International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Predicting the Symptoms of Cardio Vascular Disease using Machine Learning Technique

K.Arunmozhi Arasan, E.Ramaraj, S.Muthukumaran

Abstract: ***Heart disease is the top most reason for mortality in India. Cardio Vascular Disease is a type of heart disease and it refers to narrowed or blocked blood artery vessels that affect the flow of blood to and from the heart and rest of the parts of the human body. Globally 31% of the people died due to Cardio Vascular Disease per year. Blood pressure, diabetes, abnormal cholesterol increase, diabetes, hypertension are the risk factors for the Cardio Vascular Disease. This paper used a machine learning approach with C4.5 classifier to predict the reason for the Cardio Vascular Disease. The data used for this research was collected from private hospitals located in different districts of Tamil Nadu. The parameters which cause Cardio Vascular Disease are consulted with doctors and a dataset was formed with 18 parameters. Decision tree was obtained from the dataset using C4.5 algorithm and the reason for the cause of Cardio Vascular Disease was generated as rules from the Decision Tree. This extracted knowledge is used to predict the reason for Cardio Vascular Disease and prevents deaths due to heart disease.***

Keywords: Machine Learning, Cardio Vascular Disease, Decision Tree, C4.5.

I. INTRODUCTION

Disease which affects the functions of the heart such as Coronary Artery which affects the heart arteries, Valvular heart disease that affects the regulation of blood flows in valves, Cardiomyopathy that affects the squeezes of heart muscles, etc., are grouped under a common name called Heart Disease. There are approximately ten types of Cardio vascular Disease (CVD) are identified and labeled till date by the World Health Organization (WHO). WHO statistics report says that 17.9 million people across the world die to CVD that is 31% of the total deaths across the world. According to Indian Council of Medical Research report, in India Cardio Vascular Mortality has increased to an extend of 28.1% of the total deaths in 2016 and it has increased to 50% of the total death in 2018. Research and action has to be taken immediately to achieve sustainable development in preventing Cardio Vascular Mortality. Early prediction of Cardio Vascular Disease helps the physician for giving treatment and reduces the mortality. Large amount of data are generated in healthcare industry every day for treatment process of any particular disease. Data mining techniques are used widely in healthcare industry to process the large amount of data generated in the form images, text etc.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
K.Arunmozhi Arasan, Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, India. Email: arunlucks@yahoo.co.in
E.Ramaraj, Professor and Head, Department of Computer Science, Alagappa Univeristy, Karaikudi, India. Email:eramaraj@rediffmail.com
S.Muthukumaran, Research Scholar, Department of Computer Science, Alagappa University, Karaikudi, India. Email:muthumphil11@gmail.com

In this paper decision tree is used to classify the Cardio Vascular data and the generated rules can be used to identify the reason for the Cardio Vascular Attack.

II. LITERATURE REVIEW

Hashi, E. K., Zaman, et al [1] presented a paper to predict the diabetes disease by comparing C4.5 and K-Nearest Neighbor algorithms. They stated that C4.5 algorithm performs better when compared with KNN by giving better accuracy. Alam, J., Alam, S.Hossan et al [2] presented a paper and in it they proposed an algorithm called Multi-class Support Vector Machine classifier for detecting Lung cancer. Their proposed method gives better accuracy in detecting the probability of lung cancer. Roohallah Alizadehsania et al [3] reviewed various articles published from 1992-2019 that used machine learning algorithm for predicting Cardio Vascular Disease. In their paper they reviewed about the parameters used, sample size of the participants, geographical size stenosis of each coronary artery were studied and the performance metrics used for various machine learning techniques are also briefly analyzed. Bashir, S. et al [4] presented a paper and in it they compared three classification algorithms namely ID3, C4.5, CART for diagnose the diabetes disease. They used the performance metrics such as Bagging, Adaboosting and Majority Voting to improve the performance and accuracy. Joshi, R., Alehegn, M. [5] presented a paper and they proposed a hybrid approach combining KNN, Naïve Bayes, and Random Forest and J48 algorithms for diagnosis of diabetes.

III. PROPOSED FRAMEWORK

This paper uses a framework for the support of medical practitioner to detect the cardiovascular disease. The risk factors for cardiovascular disease such as age, genetic factors, cholesterol, obesity, glucose level, stress, smoke/alcohol/tobacco use and physical activity are identified by consulting with doctors and cardiologists. The details from the people with the age group in between 40 to 65 years are collected for the study. Medical practitioner, community health centers and hospitals are the sources for the data. The following diagram shows the framework used in this paper.

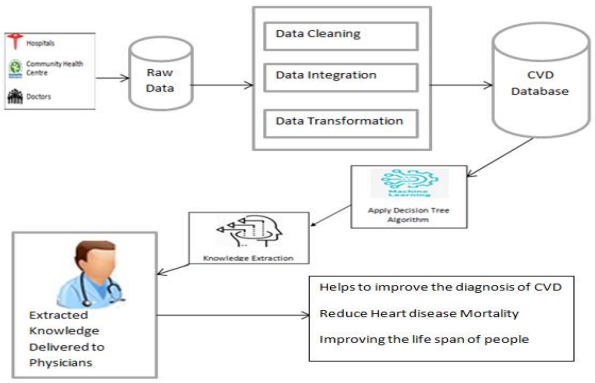

Fig. 1.Proposed Framework

The collected data are preprocessed for cleansing and filling up the missing values and then integrated into single database to form the CVD database. The C4.5 decision tree algorithm is applied on the dataset to extract the knowledge and rules. The generated rules and extracted knowledge are delivered to the medical practitioner to diagnosis of the cardiovascular disease.

A. Feature Extraction for Decision Tree Using C4.5

A Feature is an attribute contained in a Dataset and it is the most influencing attribute selected from the dataset using the selection measure. Information Gain, Gain Ratio, Gini Index are the three different attribute selection measures used by the decision tree induction algorithms. For this research the Information Gain is used as the selection measure and the expected information to classify a tuple in the Dataset is given by Info (D) $=\sum_{i=1}^{m} p_i \log_2 p_i$. The information required to classify a tuple from the dataset on partitioning is calculated by $\text{Info}_A(D)=-\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \text{Info}(D_j)$.

The gain is calculated by using the formula $\text{Gain}(A)=\text{Info}(D)- \text{Info}_A(D)$. The gain ratio is the selection measure to find the splitting attribute and the attribute which has the maximum value is taken as the splitting criterion for construction of a decision tree. The Gain ratio applies normalization on information gain and it is calculated by using the formula,

$$\text{SplitInfo}_A(D)= - \sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$\text{GainRatio}(A)=\frac{Gain(A)}{SplitInfo(A)}$$

IV. RESULTS AND DISCUSSION

The results obtained from Tanagra data mining tool shows that sys_hi (systolic pressure) is taken as the root node because it has the highest information gain value. Hence it is taken as the splitting criterion. R1: If Physical_activity is regular and sys_hi is less than 134 then Cardio is NO. (30) Tuples. R2: If family_genetic is yes and food_prefer is normal_food and physical_activity is no_activity and sys_hi value is less than 134 then cardio is YES. (16) Tuples. R3: If the use of tobacco is greater than or equal to 0.5 and family_genetic is no and food_prefer is normal_food and physical_activity is no_activity and sys_hi is less than 134 then cardio is YES. (5) Tuples. R4: If stress is very_high_stress and age is less than 22019 days and cholesterol is NORMAL_COLES and weight is greater than 62.5 kgs and dia_lo is greater than 74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal_food and physical_activity is no_activity and sys_hi is less than 134 then cardio is YES. (6) Tuples. R5: If stress is stress and age is less than 22019 days and cholesterol is NORMAL_COLES and weight is greater than or equal to 62.5 and dia_lo is greater than74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal-food and physical_activity is no_actvity and sys-hi is less than 134 then cardio is NO (8) Tuples. R6: If age is less than 18778 days and gender is MALE and stress is no_stress and age is less than 22019 days and cholesterol is NORMAL_COLES and weight is greater than or equal to 62.5 and dia_lo is greater than74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal-food and physical_activity is no_actvity and sys-hi is less than 134 then cardio is NO (6) Tuples. Similarly, R7: If age is less than 20452 days and age is > = 18778days and gender is MALE and stress is no_stress and age is less than 22019 days and cholesterol is NORMAL_COLES and weight is > = 62.5 and dia_lo is > = 74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal-food and physical_activity is no_actvity and sys-hi is less than 134 then cardio is NO (7) Tuples..R8: If height is less than 166 and age is greater than or equal to 20452 days and age is >= 18778 days and gender is MALE and stress is no_stress and age is less than 22019 days and cholesterol is NORMAL_COLES and weight is >= 62.5 and dia_lo is greater than or equal to 74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal-food and physical_activity is no_actvity and sys-hi is less than 134 then cardio is NO (7) Tuples. R: 9 If dia_lo is less than 74 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal_food and physical_activity is no_activity and sys_hi is less than 134 then cardio is NO. (55) Tuples. R10: If gender is FEMALE and stress is no_stress and age is less than 22019 and cholesterol is NORMAL_COLES and weight is greater than 62.5 and dia_lo is > = 74.5 and tobacco is less than 0.5 and family_genetic is no and food_prefer is normal_food and physical_activity is no_activity and sys_hi is less than 134 then cardio is No. (23) Tuples.R11: If food_prefer is junk_oil food and physical_activity is no_activity and sys_hi is less than 134 then cardio is YES. (12) Tuples. R12: If physical_activity is irregular and sys_hi is less than 134 then cardio is YES. (16) Tuples. R13: If sys_hi is > = 134 then cardio is NO. (113)Tuples.

The obtained result generates 18 rules for detecting the CV Disease. The attribute sys_hi (systolic blood pressure) is taken as the root node for the generated decision tree. The following figure shows the resultant decision tree obtained from Tanagra tool.

Retrieval Number: K18650981119/19©BEIESP
DOI: 10.35940/ijitee.K1865.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

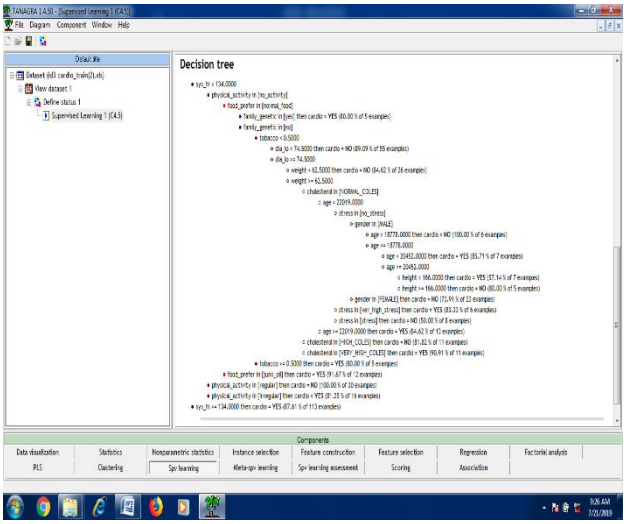

Fig. 2.Decision Tree obtained from Tanagra Tool

A. Performance Measures

The accuracy of the classifier can be measured by the number records that are classified correctly by the classifier. Confusion matrix is a tool to analyze the performance of the classifier. Confusion matrix for the CVD dataset is as follows,

Table I: Confusion Matrix for CVD dataset.

	NO	YES	SUM
NO	141	28	169
YES	23	167	190
SUM	164	195	359

In the CVD dataset cardio is taken as the target attribute and it is in discrete type (YES and NO). Accuracy is a performance metric to measure how long the classifier correctly classifies the given dataset. Error rate is calculated by using the number of tuples that are misclassified by the classifier. The performance of the classifier is measured using the quality metrics such as, accuracy, error rate (misclassification rate), recall, specificity, precision, 1-precision and the results are tabulated as given below.

Table II: Performance Metrics

Measures	C4.5
Accuracy	0.8579
Error Rate	0.1421
Recall	0.8343
Specificity	0.8789
Precision	0.8598
1- Precision	0.1402

The performance measures show that the C4.5 algorithm correctly classifies 86% of tuples from the cardiovascular dataset using Tanagra tool.

V. DISCUSSION

People with systolic blood pressure less than 134 and doing regular physical activity are having fewer chances for the cardiovascular disease. [R1]. If the systolic blood pressure is >=134 then it is the major cause among the other attributes for the CVD [R13] and it covers 113 tuples. The family_genetic is also a reason for the cardiovascular diseases because it may be inherited from the parents (heredity). If the diastolic pressure (dia_lo) is less than 74 and no genetic reason for the cardiovascular diseases then there is no chance for the cardiovascular disease. The system predicts 55 such tuples. Similarly, the use of tobacco is also detected as one of the major cause for the disease if a person with systolic pressure is less than 134 and who consumes the tobacco products. If a person having high cholesterol level and not taking any regular physical activity and the body weight which is more than the average weight are facing the CV Disease. The stress and the hypertension due to some personal, financial, family, social issues raised the possibility of the cardio disease. The results show that systolic pressure and physical activity are the major causes for the cardiovascular disease.

VI. CONCLUSION

Most of the premature deaths are happened due to the cardio diseases, especially when it is not detected during the initial stages. If it is not properly diagnosed the cardiovascular disease can lead to some other diseases such as sudden cardiac arrest, stroke, heart failure, coronary artery disease etc. The system detects the reason for the cardiovascular disease using the C4.5 decision tree algorithm and it can also be used as tool to predict symptoms forth disease and treatments to be given for the disease. This system identifies systolic blood pressure is the major cause for the cardio vascular disease. Also the system suggests doing regular physical activity which may reduce the chances for the CVD. Further studies can be extended to find the specific type of the Cardiovascular Disease and more specific reasons that causes the disease.

ACKNOWLEDGEMENTS

This article has been written with the financial support of RUSA-Phase 2.0 grant sanctioned vide Letter No.F.24-51/2014-U, Policy (TNMulti-Gen), Dep.ofEdn.Govt. Of India, Dt.09.10.2018.

REFERENCES

1. Hashi, E. K., Zaman, M. S. U., &Hasan, M. R., “An expert clinical decision support system to predict disease using classification techniques”. International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 396-400). IEEE, 2017.
2. Alam, J., Alam, S., &Hossan, A. “Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie”. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE, 2018.

3. Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., &Acharya, U. R. "Machine learning-based coronary artery disease diagnosis: A comprehensive review". Computers in biology and medicine, pp.: 103346, 2019.
4. Bashir, S., Qamar, U., Khan, F. H., &Javed, M. Y. "An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles".12th International Conference on Frontiers of Information Technology (pp. 226-231). IEEE, 2014.
5. Joshi, R., &Alehegn, M. "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach". International Research Journal of Engineering and Technology, Vol.4, Issue 10, 2017.

AUTHORS PROFILE

K.Arunmozhi Arasan is a Research Scholar in the Department of Computer Science, Alagappa University, Karaikudi. His main Research Interest includes Data Mining and Big Data Analytics. He has published 6 International Journals.

E. Ramaraj is working as the Professor and Head of the Department of Computer Science, Alagappa University, Karaikudi. He has the sound knowledge in many research fields especially in Data Mining, Network Security, Remote Sensing, and Big Data & Analytics. He has published more than 100 international journals.

S.Muthukumaran is a Research Scholar in the Department of Computer Science, Alagappa University, Karaikudi. His main Research Interest includes Data Mining, Big Data Analytics, and Internet of Things. He has published 9 International Journals.