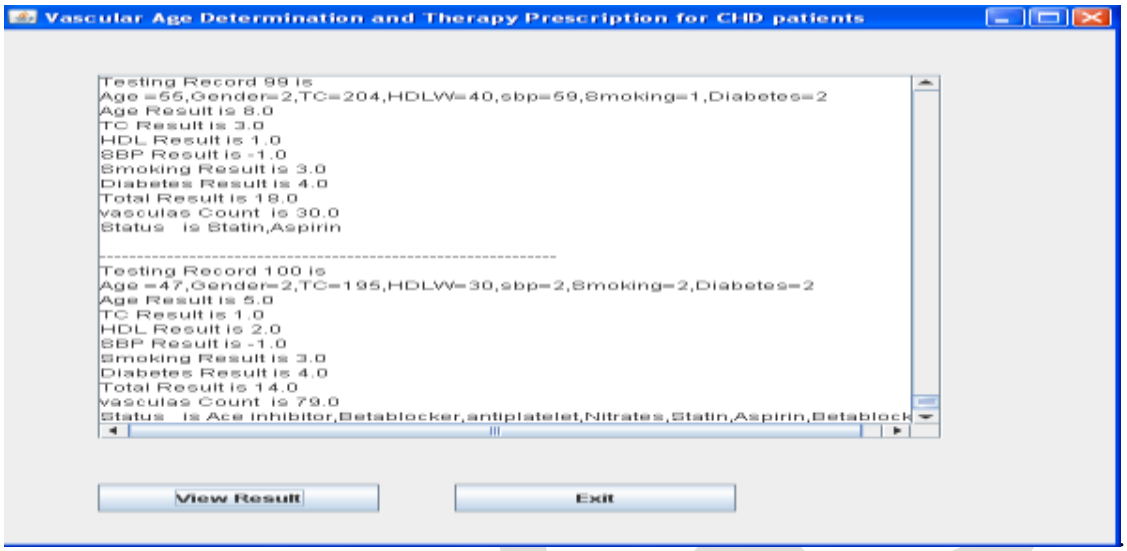


Fig.7.Vascular Age determination and therapy selection for CHD patients

DIFFERENCE BETWEEN C4.5 AND C5.0 DECISION TREE CONSTRUCTION

C4.5 DECISION TREE CONSTRUCTION	C5.0 DECISION TREE CONSTRUCTION
Handles discrete and continuous data only	Handles dates, times, timestamps, ordered discrete attributes, and categorical data
Attribute selection measures used are • Information gain • Gain ratio	Attribute selection measure used is Distance measure alone. So, reduces manipulation, time consumption.
No boosting over decision tree construction and classification.	Boosting plays a significant role in it by assigning weights to the decision tree formed and reduces misclassification error. Thus, increases accuracy.
Produces result in the form of decision tree	Produces result in the form of rule set which is more precise and ease to interpret.

CONCLUSION

The implemented methodology uses C4.5 and C5.0 decision tree for assessment of acute coronary syndrome related risk factors and reduction of the events that includes Angina Pectoris, St-elevation Myocardial Infarction and Non-St-Elevation Myocardial Infarction. C4.5 Decision tree algorithm identifies most important risk factors for the events using attribute selection measures whereas C5.0 algorithm uses attribute selection measure for classification ,boosting for increasing the accuracy over classified risk factors and rule set generation for making decision more accurate and precise. Accuracy obtained by deploying C4.5, C5.0 algorithm is 86% and 89.3% which justifies that C5.0 algorithm has highest accuracy compared to other decision tree approaches that includes ID3, CHI-SQUARED STATISTICS, GINI INDEX, CART and C4.5. Rule based classification is used for determination of vascular age based on which a specific therapy is prescribed for a patient diagnosed with the events.

FUTURE WORK

Future work involves in decision tree construction for more events instead of finding for limited number of events with large dataset values and also grouping of different diseases and generating rules separately for diagnosis, therapy prescription of different events rather than finding specific disease, such that it makes clinicians to interpret the result for several disease at once in case of emergencies.

REFERENCES

[1] ARIHITO ENDO, TAKEO SHIBATA, HIROSHITANAKA 'Comparison of Seven Algorithms to Predict Breast Cancer Survival' Biomedical Soft Computing and Human Sciences, Vol.13, No.2, pp.11-16 (2008)
[2] C. A. Pena-Reyes,(2004) 'Evolutionary fuzzy modeling human diagnostic decisions,' Ann. NY Acad. Sci., vol. 1020, pp. 190–211.
[3] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy, (1998) 'Using classification trees and logistic regression methods to diagnosemyocardial infraction,' in Proc. 9th World Congr. Med. Inf., vol. 52, pp. 493–497.
[4] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra,J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia,(2001) 'Mining constrained association rules to predict heart disease,' in Proc. IEEE Int.Conf. Data Mining (ICDM 2001), pp. 431–440.
[5] J. Han and M. Kamber, (2001) 'Data Mining, Concepts and Techniques', 2nd ed. San Francisco, CA: Morgan Kaufmann.
[6] J. R. Quinlan,(1987) 'Simplifying decision trees', Int. J. Man-Mach. Stud.,vol. 27, pp. 221–234.
[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, (1986),' Classification and Regression Trees', Belmont, CA: Wadsworth Int. Group.
[8] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis, (2009),'Association rule analysis for the assessment of the risk of coronary heart events', in Proc. 31st Annu. Int. IEEE Eng. Med. Biol. Soc. Conf., Minneapolis, MN, Sep. 2–6, 2009, pp. 6238–6241.
[9] M. Karaolis, J. A.Moutiris, and C. S. Pattichis, (2008) 'Assessment of the risk of coronary heart event based on data mining', in Proc. 8th IEEE Int. Conf.Bioinformatics Bioeng, pp. 1–5.
[10] Phayung Meesad and Kairung Hengpraprohm, (2008) 'Combination of KNN-Based Feature Selection and KNN-Based Missing-Value Imputation of Microarray Data,'the 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08) IEEE computer society.
[11] R. B. Rao, S. Krishan, and R. S. Niculescu,(2006) 'Data mining for improved cardiac care', ACM SIGKDD Explorations Newslett., vol. 8, no. 1, pp. 3–10.
[12] R. Lopez de Mantras, (1991) 'A distance-based attribute selection measure for decision tree induction', Mach. Learn., vol. 6, pp. 81–92, 1991.
[13] S. A. Pavlopoulos, A. Ch. Stasis, and E. N. Loukis, (2004) 'A decision treebased method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds', Biomed. Eng. OnLine, vol. 3, p. 21.
[14] Subrata Paramanik,Utpala Nanda Chowdhury,(2010) 'A Comparative Study of Bagging,Boosting and C4.5: The Recent Imprivements indecision Tree Learning Algorithm',Asian Journal of Information Technology.
[15] S. M. Grundy, R. Pasternak, P. Greenland, S. Smith, and V. Fuster, (1999) 'Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations', Amer. Heart Assoc., vol. 100, pp. 1481–1492, 1999.
[16] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman,(2002) 'Decision trees: An overview and their use in medicine', J. Med. Syst., vol. 26, no. 5, pp. 445– 463.
[17] http://archive.ics.uci.edu/ml/datasets/Heart+Disease
(heart disease dataset).
[18] Universal Medicare center, Coimbatore-Consulted Dr. Mohan, Cardiologist, for heart disease and risk factors related information and real time dataset collection.