INDO ASIAN JOURNAL OF INFORMATION RESEARCH AND TECHNOLOGY

VOL 1 (01), AUG 2024, PP. 26-33

AN APPROACH FOR MODELLING AND PREDICTING CARDIAC DISEASE USING CLASSIFICATION AND REGRESSION TREE (CART)

SIDHARTH GUPTA
PGDM (RESEARCH AND BUSINESS ANALYTICS), LAL BAHADUR SHASTRI COLLEGE
SIDHARTHGUPTA-E23@LBSIM.AC.IN

ABSTRACT

Heart-related illnesses are estimated to be the primary cause of about one-third of all fatalities globally, making heart disease the leading cause of death. Early detection is made possible by expanding the use of classification-based machine learning in medicine. The Classification and Regression Tree (CART) in this investigation supervised machine learning technique called algorithm has been used to forecast heart disease and extract guidelines for making decisions on how to relate input and output variables. Furthermore, the study's conclusions sort the characteristics that affect heart disease according to significance. Taking into account every performance criterion, the model's 87% prediction accuracy confirms its dependability. Conversely, extricated decision rules shown in the study can make using it for healthcare purposes easier without requiring further understanding. Overall, patients who are experiencing financial or time restrictions during the diagnosis and treatment of heart disease can benefit from the suggested strategy in addition to medical specialists.

Keywords: Machine Learning, classification, CART

1. INTRODUCTION

The field of machine learning is becoming more and more crucial for illness early detection. It focusses on finding patterns hidden in observations to draw conclusions consistent with new knowledge (Chivakula and Dao, 2022). Heart-related diseases, for instance, are among the worst illnesses globally, claiming the lives of almost 17.9 million people year(World Health Organisation, 2013). Many researchers from many backgrounds have been studying the problem using this method since early detection of cardiac disease may lower the death rate. Determining a patient's likelihood of receiving a heart disease diagnosis has gained attention again because to recent advancements in the field of machine learning(Chang et al, 2022). Decision trees are frequently employed as one of the categorisation strategies in machine learning for early identification of heart disease (Ahsan and Siddique, 2022).

Decision trees are superior to other classification algorithms in clinical contexts because they are easy to grasp by decision makers and can be used to classify data that has not yet been seen(Alanazi, 2022). Higher variations of decision trees are successfully used in this domain the more new data are generated [8,9]. There has been very little study on modelling and extracting rule sets with graphical representations, which may simplify decision-making, despite the existence of studies in several application areas that primarily focus on prediction by employing decision trees; notably, on the heart disease sub-domain.

In an effort to compensate for the aforementioned limitations, this study models and predicts cardiac disease. A large data collection comprising five data sets with 1190 observations and eleven features was employed in this investigation. First, the data were pre-processed in order to run the model more effectively and precisely. Second, the Classification and Regression Tree (CART) method was presented and investigated as a means of accurately identifying cardiac illness and presenting a graphical depiction of the model's rule inferences. Unlike most studies, this one's findings and implications could be easily integrated into clinical decision support systems in the future because of the provided rule sets and modelling outcomes.

2. MATERIALS AND METHODS

The literature has examined a wide range of machine learning methods, mostly comprising supervised and unsupervised techniques, for the prediction of heart disease. Some data sets are said to perform better in the early diagnosis of heart disease as machine learning is used in medicine more and more. Among the seventy data sets that are published and contain cases related to heart disease within three decades, Cleveland, Hungarian, Switzerland, and Statlog are the most well-known and frequently mentioned (Alizadehsaniet al, 2021). All of the aforementioned data sets have both numerical and categorical data among their attributes. These data sets contain non-medical characteristics like the age and gender of the patients, medical characteristics like blood pressure, blood sugar, cholesterol, ECG readings, maximal heart rate, etc

The majority of our knowledge regarding the use of machine learning to predict cardiac disease comes from empirical research that looks into the algorithm's performance. To compare output performances, M.M. Ali et al. used supervised machine learning techniques such as Random Forest, Decision Tree, Adaboost Classifier, K-Nearest Neighbour (KNN), Multilayer Perceptron, and Logistic Regression(Ali et al, 2021). A heart disease application that continually tracks patients with coronary heart disease and allows them to view their state was introduced by (Repaka et al, 2019). The Naïve Bayes algorithm served as the foundation for the development of this software. suggested a hybrid system made from supervised machine learning methods after combining a genetic algorithm with recursive feature reduction for feature selection(Rani et al, 2021).They looked at a few classification techniques using the Framingham and CVD data sets. Therefore, they proposed that optimising feature size could lead to a reduction in processing time and an improvement in model performance.

Decision trees, a component of widely used machine learning algorithms, provide insight into handling target variable prediction based on supplied input parameters(Sann et al, 2022). Depending on the state of the variables, decision trees can be seen from two different angles. Regression can be used to address the problem when the target variable is numerical; if not, classification is the appropriate approach [36].

Despite the fact that the literature has a large number of decision tree methods for classification problems. Among the most often cited decision tree algorithms are Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detection (CHAID) algorithms(Ghiasi and Zendehboudi, 2019). According to related studies, the CART algorithm demonstrates its dependability over all other categorisation methods(Batra and Agrawal, 2019). The CART method, like other decision tree algorithms, constructs a model to predict the target values by extracting decision rules from characteristics(Ghiasi and Zendehboudi, 2019). Both qualitative and numerical data may be included in the characteristics.
In general, the CART algorithm was implemented to guarantee that the outcomes of researchers and medical professionals might benefit from the study in the early diagnosis and strategies for care.

Five well-known heart disease data sets totalling both medical and non-medical variables were used to curate the data set for this study. The data set includes both categorical and numerical variables.
https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/input. 1190 patient records in all were included in this extensive data set. Eleven features make up the data set: maximal heart rate, exercise-induced angina, sex, age, kind of chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, oldpeak,

and ST slope. It also contains a single target variable that might be "normal" or "heart disease."

Table I Variables

Variable	Data Type
Age	int64
Sex	object
Chest_Pain_Type	object
Resting_Blood_Pressure	int64
Cholesterol	int64
Fasting_Blood_Sugar	object
Resting_Electrocardiogram	object
Max_Heart_Rate_Achieved	int64
Exercise_Induced_Angina	object
St_Depression	float64
St_Slope	object
Num_Major_Vessels	int64
Thalassemia	object

The dictation of data is as follows

1. age: age in years
2. sex: sex
    - 1 = male
    - 0 = female
3. cp: chest pain type
    - Value 0: typical angina
    - Value 1: atypical angina
    - Value 2: non-anginal pain
    - Value 3: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl)
    - 1 = true;
    - 0 = false
7. restecg: resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach: maximum heart rate achieved
9. exang: exercise induced angina
    - 1 = yes
    - 0 = no
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    - Value 0: upsloping
    - Value 1: flat
    - Value 2: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal:
    - 0 = error (in the original dataset 0 maps to NaN's)
    - 1 = fixed defect
    - 2 = normal
    - 3 = reversable defect
14. target (the lable):
    - 0 = no disease,
    - 1 = disease

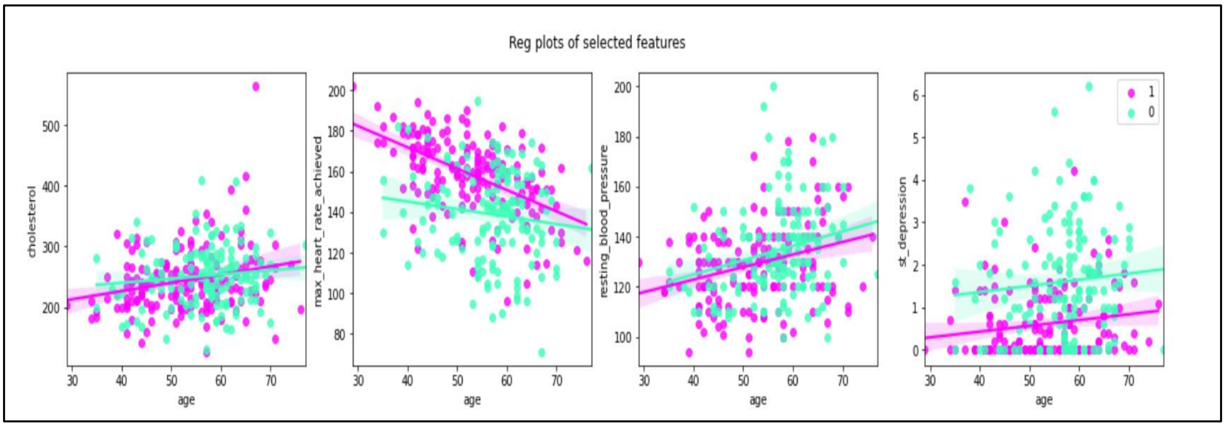
3. DISCUSSION

Table II Descriptive Statistics

Variables	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
age	296.0	54.523649	9.059471	29.0	48.0	56.0	61.00	77.0
cholesterol	296.0	247.155405	51.977011	126.0	211.0	242.5	275.25	564.0
resting_blood_pressure	296.0	131.604730	17.726620	94.0	120.0	130.0	140.00	200.0
max_heart_rate_achieved	296.0	149.560811	22.970792	71.0	133.0	152.5	166.00	202.0
st_depression	296.0	1.059122	1.166474	0.0	0.0	0.8	1.65	6.2
num_major_vessels	296.0	0.679054	0.939726	0.0	0.0	0.0	1.00	3.0

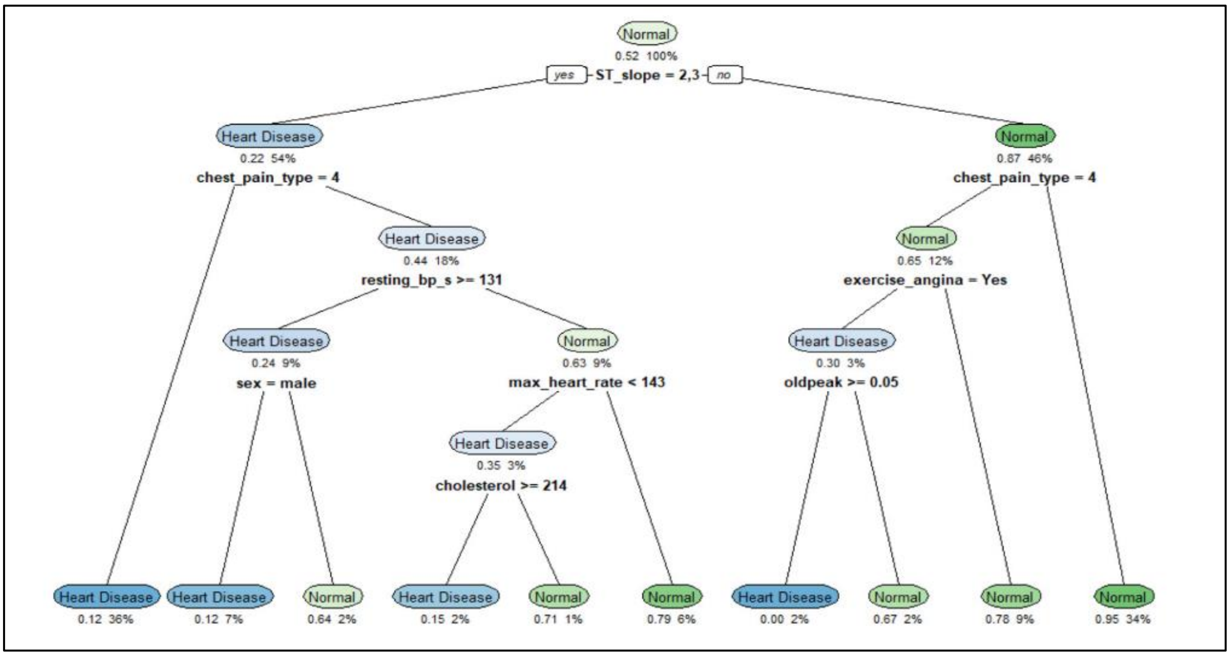
Source :Author Calculation

Figure I Reg Plot



Source :Author Calculation

Figure II CART Diagram



Source :Author Calculation

According to the results, the ST slope is flat or downsloping in about 96% of the patients who are expected to have heart disease. Furthermore, the type of asymptomatic chest pain is a crucial signal; in fact, 81% of patients with heart disease experience this form of asymptomatic chest pain. These results also suggest that if ST slope values and the type of chest discomfort are not examined, early detection of heart disease may be called into question. However, cholesterol can be viewed as an input to raise the prediction's overall accuracy, even if it is one of the features with little influence on the model.

4. CONCLUSION

One of the most deadly illnesses in the world today is heart disease. It is extremely important, particularly for patients who run the risk of having a heart attack or even dying, as well as for the medical personnel in charge of making the diagnosis as soon as is practical. It is crucial to identify the patients in advance in order to effectively address these problems.
In this way, the CART model is proposed in this study to provide healthcare providers with a more thorough and rational understanding of their patients. It also provides a logical set of rules for using the model in real-world scenarios without undue complexity. In this instance, 1190 patients' electronic medical records were used to build the model. The model was run following the preprocessing and analysis of the original data set. The model used in this investigation turned out to be validated and precise enough. A noteworthy discovery from this research is that our model has a respectable predictive ability for heart disease. The second noteworthy discovery is that features related to ECG tests, such as ST Slope and Oldpeak, are more relevant to detect heart illness based on the CART model's value of features.

REFERENCES

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, *30*, 100924.

Alizadehsani, R., Khosravi, A., Roshanzamir, M., Abdar, M., Sarrafzadegan, N., Shafie, D., ... & Acharya, U. R. (2021). Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020. *Computers in Biology and Medicine*, *128*, 104095.

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.

Batra, M., & Agrawal, R. (2018). Comparative analysis of decision tree algorithms. In *Nature Inspired Computing: Proceedings of CSI 2015* (pp. 31-36). Springer Singapore.

Chivakula, M., & Dao, K. (2022). The Relationship Between Classical Music Therapy and Heart Disease: A Systematic Review and Meta-Analysis. *Journal of Student Research*, *11*(2).

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016.

Ghiasi, M. M., & Zendehboudi, S. (2019). Decision tree-based methodology to select a proper approach for wart treatment. *Computers in biology and medicine*, *108*, 400-409.

Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263-275.

Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

Sann, R., Lai, P. C., Liaw, S. Y., & Chen, C. T. (2022). Predicting online complaining behavior in the hospitality industry: Application of big data analytics to online reviews. *Sustainability*, *14*(3), 1800.

World Health Organization. (2013). Health topics: Cardiovascular diseases. *Fact Sheet. Available online: http://www. who. int/cardiovascular_diseases/en/(accessed on 11 December 2020).*