HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

by Raju Talari

Submission date: 23-Jun-2023 10:44AM (UTC-0400)

Submission ID: 2121305892

File name: 21121F0016_main_paper_2.docx (819.34K)

Word count: 2481

Character count: 14837

HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

BUSETTI SIVAIAH

21121F0016

ABSTRACT

Heart disease is among the conditions that people suffer from most frequently. Millions of people worldwide pass away each year as a result of it, making it one of the main causes of mortality. Heart disease can be characterised by issues with the heart valves, heart failure, arrhythmias, and coronary artery disease. Heart disease comes in more than 30 distinct forms. By allowing for prompt intervention and the right kind of care, early and precise cardiac disease prediction can greatly improve patient outcomes. In this model, we investigate the application of machine learning techniques for anticipating cardiac disease. We investigate a large dataset made up of patient details, such as demographics, medical histories, and clinical measures. The prospect of using machine learning algorithms to accurately predict the onset and diagnosis of heart illness is astounding. Predictive models are created using a variety of machine learning methods, such as logistic regression, decision trees, gradient boosting, XGBoost, random forests, SVM and ANN. The forecasting model is a hybrid model that ANN, Gradient Boosting, Decision Tree, SVM, Random Forests, and Logistic Regression. To increase the model's accuracy. To manage missing values, normalise features and solve class imbalance. The dataset has been pre-processed. The best accurate heart disease predictions are found using feature selection approaches. Performance indicators including accuracy, precision, recall, and area under the receiver operating the characteristic curve are used to train, validate, and assess the models. The major goal of this model is to put out a novel strategy for creating a model that successfully solves practical issues.

Keywords: performance analysis, gradient boosting, SVM, XGBoost, random forests, ANN and decision trees.

INTRODUCTION

Heart disease is a serious public health problem and the number one killer in the globe. Another name for it is cardiovascular disease. The World Health Organisation projects 17.9 million deaths annually, or nearly 31% of all deaths globally [5]. The early identification of people who are at a high risk of contracting the illness is essential for effective heart disease prevention and management.

There is a need for more precise and effective prediction models because traditional risk factors, including age, family history and gender have low predictive ability. Adopting a healthy lifestyle is key to preventing and managing heart disease. Depending on the individual situation, there are many different ways to treat heart disease, including medication, dietary modifications, medical treatments, and surgery.

The dataset, which is accessible in UIC machine learning repository. The Target labelled the dataset is made up of 303 rows and 14 columns. Data consist of both categorical and continuous information. The dataset contains the following variables: Age, Sex, Chest Pain, Resting Blood Pressure, Cholesterol, Fasting Blood Pressure, Resting Electrocardiogram, Max Heart Rate, ExAng, Oldpeak, Slope, Calcium, Thal, and Target Parameters [6].

Machine learning algorithms have shown great promise in their capacity to predict heart illness. In order to create more precise prediction models for identifying those at high risk of heart disease, these algorithms can examine vast volumes of data, including medical history, lifestyle choices, and clinical test results. Healthcare practitioners may be able to intervene sooner in the course of an illness by using Machine learning algorithms, which has the potential to enhance patient outcomes. However, creation of precise Machine learning models necessitates availability of high-quality data, wise feature selection, and cautious model training and validation. Furthermore, there are moral questions raised by the application of machine learning algorithms in the healthcare sector including data privacy, prejudice, and transparency.

Artificial neural networks (ANN), Gradient Boosting, XGBoost, and Support vector machines (SVM) are just a handful of the machine learning techniques we use in this model.

A popular statistical model used for binary classification tasks is logistic regression. It is a regression approach, despite the name and is frequently employed for classification jobs. By calculating the odds that a binary target variable would belong to each class, logistic regression models the connection between a collection of input variables (features) and the target variable.

The decision tree is one well-known and widely used supervised machine learning technique used for both classification and regression applications. Using a predetermined set of input characteristics, it creates a model resembling a tree of decisions and probable outcomes. Because they are simple to comprehend and use, decision trees are frequently used for both analytical and predictive purposes.

As part of an ensemble learning technique called Random Forest, many decision trees are combined to produce a reliable and precise prediction model. It is a supervised learning method that is used in applications for classification and regression. Random Forest improves the decision tree approach by reducing overfitting and improving prediction accuracy.

Both classification and regression tasks are carried out using Support Vector Machines (SVM), a well-known and highly effective supervised machine learning technique. By creating decision boundaries in a high-dimensional space depending on the input attributes given, it successfully separates and categorises data points. It works especially well in situations when the data in the input feature space cannot be separated linearly.

A machine learning approach called gradient boosting combines the forecasts of several weak learners, often decision trees, to produce a robust predictive model. It operates by incrementally developing new models that concentrate on correcting mistakes caused by earlier models, thus lowering the total prediction error.

Extreme Gradient Boosting, commonly referred to as XGBoost (Extreme Gradient Boosting), is an improved version of the gradient boosting, an ensemble learning technique that combines a number of weak models (typically decision trees) to produce a powerful prediction model. The speed and scalability of XGBoost have made it a popular choice for machine learning contests.

LITERATURE SURVEY

SALAH UD DIN, ASIF KHAN, AMIN UL HAQ, JALALUDDIN KHAN AND JALALUDDIN KHAN: The authors work This project at the University of Electronic Science and Technology in China, this model main intention is low Redundancy and high Relevance. In this model the authors use the algorithm like ANN, KNN and decision tree that can be used in conjunction with classification algorithms like SVM and Logistic Regression. The decision tree may be improved by adopting these feature selection strategies by concentrating on the most relevant and informative characteristics, increasing its accuracy and understandability [1].

MIRJAM JONKMAN, ABHIJITH REDDY BEERAVOLU, F. M. JAVED MEHEDI SHAMRAT, ASIF KARIM, EVA IGNATIOUS, SHAHANA SHULTANA, SAMI AZAM, AND FRISO DE BOER: The researchers done on Daffodil International University's in Dhaka, Bangladesh (1225). The right features can be selected using the Relief and Least Absolute Shrinkage and Selection Operator (LASSO) techniques by combining the conventional classifiers with bagging techniques like the Decision Tree, KNN, Random Forest, AdaBoost and Gradient Boosting.

SYED MD. HUMAYUN AKHTER, HIRA FATIMA, GHULAB NABI AHMAD, SHAFIULLAH, AND ABDULLAH ALGETHAMI: the authors do this work on Mangalayatan University, Uttar Pradesh, India. It was the site of the authors' work on this project. Researchers used a number of methods to create this model to predict cardiac illness, including Decision Tree, SVM, Linear Discriminant Analysis, Random Forest, Gradient Boosting Classifier, and K-Nearest Neighbours . To choose pertinent characteristics for the prediction job, they also included the sequential feature selection method [3].

SHAFIULLAH, ABDELAZIZ SALAH SAIDI, GHULAB NABI AHMAD, AND IMDADULLAH: The authors prepare this model at the Mangalayatan University, Uttar Pradesh, India. They used the GridSearchCV and many Machine Learning algorithms to predict cardiac illness, including Logical regression, KNN and SVM. Verification is done using 5-fold cross-validation approach to assess performance of models. The effectiveness of the various models in predicting and diagnosing heart illnesses is examined and compared using these datasets. The study intends to evaluate the generalization and robustness capabilities of the models across various populations and data sources [4] by utilizing a variety of datasets.

EXISTING SYSTEM

The majority of cardiac problems are mostly avoidable, and early diagnosis and treatment considerably improve the prognosis. Examples of such changes in lifestyle include quitting smoking, eating healthier, exercising, and avoiding obesity. Due to the multifactorial nature of many contributing risk factors such as diabetes, BP, cholesterol, etc., it is challenging to identify high-risk individuals [7].

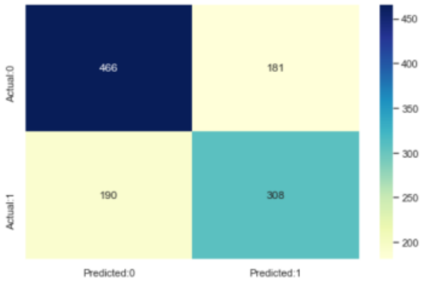
Due to its superiority in pattern identification and classification, ML has demonstrated that it is useful in helping to make judgements and predictions from huge amount of data produced by the health sector about cardiac disease [7]. Using the Framingham dataset, they will investigate various machine learning techniques in this model to determine if a patient has a chance of getting coronary disease (CHD). The dataset may be found on the Kaggle website. [8].

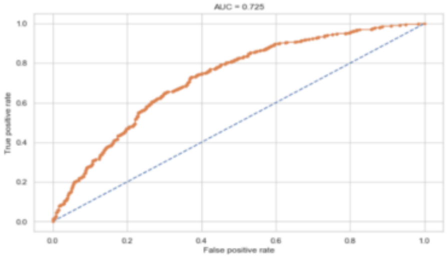
In order to select the most accurate outcome, we analyse four computational methods for this model: logistic regression, KNN, decision trees, and SVM.

Logistic regression

We get a 67.6% accuracy rate.

The logistic regression f1 score is 62.41%

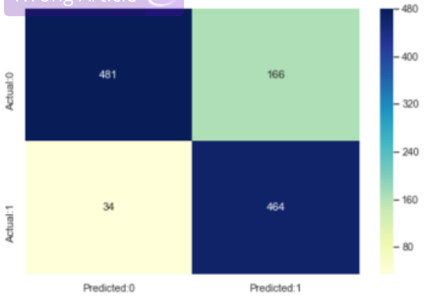


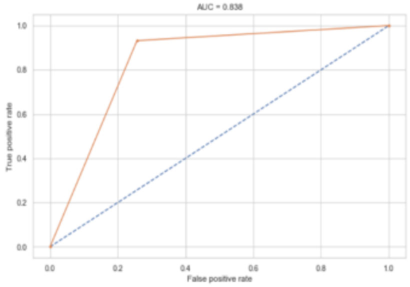


k-Nearest Neighbors

Accuracy rate is 82.53%

f1 score is 82.27%.

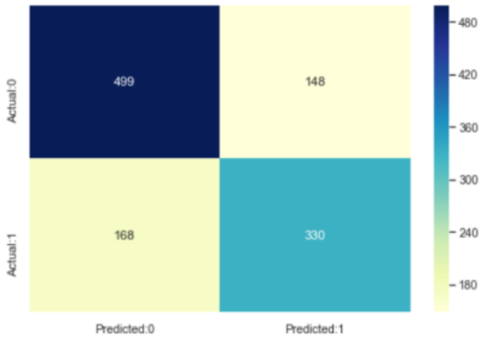


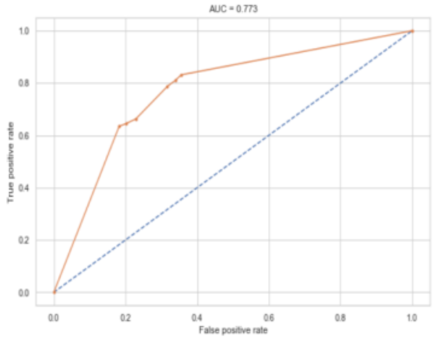


Decision Trees

Accuracy rate is 72.4%.

f1 score is 67.62%.

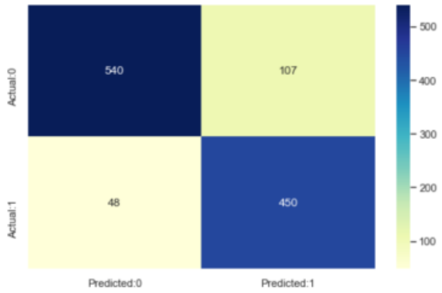


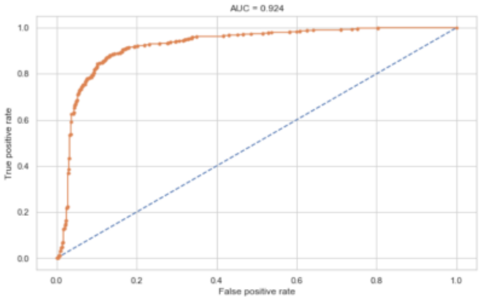


Support Vector Machine

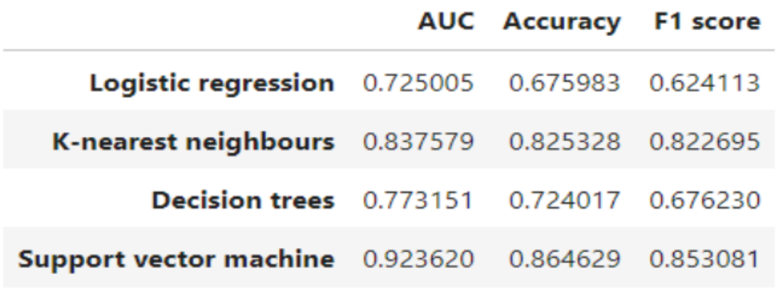
Accuracy rate 86.46%.

f1 score is 85.31%.





Model Comparison



	AUC	Accuracy	F1 score
Logistic regression	0.725005	0.675983	0.624113
K-nearest neighbours	0.837579	0.825328	0.822695
Decision trees	0.773151	0.724017	0.676230
Support vector machine	0.923620	0.864629	0.853081

Disadvantages

1. Time complexity and space complexity is more, because the model use all the algorithm models each and every time.
2. Developing accurate machine learning models requires access to high-quality data. If the data used to train the model is incomplete, inaccurate, or biased, it can lead to unreliable or incorrect predictions.

PROPOSED SYSTEM

The suggested technique uses a hybrid machine learning algorithm, which combines many methods, to predict cardiac disease. In this model, we combine gradient boosting, SVM, decision trees, ANN, random forests, logistic regression, and XGBoost.

Logistic Regression

In fact, supervised learning techniques such as logistic regression are frequently employed for binary classification issues. In order to determine the likelihood that a binary event will occur, such as determining whether a result will be "yes" or "no," "true"or "false,"or "positive"or "negative," binary logistic regression is used.

Logistic regression's mathematical foundations

Output is always in the range of 0 (does not occur) to 1 (occurs).

$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X)$$

***'hΘ(x)'** is output of logistic function , where $0 \leq h\Theta(x) \geq 1$*
***'β1'** is the slope*
***'βo'** is the y-intercept*
***'X'** is the independent variable*

$(\beta o + \beta 1{*}x)$ - derived from equation of a line $Y(predicted) = (\beta o + \beta 1{*}x) + Error\ value$

Decision Tree

Decision support uses a hierarchical model known as a decision tree to represent actions and their different outcomes, including chance occurrences, resource costs, and utility. Conditional control statements are employed in this non-parametric, supervised learning algorithmic paradigm, which may be applied to both classification and regression problems. The tree structure, which is arranged hierarchically to resemble a tree, is made up of a root node, branches, internal nodes, and leaf nodes. Output from decision trees frequently consists of binary options, such as "yes" or "no."

Random Forest

Random forest is used for both classification and regression. Random Forest combines the predictions of a group of decision trees to provide predictions or classifications that are more reliable and accurate. By decreasing overfitting and boosting the variety of each tree's predictions, this method enhances the model's functionality and generalisation.

Support Vector Machine

Although it also does well on smaller datasets, the robust supervised method Support Vector Machine (SVM) shows its strength most prominently on complicated datasets. SVM used for classification and regression tasks, although it is well known for its prowess in handling classification issues. SVM can handle datasets with complicated decision boundaries and nonlinear interactions between variables by determining the best hyperplane that maximum separates various classes. As a result, SVM is an effective technique for solving classification issues since it offers excellent accuracy and resilience across a range of domains.

Gradient Boosting

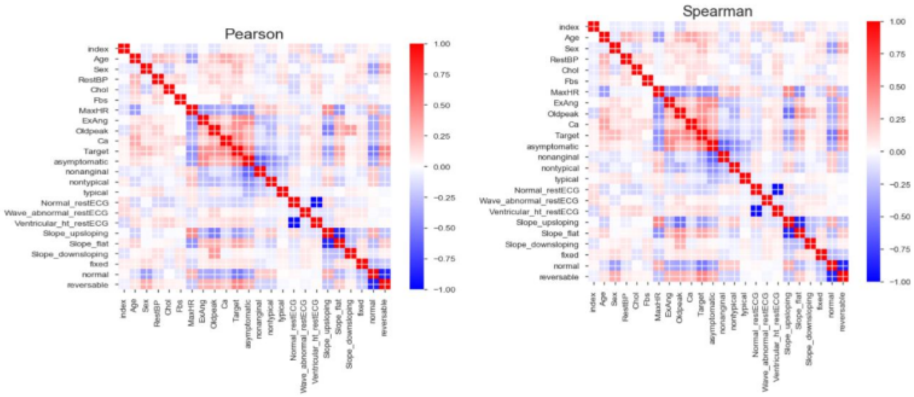
A machine learning approach called gradient boosting combines the forecasts of several weak learners, often decision trees, to produce a robust predictive model. It operates by incrementally developing new models that concentrate on correcting mistakes caused by earlier models, thus lowering the total prediction error.

XGBoost

Extreme gradient boosting, commonly referred to as XGBoost, is a better form of gradient boosting, an ensemble learning approach that combines a number of weak models to build a potent proposed model. The speed and scalability of XGBoost have made it a popular choice for machine learning contests.

In this model we use several steps is to collect relevant data on heart disease from different sources such as hospitals, medical records, and research studies. Tha data consist of various fields, based on those fields the model should work.

Correlations



Advantages

1. Increased Accuracy: The hybrid approach used by the suggested model will assist to increase accuracy.

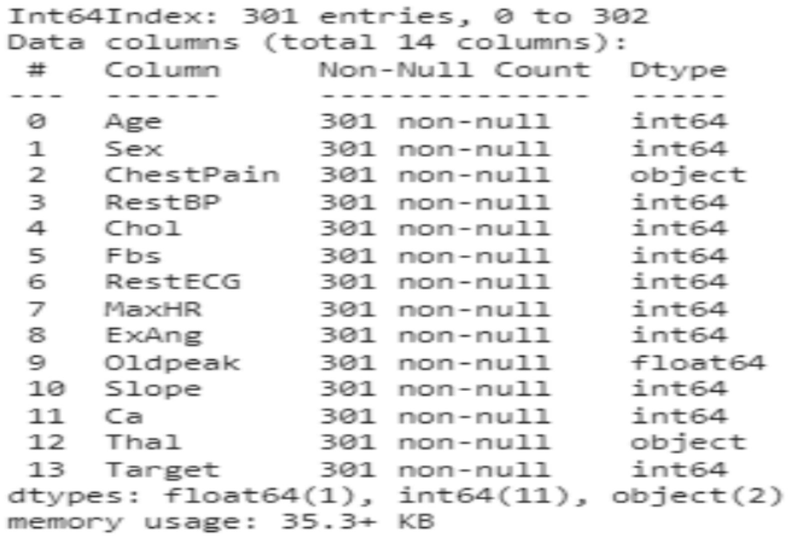
2. Low latency: When compared to the existing model, the suggested system has a low time complexity.

EXPREMENTAL RESULT

We employ a mixed machine learning algorithums in our model to predict cardiac disease. Decision trees, random forests, SVM, gradient boosting, XGBoost, ANN and logistic regression are some of the methods that are combined in the hybrid model. A dataset gathered from numerous sources, including hospitals, medical records, and research projects.

We used Data Cleansing, Data Visualisation, Overview, calculate Training Accuracy, and calculate Test Accuracy to assess the performance of our hybrid model.Cross-validation was employed to guarantee robustness and reduce overfitting. Each of the k subsets of the dataset was utilised as a test set, while the remaining subsets were used as the training set. The outcomes were averaged after this process was carried out k times.

Data Visualization



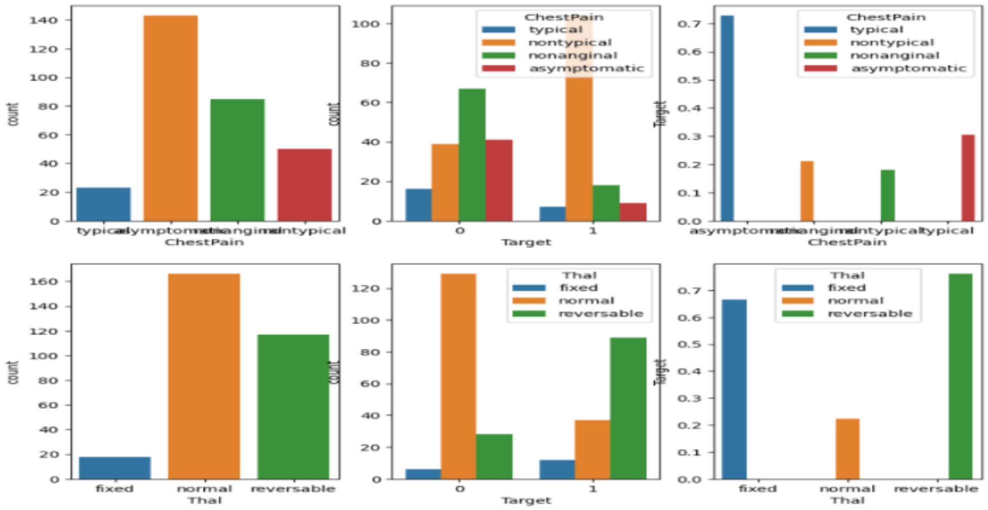
```
Int64Index: 301 entries, 0 to 302
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Age        301 non-null    int64  
 1   Sex        301 non-null    int64  
 2   ChestPain  301 non-null    object 
 3   RestBP     301 non-null    int64  
 4   Chol       301 non-null    int64  
 5   Fbs        301 non-null    int64  
 6   RestECG    301 non-null    int64  
 7   MaxHR      301 non-null    int64  
 8   ExAng      301 non-null    int64  
 9   Oldpeak    301 non-null    float64
 10  Slope      301 non-null    int64  
 11  Ca         301 non-null    int64  
 12  Thal       301 non-null    object 
 13  Target     301 non-null    int64  
dtypes: float64(1), int64(11), object(2)
memory usage: 35.3+ KB
```

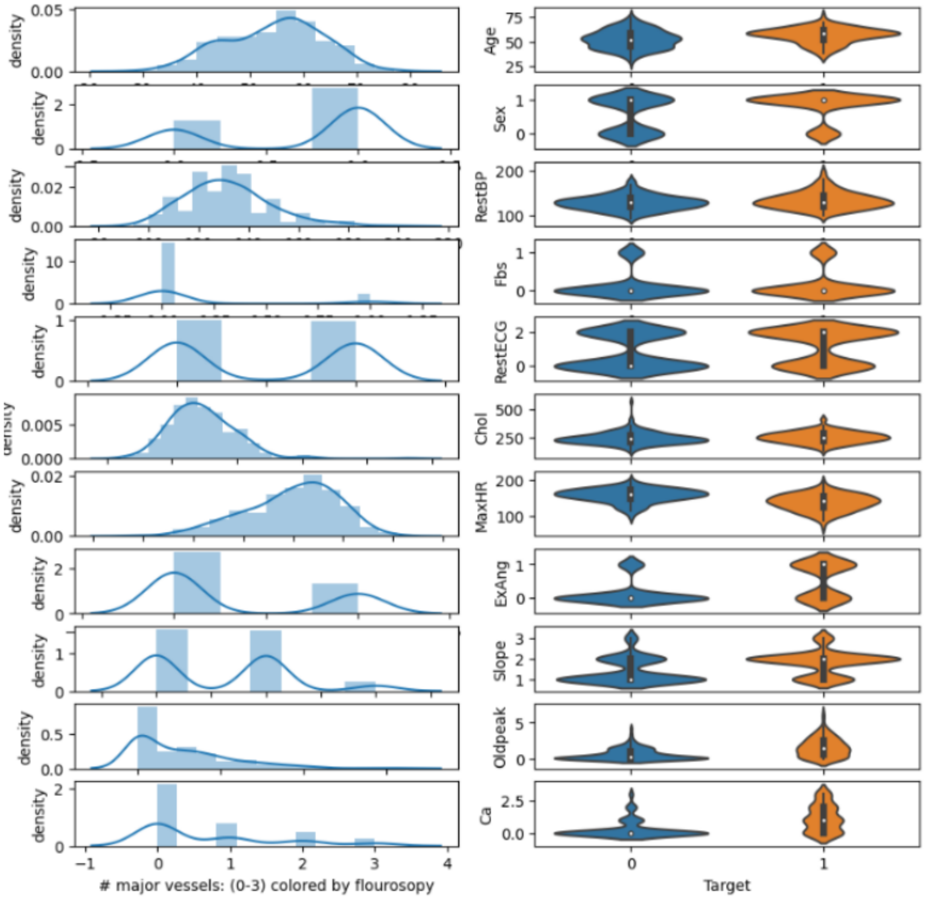
Pandas Profiling Provide the Report of data set including Corelation

S/V

```
Summarize dataset: 100%  48/48 [00:12<00:00, 1.64it/s, Completed]
Generate report structure: 100%  1/1 [00:18<00:00, 18.80s/it]
Render HTML: 100%  1/1 [00:02<00:00, 2.88s/it]
```

Ploting Function For Categorical Data " Chest Pain" && "Thalassemia"





According to the findings of our experiments, the hybrid machine learning algorithm performed effectively in predicting heart disease. We were able to increase the overall prediction accuracy by combining decision trees, SVM, random forests, gradient boosting, XGBoost, and ANN are examples of data analysis techniques. Comparing the hybrid model to the component algorithms by themselves, it showed improved accuracy and fast response times.

CONCLUSION

This model highlights the effectiveness of the hybrid machine learning algorithm in the prediction of heart disease. The combination of multiple algorithms and the incorporation of various risk factors enhance the accuracy and reliability of predictive model, potentially aiding in early detection and proactive management of heart disease.

REFERENCES

[1] JIAN PING LI, AMIN UL HAQ, SALAH UD DIN, JALALUDDIN KHAN, ASIF KHAN, AND ABDUS SABOOR, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" and Digital Object Identifier 10.1109/ACCESS.2020.3001149

[2] PRONAB GHOSH, SAMI AZAM, MIRJAM JONKMAN, (Member, IEEE), ASIF KARIM, F. M. JAVED MEHEDI SHAMRAT, EVA IGNATIOUS, SHAHANA SHULTANA, ABHIJITH REDDY BEERAVOLU, AND FRISO DE BOER," Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques" and Digital Object Identifier 10.1109/ACCESS.2021.3053759

[3] GHULAB NABI AHMAD, SHAFIULLAH, ABDULLAH ALGETHAMI, HIRA FATIMA, AND SYED MD. HUMAYUN AKHTER, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection" and Digital Object Identifier 10.1109/ACCESS.2022.3153047

[4] GHULAB NABI AHMAD, HIRA FATIMA1, SHAFIULLAH, ABDELAZIZ SALAH SAIDI, (Member, IEEE), AND IMDADULLAH, (Senior Member, IEEE), "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV" and Digital Object Identifier 10.1109/ACCESS.2022.3165792

[5]WORLD HEALTH ORGANIZATION https://www.who.int/india/health-topics/cardiovascular-diseases

[6] UCI machine learning repository, Heart Disease and https://archive.ics.uci.edu/dataset/45/heart+disease

[7] amayomode repository, HEART DISEASE PREDICTION USING MACHINE LEARNI and https://github.com/amayomode/Heart-Disease-Risk-Prediction

HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

ORIGINALITY REPORT

9%	5%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Marmara University Student Paper	2%
2	www.euro-online.org Internet Source	1%
3	Submitted to University of Chichester Student Paper	1%
4	Submitted to University of Kent at Canterbury Student Paper	1%
5	Submitted to Indian Institute of Engineering Science and Technology Student Paper	1%
6	ijircce.com Internet Source	1%
7	G. Krishna Lava Kumar, S. Asif, U. Veeresh. "A study on heart disease prediction using supervised machine learning models", AIP Publishing, 2021 Publication	1%

8	ash.confex.com Internet Source	1 %
9	www.npmjs.com Internet Source	<1 %
10	JianPing Li, Amin ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, Abdus Saboor. "Heart Disease Identification Method Using Machine Learning classification in E-Healthcare", IEEE Access, 2020 Publication	<1 %
11	www.jetir.org Internet Source	<1 %
12	www.researchgate.net Internet Source	<1 %
13	Ran Wang, Shilei Lu, Qiaoping Li. "Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings", Sustainable Cities and Society, 2019 Publication	<1 %
14	www.journaltocs.ac.uk Internet Source	<1 %
15	Swati Srivastava, Sanjana Solomon, P. Madhavan. "Effective Analysis of Human Cardiovascular Disease via Ensemble Learning", 2022 International Conference on	<1 %

	Power, Energy, Control and Transmission Systems (ICPECTS), 2022 Publication	
16	www.science.gov Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

PAGE 1

PAGE 2

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Missing "," You may need to place a comma after this word.

PAGE 3

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word. Consider using the article **the**.

Article Error You may need to use an article before this word.

Missing "," You may need to place a comma after this word.

Missing "," You may need to place a comma after this word.

PAGE 4

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Possessive This word may be a plural noun and may not need an apostrophe.

PAGE 5

Article Error You may need to remove this article.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Prep. You may be using the wrong preposition.

Article Error You may need to use an article before this word.

PAGE 6

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Article Error You may need to use an article before this word. Consider using the article **the**.

PAGE 7

Article Error You may need to use an article before this word.

PAGE 8

Article Error You may need to use an article before this word.

Proofread This part of the sentence contains a grammatical error or misspelled word that makes your meaning unclear.

PAGE 9

Article Error You may need to use an article before this word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

PAGE 10

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

PAGE 11

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

PAGE 12

PAGE 13