International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 196–206

Classification and Regression Tree (CART) Algorithm for the Prediction of Ischemic Heart Disease

Chaithra N[1], Madhu B[2*], Umamaheswari K[3], Balasubramanian S[4]

[1]*Assistant Professor, Division of Medical Statistics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India., Email id: chaithra.mstats@jssuni.edu.in*

[2*]*Associate Professor, Department of Community Medicine, JSS Medical College & Assistant Director Research, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India, Email id: madhub@jssuni.edu.in*

[3]*Professor, Department of Information Technology, PSG College of Technology, Coimbatore, Tamil Nadu, India, Email id: uma@ity.psgtech.ac.in*

[4]*Director (Research) and Dean, Department of Water and Health, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India, Email id: director_research@jssuni.edu.in*

Abstract

CART is a classification technique, which generates a binary decision tree, namely every single interval node having only two children To decide which attributes requires splitting and when it has to be done, CART searches all possible splits of all the attributes and chooses the best split based on the impurity measure of the Gini index method. A Retrospective study contains a total of 7304 patients echocardiography records with one dependent variable and twenty-five independent variables. The "diagnosis" attribute was recognized as a predicted attribute with the value of "1" for patients with IHD and value "0" for the patients with no IHD. There are a maximum of five tree depths, as well as a minimum of 40 and 20 cases in the parent node and in the child node respectively, used in the analysis in order to construct the model and the model thus resulted from the process, portrays information based on the 21 of the total number and 11 terminal nodes, 5 tree depths and 9 parameters were include in the final model for disease prediction. Node 0 is the overall probability of diagnosis; it shows the 1113(15%) proposition of the patients have with IHD and 6191(85%) patients without IHD. FS was identified to be the highly influencing factor for IHD and others attributes i.e. EF, ESV, MR, LVID_s, LA, age, LVID_d, and AV _max are key factors in determining patients with heart disease and strongest interaction with the response variable. This result shows that the IHD of a patient are predicted successfully with an acceptable ratio of 94 %. Furthermore, the true negative rate of the resulting model is high and significant rules were extracted from a dataset that makes the application of Decision tree in predicting IHD in healthcare.

Keywords: *Echocardiography data, CART algorithm, Gini index, Prediction performance measures*

1. INTRODUCTION

The World Health Organization (WHO) estimates that 17.6 million deaths occur from cardiovascular diseases (CVDs) and this was expected to increase up to 24.2 million by 2030, over three million of these mortalities took place before 60 years of age and it could have been avoided to a great extent. CVDs include diseases of the heart and blood vessels [1]. Ischemic Heart Disease (IHD) is characterized as insufficient circulation of blood to a confined region, which is the result of blockage in the blood vessels that supplies to the heart muscle and it can be diagnosed in several ways. An Echocardiography

ISSN: 2233-7857 IJFGCN

(ECHO)**,** is an ultrasound test used to view moving pictures of the heart on a screen**.** It is used to detect and evaluate a variety of conditions, including heart valve problems, abnormal heart rhythms, congenital heart disease, heart murmurs or infections involving the heart.

Classification and Regression Trees (CART) are extensively being used in the medical domain in developing a prediction rule and it was evolved in the early 1980s by statisticians, Leo Breiman [2]. It is useful in explaining the dependence of Y variable on several independent variables $(X_i, i = 1, 2, ..., n)$ through binary division of a set of values of Y, recursively in accordance with X-values, the explanatory variable X_i may be a mixture of categorical and continuous variables [3]. CART is a classification technique, which generates a binary decision tree, namely every single interval node having only two children. Decision tree outcomes are depicted as a tree diagram utilizing a set of basic if-then rules [4,5]. The CART methodology refers to the two types of decision trees are classification and regression, the decision trees are split into regression trees when the dependent variable I s continuous and the response variable is categorical, it splits into classification trees [6,7]. For deciding which attributes requires splitting and when it has to be done, CART searches all possible splits of all the attributes and chooses the best split based on the Gini index method [5,8]. This procedure continues until all child nodes are homogeneous. CART analysis is an effective procedure with great capability and clinical utility [9].

2. METHODOLOGY

2.1 Study subjects and data set

A Retrospective study contains a total of 7304 patients echocardiography records with one dependent variable and twenty-five independent variables such as AO, Aortic Root; LA, Left Atrium; RV, Right Ventricle, LVID_D, LVID_S, Left Ventricle Internal Diastole during Systole; Left Ventricle Internal Diameter during Diastole; IVS_S, Intact Ventricular Septum Diastole during Systole; IVS_D, Intact Ventricular Septum Diameter during Diastole; LVPW_S, Left Ventricular Posterior Wall Diastole during Systole; LVPW_D, Left Ventricular Posterior Wall Diameter during Diastole; ESV, End Systolic Volume; EDV, End Diastolic Volume; EF (%), Ejection Fraction; FS(%),SV, Stroke Volume; Fractional Short; MV_E, MV_A, Mitral valve - ratio of the early (E) to late (A) ventricular filling velocities; MR, Mitral regurgitation; TV_E, TV_A, Tricuspid valve- ratio of the early (E) to late (A) ventricular filling velocities; TR, Tricuspid regurgitation ; AV_VMAX, Aortic Valve- The maximal aortic jet velocity; AR, Aortic regurgitation; PV_VMAX, Pulmonary Vascular - The maximal Pulmonary jet velocity; PR Pulmonary regurgitation were obtained from the echocardiography measurements was recorded at Cardiology Department, JSS Hospital. To develop a prediction model that can predict IHD cases based on the collected information. The “diagnosis” attribute was recognized as a predicted variable with the value of “1” for patients with IHD and value “0” for the patients with no IHD. The models were developed with CART decision Tree using SPSS and machine learning software, WEKA 3.6.4.

2.2 Classification and Regression Trees Algorithm

The CART analysis makes use of binary recursive partitioning for creating a tree with each node T, as the partition cell and A, B, C representing the leaves or terminal nodes, implying that after this split, further splitting of the data does not describe sufficiently the variance to be relevant to the Y description, as represented in figure 1. In descending order of priority, the graphic is considered as the domain of all the variables connected with our Y. The term “binary” implies that each group of patients, represented by a “node” in a decision tree, can only be split into two groups. As a result, every single node can be divided into 2 child nodes, where the original node is considered the parent node. Thus, every single parent node can be revived to 2 child nodes, and each of them, in turn, may split themselves,

ISSN: 2233-7857 IJFGCN

forming additional children. The word "partitioning" particularly denotes that the dataset is partitioned or broken down into sections [9,10].

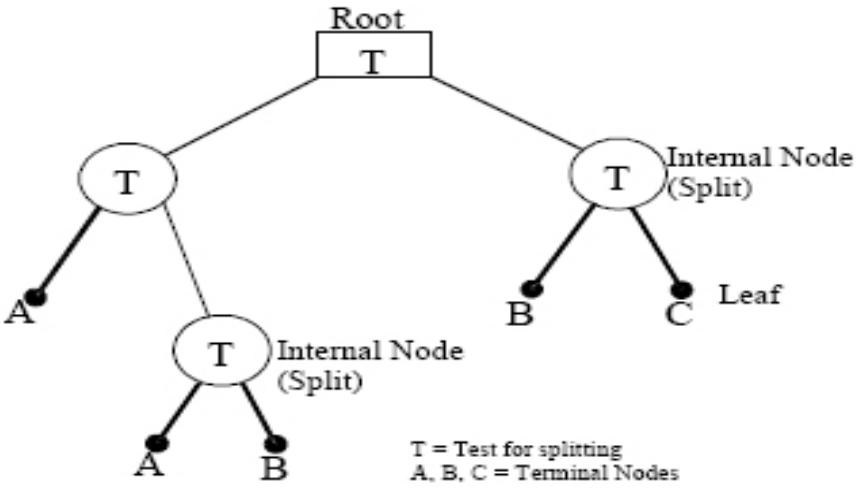

Figure 1: Structure of CART.

2.3 Attribute Selection Measures by using Gini Index

The Gini index or Gini impurity is a criterion focused on the impurity that measures the divergences between the probability distributions of the independent value. To select the best split of attributes, CART uses the Gini method to create binary splits and it is used for the measurement of data impurity.

The Formula for the calculation of the Gini Index is given below.

$$GiniIndex = 1 - \sum_j p_j^2$$

Wher $p_j = n(i \setminus t) / n(t)$, where i refers to a target variable class (in this study j = 0 implies, with IHD and j = 1 denotes without IHD), $n(i \setminus t)$ is the aggregate records of node t that belongs to the class i, and n(t) is the overall record number in the t node [11,12].

$$Gini(t)_{split} = \frac{n(t_L)}{n(t)} Gini(t_L) + \frac{n(t_R)}{n(t)} Gini(t_R)$$

Where t_R and t_L are the left and right child nodes of node t. The attribute that minimizes the $\text{Gini(t)}_{\text{split}}$ is chosen to split the node.

2.4 Pseudocode for tree construction

1. Begin a single tree with a root node.

2. Determine the set S for which the node impurities sum in both the child nodes is minimum, for a given X and pick the split $\{X^* \in S^*\}$ giving the minimum X and S altogether.

3. Exit, once the stopping criterion is attained or else continue with the application of step 2 to every single child node in turn.

ISSN: 2233-7857 IJFGCN

3. RESULTS AND DISCUSSION

The purpose of this study is to predict the IHD using classification and regression tree, the experiments were conducted on the dataset using 6191 patients without IHD and 1113 patients with IHD. The process of creating a decision tree works by greedily selecting the best split point to make predictions and repeating the process until the tree is a fixed depth. After the tree is constructed, it is pruned to improve the model's ability to generalize to new data. The model summary table1 indicates information about the specifications used to generate the CART tree model, including the one dependent variable i.e. diagnosis and twenty-six independent variables were specified as Age, AO, LA, RV, L VID_d, L VID_s, IVS_d, IVS_S, LVPW_d, LVPW_s, EDV, ESV, SV, EF(%), FS(%), MV_E, MV_A, MR, TV_E, TV_A, TR, AV_VMAX, AR, PV_VMAX, PR and five maximum tree depth, 40 minimum cases in parent node and 20 minimum cases in child node used in the analysis to build the model and the resulting model displays information on the 21 number of total and 11 terminal nodes, 5 depth of the tree and 9 parameters were included in the final model to predict the diseases.

Table1: Model summary in the SPSS output.

Specifications	Growing Method	CRT
	Dependent Variable	Diagnosis
	Independent Variables	Age, AO, LA, RV, L VID_d, L VID_s, IVS_d, IVS_S, LVPW_d, LVPW_s, EDV, ESV, SV, EF(%), FS(%), MV_E, MV_A, MR, TV_E, TV_A, TR, AV_VMAX, AR, PV_VMAX, PR
	Validation	None
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	40
	Minimum Cases in Child Node	20
Results	Independent Variables Included	FS(%), EF(%), L VID_s, ESV, EDV, L VID_d, LA, MR, MV_A, IVS_S, TR, IVS_d, MV_E, LVPW_d, TV_E, PV_VMAX, AV_VMAX, TV_A, RV, LVPW_s, AR, SV, AO, Age, PR
	Number of Nodes	21
	Number of Terminal Nodes	11
	Depth	5

Table 2: Classification table.

	Predicted		
Observed	**Yes**	**No**	**Percent Correct**
Yes	908	205	81.6%
No	231	5960	96.3%
Overall Percentage	15.6%	84.4%	94.0%

The table 2 shows the number of cases classified correctly and incorrectly for each category of the dependent variable where the confusion matrix provides overall accuracy of classification is 94 %. The CART model correctly classifies 908 patients are having the disease, but misclassified 205 patients were identified as not having the disease but they had the disease, it correctly classified 81.6% of cases and 231 patients were classified as having the disease but were free from disease and 5960 patients were rightly diagnosed and were predicted as free from disease, it correctly classified 96.3 % of cases.

ISSN: 2233-7857 IJFGCN

The table 3 provide a quick evaluation of how well the model fits. The Estimate value of 0.06 indicates that the proportion of cases incorrectly classified after adjustment for prior probabilities by the model is wrong for six percent of the cases with Std. Error value is 0.003. So, the "risk" of IHD patients is approximately 6 %. The prediction performance measures of the model were evaluated using the standard metrics with 0.94% (95% CI = 0.935 - 0.945) of accuracy in predicting ischemic heart disease. True Positive and True Negative is the correctly classified positives and negatives, False Positive and False Negative is the incorrectly classified positives and negatives. The result was obtained with 6868 correctly classified instances and 436 incorrectly classified instances which represent 0.94 and 0.6 respectively. Overall True Positive rate of 0.826 (95% CI = 0.797 - 0.833), True Negative rate of 0.963 (95% CI = 0.959 - 0.966), Negative Predicted Value of 0.967 (95% CI = 0.963 - 0.970), Positive Predicted Value of 0.797 (95% CI = 0.779 - 0.814), Kappa statistics is 0.771 (95% CI =0.750 - 0.791) , F – Score of 0.806, Gini index value of 0.88and AUC of 0.94. Summary information for each node in the tree, including parent node number, counts and percentages for categorical dependent Variables, predicted category, improvements, split values and independent factors for final model fit are display in table 4.

Table 3: Prediction performance measures.

Estimate	0.06
Std. Error	0.003
True Positive rate (Sensitivity)	0.826 (95% CI = 0.797 - 0.833)
True Negative rate (Specificity)	0.963 (95% CI = 0.959 - 0.966)
Negative Predicted Value	0.967 (95% CI = 0.963 - 0.970)
Positive Predicted Value	0.797 (95% CI = 0.779 - 0.814)
Accuracy	0.940 (95% CI = 0.935 - 0.945)
Kappa	0.771 (95% CI =0.750 - 0.791)
Precision	0.797
F - Score	0.806
Correctly classified instances	6868(0.94)
Incorrectly classified instances	436(0.6)
Gini	0.88
Area under (ROC)	0.94

ISSN: 2233-7857 IJFGCN

Table 4: CART Tree in table format.

Node	Yes		No		Total		Predicted Category	Parent Node	Primary Independent Variable		
	N	Percent	N	Percent	N	Percent			Variable	Improvement	Split Values
0	1113	15.2%	6191	84.8%	7304	100.0%	No				
1	88	1.6%	5567	98.4%	5655	77.4%	No	0	FS(%)	.128	>30
2	1025	62.2%	624	37.8%	1649	22.6%	Yes	0	FS(%)	.128	<30
3	51	1.0%	5253	99.0%	5304	72.6%	No	1	EF(%)	.001	>60
4	37	10.5%	314	89.5%	351	4.8%	No	1	EF(%)	.001	<60
5	387	42.8%	517	57.2%	904	12.4%	No	2	ESV	.021	<= 46.5
6	638	85.6%	107	14.4%	745	10.2%	Yes	2	ESV	.021	> 46.5
7	146	26.7%	400	73.3%	546	7.5%	No	5	MR	.010	Nil
8	241	67.3%	117	32.7%	358	4.9%	Yes	5	MR	.010	0.25; 0.5; 0.75
9	289	75.3%	95	24.7%	384	5.3%	Yes	6	L VID_s	.002	<= 39.5
10	349	96.7%	12	3.3%	361	4.9%	Yes	6	L VID_s	.002	> 39.5
11	33	15.0%	187	85.0%	220	3.0%	No	7	LA	.001	<= 28.5
12	113	34.7%	213	65.3%	326	4.5%	No	7	LA	.001	> 28.5
13	40	51.3%	38	48.7%	78	1.1%	Yes	9	Age	.002	<= 45.5
14	249	81.4%	57	18.6%	306	4.2%	Yes	9	Age	.002	> 45.5
15	74	29.0%	181	71.0%	255	3.5%	No	12	L VID_d	.001	<= 46.50
16	39	54.9%	32	45.1%	71	1.0%	Yes	12	L VID_d	.001	> 46.50
17	30	69.8%	13	30.2%	43	0.6%	Yes	13	AV_VMAX	.001	<= 118.0
18	10	28.6%	25	71.4%	35	0.5%	No	13	AV_VMAX	.001	> 118.0
19	63	66.3%	32	33.7%	95	1.3%	Yes	14	MR	.001	Nil
20	186	88.2%	25	11.8%	211	2.9%	Yes	14	MR	.001	0.25; 0.5; 0.75; 1

ISSN: 2233-7857 IJFGCN

The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. The list of predictors with relatively decreasing normalized importance is as follows: FS(%), EF(%), LVID_s, ESV, LVID_d, MR, EDV, LA, IVS_d, IVS_S, LVPW_d, MV_A, TR, PV_ VMAX, LVPW_s, Age, MV_E, AV_VMAX, RV, SV, TV_E, AR, PR, and AO. Normalized importance is simply the importance values divided by the largest importance values and expressed as percentages. Table 5 illustrate FS (%) has high significant independent variable esteem 0.128 (importance value) with standardized significance 100% (normalized importance) and AO has the least important value with 0.3 % normalized importance in predicting IHD using CART model. The bar chart of model ranks each predictor variable according to its normalized importance with importance value is display in figure 2.

Table 5: Independent Variable Importance.

Independent Variable	Importance	Normalized Importance
FS (%)	0.128	100.00%
EF (%)	0.111	86.30%
L VID_s	0.107	83.00%
ESV	0.105	81.90%
L VID_d	0.057	44.20%
MR	0.054	42.00%
EDV	0.053	41.20%
LA	0.034	26.50%
IVS_d	0.014	10.60%
IVS_S	0.013	10.20%
LVPW_d	0.011	8.70%
MV_A	0.009	6.80%
TR	0.008	6.50%
PV_VMAX	0.005	4.20%
LVPW_s	0.005	4.10%
Age	0.005	3.50%
MV_E	0.003	2.70%
AV_VMAX	0.003	2.70%
RV	0.003	2.70%
SV	0.003	2.20%
TV_E	0.003	2.10%
TV_A	0.001	1.10%
AR	0.001	0.80%
PR	0	0.30%
AO	0	0.30%

ISSN: 2233-7857 IJFGCN

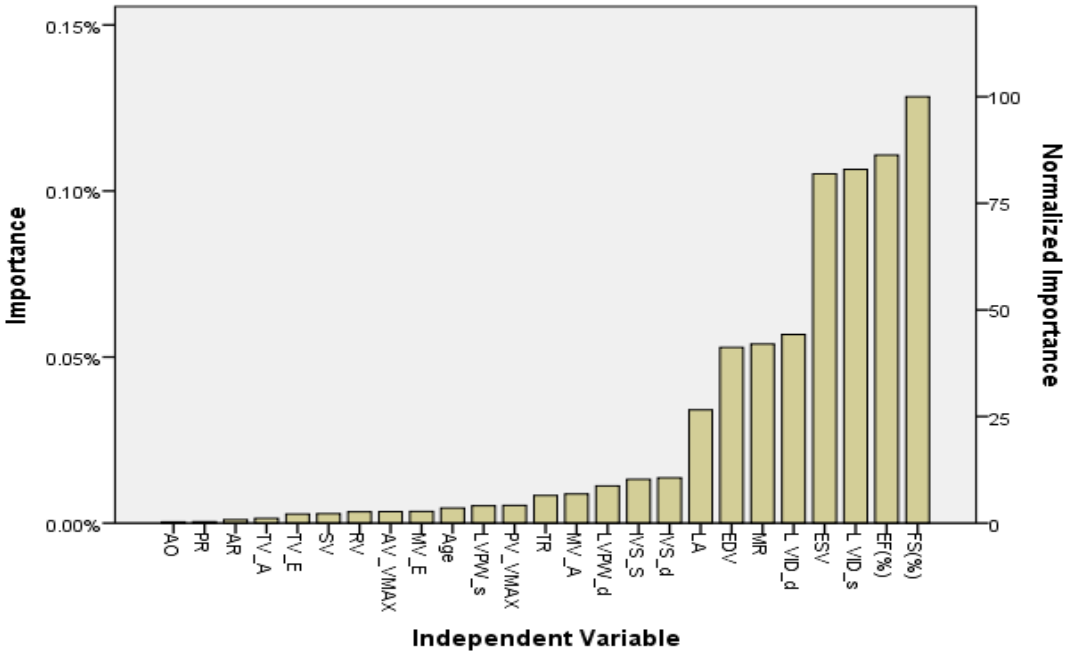

Figure 2: Normalized importance of echo parameters.

Results of the CART tree, extracted rules for determining independent factors significantly influenced on a prediction of IHD. The model spanning 21 nodes with a total depth of 5 nodes. Starting from the root node, a path of successive decisions is realized until a leaf node is reached to predict the categorical variables. Node 0 is the overall probability of diagnosis; it shows the 1113(15%) proposition of the patients have with IHD and 6191(85%) patients without IHD as illustrated in figure 3. Node 0 was divided into two arms, Node 1 and Node 2 respectively, FS was found to be the most influential factor for IHD and others attributes i.e EF, ESV, MR, LVID_s, LA, age, LVID_d, and AV _max are key factors in determining patients with heart disease and strongest interaction with the response variable. The following rules associated with IHD in the CART-based decision model were found to be statistically significant based on the CART algorithm.

1. IF FS > 30 AND EF >60, THEN Diagnosis = "NO"
   This rule is strong in identifying normal patients who are free from IHD.
2. IF FS <30 AND ESV<= 46.5 AND MR =Nil AND LA<= 28.5, THEN Diagnosis = "NO"
3. IF FS <30 AND ESV<= 46.5 AND MR =Nil AND LA > 28.5 AND, LVID_D <= 46.50 THEN Diagnosis = "NO"
   Based on rule2 and rule3, patients are free from the Disease.
4. IF FS <30 AND ESV > 46.5 AND LVID_s > 39.5, THEN diagnosis = "YES
5. IF FS <30 AND ESV > 46.5 AND LVID_s <=39.5 AND Age > 45.5 AND MR <=1 THEN diagnosis = "YES"
6. IF FS <30 AND ESV > 46.5 AND LVID_s <=39.5 AND Age <= 45.5 AND AV_VMAX <= 118 THEN diagnosis = "YES"
   Rule 4, 5 and 6 indicates patients are in high risk of having Disease.

ISSN: 2233-7857 IJFGCN

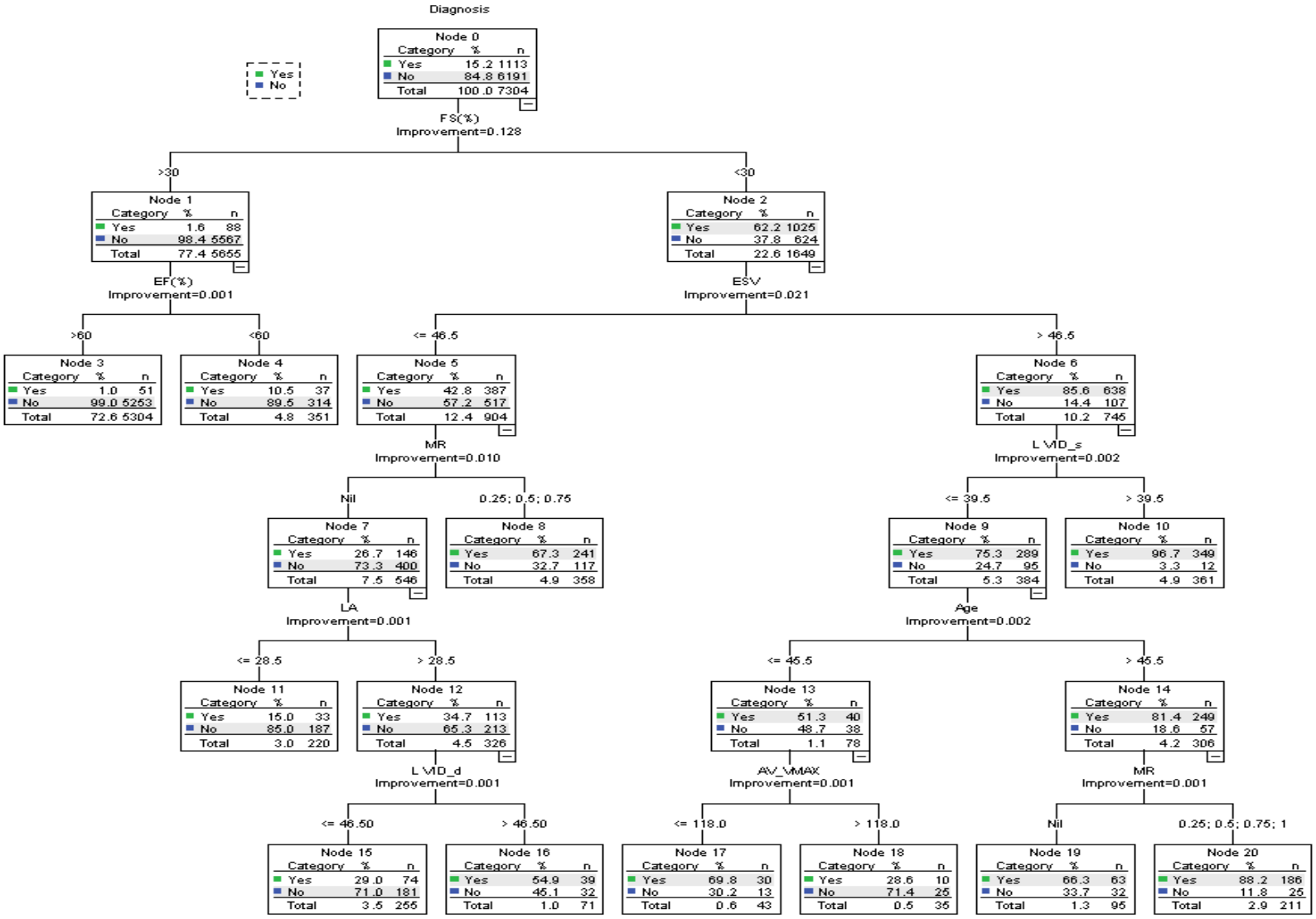

Figure 3: CART tree model for prediction of IHD.

ISSN: 2233-7857 IJFGCN

CART Algorithm

FS (%) = >30: No (5567/98.4%)
| EF (%) = >60: No (5253/99%)
| EF (%) = <60
FS (%) = <30
| ESV <= 46.50
| | MR = Nil
| | | LA <= 28.5: No (187/85%)
| | | LA > 28.5
| | | | LVID_d <= 46.50: No (181/71%)
| | | | LVID_d > 46.50: Yes (39/54.9%)
| | MR = 0.2;0.5;0.75: Yes (241/67.3%)
| ESV > 46.50
| | LVID_s <= 39.5: Yes (289/75.3%)
| | | Age <= 45.5: Yes (40/51.3%)
| | | | AV_VMAX <= 118.0: Yes (30/69%)
| | | | AV_VMAX > 118.0: No (25/71.4%)
| | | Age > 45.5: Yes (249/81.4%)
| | | | MR = Nil: Yes (63/66.3%)
| | | | MR = 0.25;0.5;0.75;1: Yes (186/88.2%)
| | LVID_s > 39.5: Yes (349/96.7%)

4. CONCLUSION

CART analysis can guide medical researchers to isolate which of the variables is most important as a potential site of intervention to make the decision. This result shows that the Ischemic heart diseases of a patient are predicted successfully with an acceptable ratio of 94 %. Furthermore, the resulting model has a high true negative rate and significant rules were extracted from a dataset that makes the application of Decision tree in predicting IHD in healthcare.

REFERENCES

1. Global status report on noncommunicable diseases. 2010.
2. Kalmegh S. Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News. 2015;2(2):438–46.
3. Barros RC, de Carvalho ACPLF, Freitas AA. Automatic Design of Decision-Tree Induction Algorithms [Internet]. Cham: Springer International Publishing; 2015 [cited 2020 Jul 7]. Available from: http://link.springer.com/10.1007/978-3-319-14231-9.
4. Abdul Kareem S, Raviraja S, A Awadh N, Kamaruzaman A, Kajindran A. Classification And Regression Tree In Prediction Of Survival Of AIDS Patients. MJCS. 2010 Dec 1;23(3):153–65.
5. Thenmozhi K, Deepika P. Heart Disease Prediction Using Classification with Different Decision Tree Techniques. 2014;2(6):6–11.
6. Shouman M, Turner T, Stocker R. Using Decision Tree for Diagnosing Heart Disease Patients. 2011;23–9.
7. Reddy RVK, Raju KP, Kumar MJ, Sujatha CH, Prakash PR. Prediction of Heart Disease Using Decision Tree Approach. 2016;6(3):530–2.
8. Lewis RJ, Ph D, Street WC. An Introduction to Classification and Regression Tree (CART) Analysis. 2000;(310).
9. Nahar N, Ara F. L IVER D ISEASE P REDICTION BY U SING D IFFERENT. 2018;8(2):1–9.
10. Vairavan PM. Classification Using Decision Tree Approach towards Information Retrieval Keywords Techniques and a Data Mining Implementation Using WEKA Data Set. 2017.

ISSN: 2233-7857 IJFGCN

11. Ahmad F. Implementing WEKA as a Data Mining Tool to Analyze Students ' Academic Performances Using Naïve Bayes Classifier Implementing WEKA as a Data Mining Tool to Analyze S tudents ' Academic Performances Using Naïve Bayes Classifier Nur Hafieza Ismail ,. 2013. .
12. Suthaharan S. Machine Learning Models and Algorithms for Big Data Classification [Internet]. Boston, MA: Springer US; 2016 [cited 2020 Jul 7]. (Integrated Series in Information Systems; vol. 36). Available from: http://link.springer.com/10.1007/978-1-4899-7641-3

ISSN: 2233-7857 IJFGCN