MULTI-ALGORITHMIC MODEL FOR PREDICTING CARDIOVASCULAR DISEASE :A MACHINE LEARNING APPROACH

Sushma S
Department of CSE
RNSIT, Bengaluru, India
1rn21cs169.sushmas@rnsit.ac.in

Thrishala S
Department of CSE
RNSIT, Bengaluru, India
1rn21cs174.thrishalas@rnsit.ac.in

Mrs.Anupama
Asst. Professor, Department of CSE
RNSIT, Bengaluru, India
anupamachinmay@gmail.com

***Abstract*—Cardiovascular diseases (CVDs) are a major cause of global mortality, underscoring the necessity of precise predictive models for the earlier detection and intervention. This paper presents a multi-algorithmic strategy utilizing machine learning to assess CVD risk. Our method combines several algorithms, including K-NN, Logistic Regression, Naive Bayes, and Decision Trees, to harness their individual strengths and address their weaknesses. We use a comprehensive dataset that covers lifestyle, clinical, and demographic variables to train and evaluate our model, ensuring robust performance across diverse patient groups. Our extensive experiments and validations shows that our multi-algorithmic model outperforms individual algorithms in concern with predictive accuracy. Additionally, we offer perpectives on the relative importance of various risk factors in CVD prediction, aiding physicians in targeted risk evaluation and personalized intervention strategies. Our findings underscore the effectiveness of a multi-algorithmic approach in enhancing accuracy and consistency of cardio disease prediction, leading to more proactive healthcare management and improved patient outcomes.**

***Index Terms*—CVD(cardiovascular Disease), Precision, Supervised learning, Classification, Regression,**

I. INTRODUCTION

Cardiovascular disease is a major health challenge globally, and early identification is crucial for effective treatment and prevention of complications. CVDs are the world's largest cause of death, with an estimated 18 million fatalities each year, or roughly 32% of all deaths worldwide. The World Health Organization(WHO) estimates that low- to middle-income nations account for 75 percent of deaths from CVD. These figures highlight the critical necessity of cardio disease prevention and early identification, which can dramatically lower death rates.

People in low- and middle-income nations frequently lack access to primary healthcare programs that would enable early identification and treatment of those who have cardiovascular disease risk factors. Individuals with CVDs who live in low- and middle-income nations have reduced access to quality, affordable healthcare that meets their needs. Because of this, many people in these nations experience late-stage disease identification, which causes them to pass away from CVDs and other diseases at a younger age—often during their prime working years.

Recently, it has become a notable rise in interest in predictive developing methods for early recognition of cardio disease, fueled by the availability of health data and advancements in machine learning. This trend has resulted in a growing preference for machine learning techniques to improve the prediction precision.

Notable algorithms in this area include Logistic Regression, K-Nearest Neighbor (KNN), decision trees (CART), and naive Bayes. Logistic regression is frequently employed for binary classification tasks, estimating the likelihood of cardio disease. KNN relies on proximity for classification, offering flexibility. Decision trees, such as CART, provide interpret-ability based on input features. The redundancy in logistic regression aside, the inclusion of naive Bayes highlights the exploration of probabilistic models, known for simplicity and efficiency in high-dimensional datasets.

To address existing limitations, ongoing research seeks to improvise machine learning algorithms for cardio disease prediction. One approach is to integrate additional data sources, like genetic and environmental information, to boost predictive accuracy. Another focus is on creating interpretable methods that will uncover the underlying mechanisms of cardio disease. Despite promising results in means of cost-effectiveness and practical use, significant limitations remain. These models often fail to consider critical risk factors such as family history, lifestyle choices, and medication history. While they can be valuable for estimating heart disease risk, they should not be solely relied upon for diagnosis or treatment decisions.

Plans to assess 4 widely used machine learning algorithms for predicting heart disease. We will evaluate their predictive accuracy, computational efficiency, and robustness using a publicly available dataset and a range of evaluation metrics. The findings will offer important understanding of the performance of these methods, helping healthcare professionals choose the most effective models for risk assessment and personalized treatment. Ultimately, the research seeks to advance predictive analytics in cardiovascular medicine, improving patient outcomes in heart disease management.

II. LITERATURE REVIEW

Heart disease poses a major global health challenge, making early detection and intervention crucial for better patient outcomes. Lately, there has been a growing interest in using machine learning algorithms to enhance heart disease prediction, potentially advancing risk assessment and personalized healthcare. This literature review offers a comprehensive summary of machine learning methods for predicting heart disease. By systematically analyzing the strengths, limitations, and future prospects of these algorithms, the review highlights the evolving field of predictive analytics in cardiovascular medicine. It intends to guide future research efforts to improve patient outcomes.

Md. Julker Nayeem, Sohel Rana, Md. Rabiul Islam[12] found that Random Forest model has the best accuracy (95.63%) relative to other methods such as K-NN(87.36%) and Navie Bayes(88,89%).

Saladi Novya Sree 1, K.Balaji Pranav Reddy 2, Bangarulakshmi Mahanthi3, Dr. S S Nandini4[14] found out that the Random Forest model got the highest accuracy (82%) relative to other predictive models like K-NN(74%),SVM(52%),Neural Network(76%),Logistic Rgression(65%) and Decision Tree(75%).

Prasanna M, Shrijith Shetty P, and Mamatha K found that K-NN obtained the best accuracy at 88% compare to other predictive models, such as SVC at 84% and Decision Tree at 78%.

S.Rajathi Dr.G.Radhaman[8] found that the K-NN with ACO(70.26%) had highest accuracy in comparison with other predictive methods which are K-NN(68.05%), SVM(65.12%) and Deicison Support(60.05%).

I Ketut Agung Enriko, [1] found that the K-NN with ACO(70.26%) had highest accuracy in compared to other predictive models like K-NN(68.05%), SVM(65.12%) and Decision Support(60.05%).

III. PROBLEM STATEMENT

Heart disease poses a major global health challenge, and early detection is crucial for effective treatment and prevention. Machine learning approaches have exhibited promise in predicting cardio disease upon various patient attributes. However, there is a lack of comprehensive studies comparing the efficacy of distinct machine learning algorithms for this task. This study aims to perform a comparative analysis of four popular machine learning algorithms—K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Classification and Regression Trees (CART)—to evaluate their prodiciency in predicting heart disease.

IV. METHODOLOGIES

A. *Machine Learning Techniques*

1.K-nearest neighbors(KNN):

K-Nearest Neighbors (KNN) is a supervised machine learning technique used for both classification and regression. It determines the class or prediction for a data point based on its proximity to neighboring points. As a non-parametric method, KNN does not presume any specific data distribution. It is also known as Lazy Learning because it does not build a model during training but rather stores the data source and performs calculations at the instance of classification.

Algorithm:

Input: Training Dataset T, Test Instance t, Number of nearest neighbors k. Output: Predicted class

Steps:

a. For each instance i in T, calculate Euclidean distance between the two given points (x_1, y_1) and (x_2, y_2) in a two-dimensional space is given by:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

b. Sort the distance in ascending order ans select the first k nearest training data instances to the test instance.

c.Determine the class of the instance by using majority voting or by calculating the mean.

2. Logistic Regression:

Refers to a supervised learning model aimed at predicting a discrete target variable based on several independent variables. This approach will produce probabilities and classify new data from both continuous and discrete datasets.

Formula:

The Logistic Regression model equation is given by:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

where:

- $P(Y = 1|X)$ is likelihood of the event Y occurring given the input variables X.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the co-efficients of the logistic regression model.
- $X_1, X_2, \ldots, X_n$ are the input variables.
- e represents the base of natural logarithm.

3. Naive Bayes:

It is the supervised learning technique based on Bayes' Theorem, used for addressing classification problems. It is particularly weel-suited for text classification tasks involving high-dimensional datasets.

Algorithm:

a. Compute the prior probability for target class.

b. Compute frequency matrix and likelihood probability for each feature.
c. Calculate the probability of all hypothesis using Bayes Theorem.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

where:

- $P(A|B)$ is the likelihood of event A occurring given that event B has occurred.
- $P(B|A)$ is the likelihood of an event B occurring given that event A has occurred.
- $P(A)$ is the likelihood of event A occurring.
- $P(B)$ is the likelihood of event B occurring.

d. Use map hypothesis h-map to classify the test object to the hypothesis with yielding the best probability. The Naive Bayes classifier calculates the likelihood of a class C_k given the input data points $x_1, x_2, \ldots, x_n$ using Bayes' theorem:

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k) \cdot P(x_1|C_k) \cdots P(x_n|C_k)}{P(x_1) \cdot P(x_2) \cdots P(x_n)}$$

where:

- $P(C_k|x_1, x_2, \ldots, x_n)$ is the likelihood of class $C_k$for the given input features.
- $P(C_k)$ is the prior probability of class C_k.
- $P(x_i|C_k)$ is the likelihood of feature given certain conditions x_i given class C_k.
- $P(x_i)$ is the marginal probability of feature x_i.

4. Decision Tree-CART(Classification and Regression Tree)

Refers to a predictive technique used in machine learning that forecasts the figures of the target variable. It operates as a decision tree, where every node represents a split the dataset according to the predictor variable.

Algorithm:

a. Compute Gini-index for whole training dataset based on target attribute.
b. Compute Gini-index for each of attribute and subset of each attribute.
c. Choose Best splitting subset which has min Gini-index for an attribute.
d. Compute the best splitting feature that possesses max ΔGini.
e. Continue to apply the same operation recursively to the subset until a terminal node is reached. The CART technique uses Gini impurity as the splitting criterion for classification trees and variance reduction for regression trees. For a binary split at node m, the Gini impurity can be calculated as:

$$\text{Gini}(m) = \sum_{k=1} p_{mk}(1 - p_{mk})$$

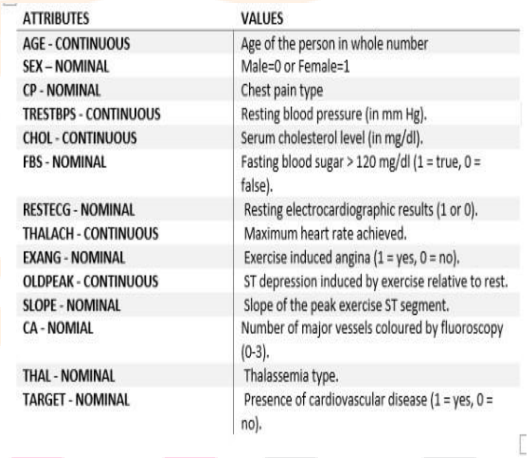
where:

- $\text{Gini}(m)$ is the Gini impurity at node m.
- K refers to number of distinct classes.
- p_{mk} is the amount of samples of class k in node m.

V. KEY COMPONENTS

A. Data Set:

Our dataset comprises 14 attributes: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, thalach, exercise induced angina, oldpeak, slope, ca, thalassemia, and target. It includes 1026 patient records related to cardiovascular disease, which were processed automatically to capture various diagnostic features. Each record was independently classified by two expert cardiologists, leading to a consensus label. The classification system incorporated in this study categorized patient conditions into two classes: NORMAL (0) and SUSPECT (1). Our focus was on two-class classification for predicting cardiovascular disease using this dataset.



ATTRIBUTES	VALUES
AGE - CONTINUOUS	Age of the person in whole number
SEX – NOMINAL	Male=0 or Female=1
CP - NOMINAL	Chest pain type
TRESTBPS - CONTINUOUS	Resting blood pressure (in mm Hg).
CHOL - CONTINUOUS	Serum cholesterol level (in mg/dl).
FBS - NOMINAL	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
RESTECG - NOMINAL	Resting electrocardiographic results (1 or 0).
THALACH - CONTINUOUS	Maximum heart rate achieved.
EXANG - NOMINAL	Exercise induced angina (1 = yes, 0 = no).
OLDPEAK - CONTINUOUS	ST depression induced by exercise relative to rest.
SLOPE - NOMINAL	Slope of the peak exercise ST segment.
CA - NOMIAL	Number of major vessels coloured by fluoroscopy (0-3).
THAL - NOMINAL	Thalassemia type.
TARGET - NOMINAL	Presence of cardiovascular disease (1 = yes, 0 = no).

Fig. 1. Attributes related to dataset

B. Correlation Matrix:

Correlation defines the statistical relationship or association between two or more variables. It quantifies the extent to which changes in one variable are associated with changes in another variable. In essence, correlation gauges how closely variables move in tandem or vary together. To visualize these relationships between variables, a Heatmap is used. It graphically displays the correlation matrix, simplifying the identification of patterns and connections among variables.

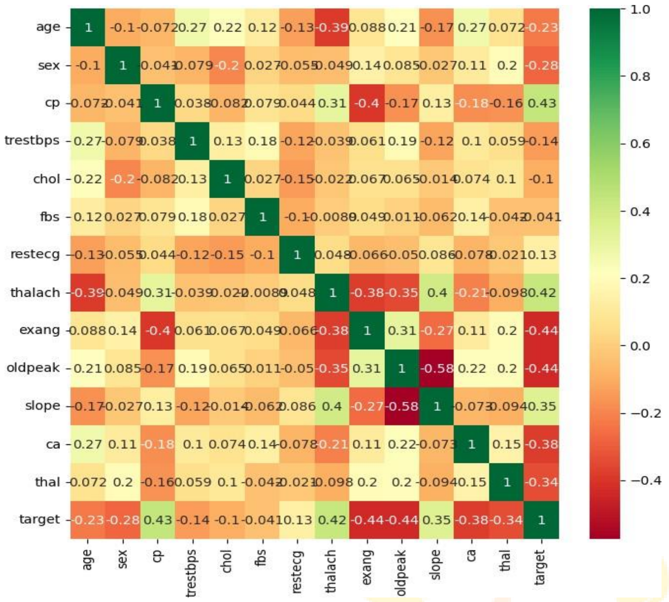


Fig. 2. Correlation between variables

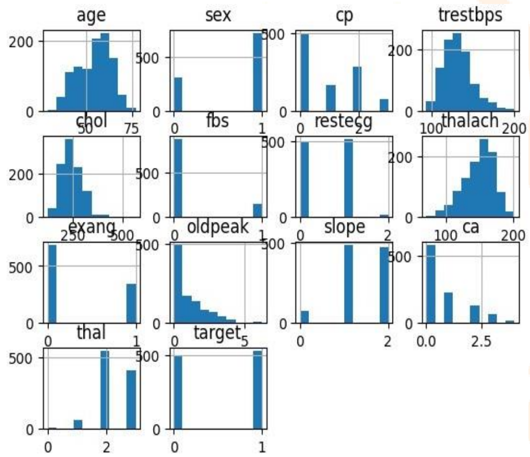


Fig. 3. Histograms of all attributes

C. Histograms of all attributes:

Histograms help visualize how data is distributed across different intervals or bins. By analyzing the shape of a histogram, one can infer the central tendency of the data, including measures like the mean, median, or mode. The width of the histogram reflects the data's spread or variability—wider histograms suggest greater variability, while narrower ones indicate less variability. Additionally, histograms offer insights into the skewness, or asymmetry, regarding the data distribution, and the kurtosis, which describes how peaked or flat the distribution is.

VI. EXPERIMENTATION

Data pre-processing stands as a fundamental preliminary phase in data mining, converting raw data into precise and practical information. It encompasses the resolution of missing values, outlier detection, and normalization of inconsistent data. Missing values are rectified through suitable methods, outliers are addressed using specialized techniques, and inconsistent data is normalized to ensure uniformity. Result is refined, standardized data primed for utilization in training and testing predictive models. Traditionally, dataset is divided into training and testing subsets, with this research allocating 80% for the training purpose and 20% for testing purposes.

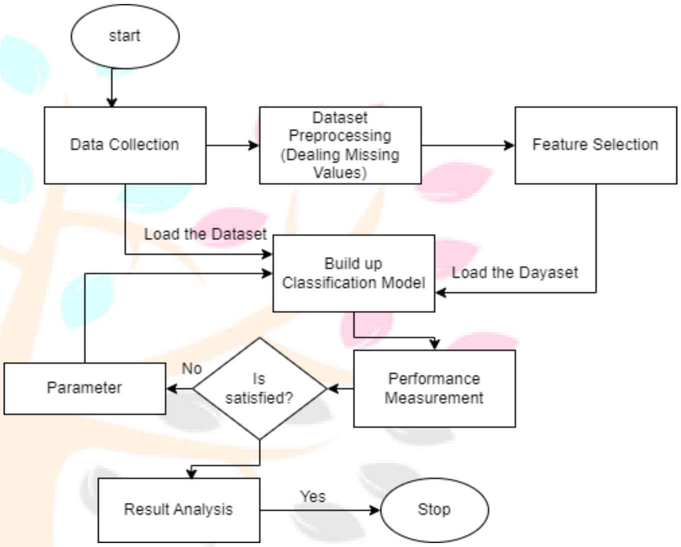
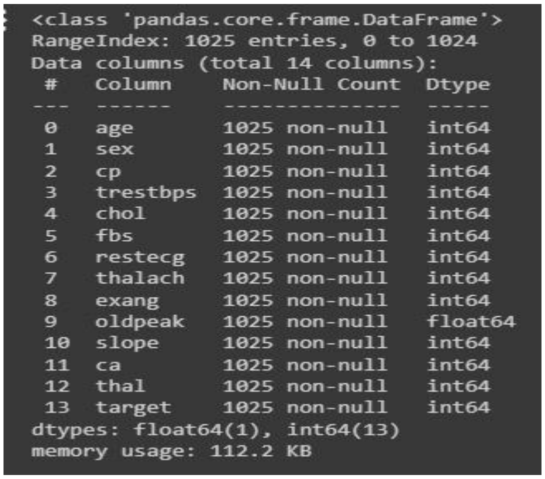


Fig. 4. Flow chart of the work process

The model framework is divided into two primary phases: data preprocessing and predictive modeling. Data preprocessing is vital for converting raw data into the format suitable for analysis. This essential step in data mining involves refining and organizing data to ensure it is standardized for further processing. During preprocessing, three main issues are addressed: handling missing values, managing outliers, and correcting inconsistencies. Missing values, which will result from errors or incomplete data collection, are addressed using appropriate techniques. Outliers, which include erroneous or irrelevant data, are managed through methods designed to detect and address anomalies. Inconsistent data, which can stem from varying formats or errors, is standardized to ensure accuracy and consistency for future analyses. This comprehensive preprocessing ensures that the data is well-prepared for effective and reliable predictive modeling.

Model Training:In this stage, a training dataset with labeled features and targets is created. Various methods, like Logistic Regression, Decision Trees, Naive Bayes, and K-Nearest Neighbors (KNN), are applied. These algorithms are tested empirically on diverse datasets with numerical and categorical



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Fig. 5. Data preprocessing

data. They are effective for classification tasks and are capable of efficiently managing large datasets.

Model Testing: In this phase, a testing dataset, which has the same features as the training dataset but without target labels, is used. The trained model then forecasts the target labels for the testing dataset drawn from patterns and characteristics it has learned.

VII. RESULTS

The system output will provide a prediction regarding whether the individual has heart disease, indicated by a ”Yes” or ”No” outcome. This prediction offers insight into the individual’s heart condition, potentially indicating the likelihood of CVD in advance. If the individual is deemed susceptible to cardio disease, the result will indicate ”Yes,” and conversely, if not, the result will indicate ”No.”

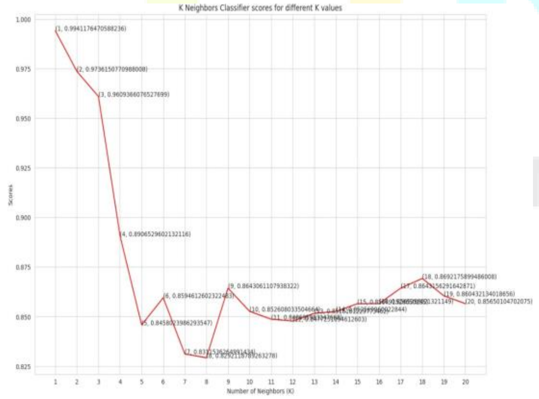


Fig. 6. Graph of K-NN

In fig.6, algorithm achieved the correctness of 84.7%, with precision of 72.55%, recall of 73.27%, and F1-Score of 72.91%. Effective in differentiating between normal and pathological instances, with balanced precision and retrieval rate.

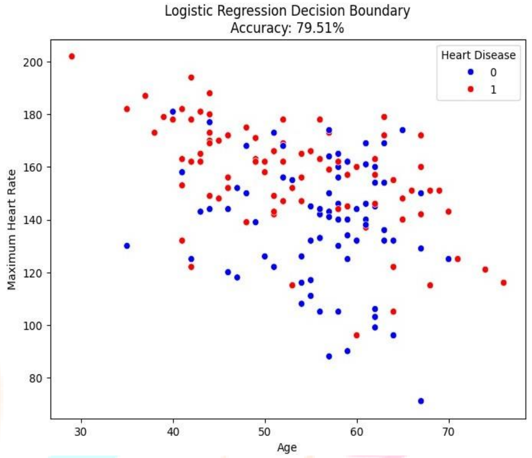


Fig. 7. Logistic Regression graph

In fig.7, the algorithm achieved the correctness of 79.5%. Precision of 71.57%, recall of 84.88%, and F1-Score of 77.66%. Demonstrates slightly lower accuracy compared to KNN, but with relatively higher recall.

In Naive Bayes, this achieved with an accuracy of 80.0%. Precision of 70.59%, recall of 86.74%, and F1-Score of 77.84%. Similar performance to Logistic Regression, with high recall and balanced precision.

In fig.8, the algorithm achieved an accuracy of 98.5%. This performance surpasses other algorithms, highlighting both flawless precision and high recall.

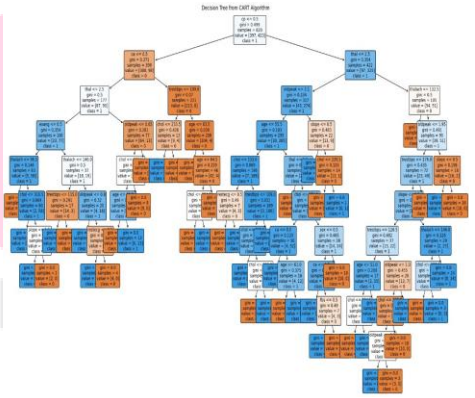
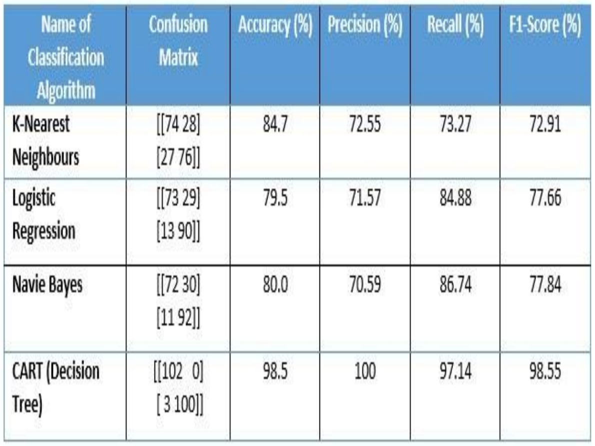


Fig. 8. Graph of Decision Tree

The CART (Decision Tree) algorithm achieved higher performance relative to other algorithms, demonstrating the



Name of Classification Algorithm	Confusion Matrix	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbours	[[74 28] [27 76]]	84.7	72.55	73.27	72.91
Logistic Regression	[[73 29] [13 90]]	79.5	71.57	84.88	77.66
Navie Bayes	[[72 30] [11 92]]	80.0	70.59	86.74	77.84
CART (Decision Tree)	[[102 0] [3 100]]	98.5	100	97.14	98.55

Fig. 9. Performance Table

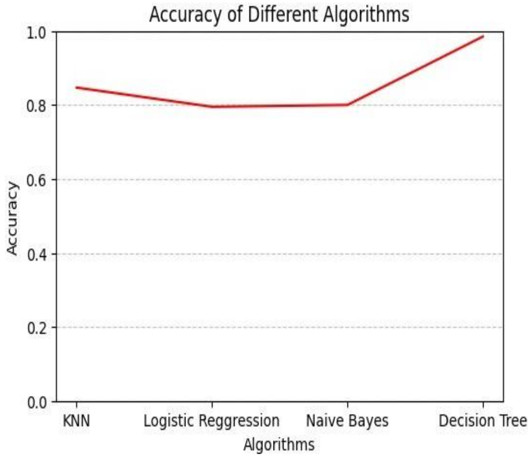


Fig. 10. Accuracy Graph

good accuracy and flawless precision.

Accuracy evaluates a model's performance by determining The proportion of accurate predictions to the overall number of cases

$$\text{Accuracy} = \frac{TP + TN}{FP + TP + TN + FN}$$

Precision evaluates accuracy of the model's positive predictions. It represents the fraction of True Positive predictions to the total Positive predictions made according to model.

$$\text{Precision} = \frac{TP}{FP + TP}$$

Recall assesses how effectively a classification model identifies all relevant instances in a dataset. Given by the ratio of True Positive (TP) instances to the sum of True Positive (TP) and False Negative (FN) instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score assesses overall efficacy of the classification model by calculating unified average of the both precision and the recall.

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In summary, our analysis of different classification procedures for anticipating cardio disease provides key points into their performance. K-Nearest Neighbors, Logistic Regression, and Naive Bayes each show competitive accuracy with optimal equilibrium of precision and recall. However, the CART (Decision Tree) algorithm stands out as the best performer, achieving highest accuracy and perfect precision. This highlights the need to choose a classification algorithm based on specific needs and priorities. The superior performance of CART suggests it could be a valuable tool for cardio disease prognosis, assisting clinicians in risk assessment and personalized treatment. Additional research and validation across various datasets are necessary to confirm these results and improve the reliability and applicability of predictive models in cardiovascular medicine.

VIII. CONCLUSION AND FUTURE WORK

In conclusion, our study assessed the impact of different classification methods for predicting cardio disease. Our evaluation revealed that the CART (Decision Tree) algorithm stands out as foremost effective, achieving both the highest accuracy and perfect precision. Although K-Nearest Neighbors, Logistic Regression, and Naive Bayes also showed strong performance, the CART's results highlight its potential to support clinicians in accurate risk assessment and tailored patient interventions. These results provide significant insights for cardiovascular medicine, highlighting the relevance of advanced machine learning techniques in enhancing predictive analytics and patient outcomes.

Future studies could explore ensemble methods, like Gradient Boosting, to potentially boost predictive reliability and robustness. Additionally, incorporating additional features or biomarkers, such as genetic information or advanced imaging data, may enhance the predictive power of models.

IX. REFERENCES

[1]. Implementation of Na¨ıve Bayes Classification Method for Predicting Purchase Fitriana Harahap1, Ahir

Yugo Nugroho Harahap2, Evri Ekadiansyah3, Rita Novita Sari4, Robiatul Adawiyah5, Charles Bronson Harahap6 Universitas Potensi Utama JL. K.L. Yos Sudarso Km. 6,5 No 3 A-Medan, 20241, Indonesia Email: fitriana@potensi-utama.ac.id1, ahiryugo.potensi@gmail.com2, evri@potensi-utama.ac.id3rita@potensi-utama.ac.id4, robiatul@potensi-utama.ac.id5, charlesharahap07@gmail.com6.

[2].Designing And Implementing Heart Disease Prediction with Naives Bayesian Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)

[3].Heart Disease Prediction System using a hidden Na¨ıve Bayes Classifier M.A.Jabbar Shirina samreen Professor, Professor Vardhaman College of Engineering, Anurag Group of Institutions Hyderabad, Telangana, INDIA Hyderabad,Telangana email:jabbar.meerja@gmail.com shirina.samreen@gmail.com.

[4].Web Analytics Support System for Prediction of Cardio Disease Using Na¨ıve Bayes Weighted Approach (NBwa)Priyanga Department of CSE K.S. Institute of Technology Bengaluru, Karnataka, India p.priyanga@gmail.com Dr. Naveen Department of CSE JSS Academy of Technical Education Bengaluru, India ncnaveen@gmail.com

[5].Predicting Multiple Heart Diseases with Logistic Regression enhanced by Ensemble methods and Hyperparameter tuning Techniques 978-1-7281-6823-4/20//c 2020 IEEE.

[6].Estimation of Heart Disease Risk Using a Logistic Regression Machine Learning Model By: Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal Department of Computer Science and Engineering IIIT Nagpur, India montu.saw@cse.iiitn.ac.in, rahul.yadav@cse.iiitn.ac.in, nidhi.lal@cse.iiitn.ac.in

[7].Human Heart Disease Prediction System using Data Mining Techniques Theresa Princy. R Research Scholar, Department of Information Technology Christ University, Faculty of Engineering, Bangalore, India-560060. princy.aida@gmail.com J. Thomas Department of Computer Science and Engineering Christ University, Faculty of Engineering, Bangalore, India-560060

[8].Prediction and the Analysis of Rheumatic Cardio Disease using kNN Classification with ACO S.Rajathi Dr.G.Radhamani Research Scholar, Part-Time Ph.D, Category –B Professor and Director Research and Development Center School of IT and Sciences Bharathiar University Dr.G.R.Damodaran College of Science Coimbatore-46 Coimbatore-14 srajathi@gmail.com radhamanig@hotmail.com

[9].Heart Disease Prediction with Machine Learning Techniques 2020 2nd International Conference on Advances in the Computing, Communication, the Control and Networking (ICACCCN) — 978-1-7281-8337-4/20/$31.00 ©2020 IEEE — DOI: 10.1109/ICACCCN51052.2020.9362842

[10].Heart Disease Prediction with the Logistic Regression Nor Fatihah Zulkiflee1 , Mohd Saifullah Rusiman1* 1Faculty of the Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, UTHM Pagoh, 86400 Muar, Johor, MALAYSIA DOI: https://doi.org/10.30880/ekst.2021.01.02.021 Received 11 June 2021; Accepted 19 July 2021; Available online 29 July 2021

[11].Heart Disease Prediction System Using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan Dept. of Electrical Engineering, Universitas Indonesia, Indonesia. i.ketut42@ui.ac.id.

[12].Prediction of Heart Disease with the Machine Learning Algorithms Md. Julker Nayeem, Sohel Rana, and Md. Rabiul Islam Published Online: November 30, 2022 ISSN: 2796-0072 DOI: 10.24018/ejai.2022.1.3.13

[13].Prediction of Heart Disease Using Machine Learning Prasanna M1, Shrijith Shetty P2, Mamatha K3 1,2 UG Scholars, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, Karnataka, India

[14].Analysis of Heart Disease Prediction with Machine Learning Saladi Novya Sree 1, K.Balaji Pranav Reddy 2, Bangarulakshmi Mahanthi3, Dr. S S Nandini4 1,2 Student, CSE, GITAM School of Technology, Visakhapatnam, India 3,4Assistant Professor, GITAM School of Technology, Visakhapatnam, India.