ZIDIO DEVELOPMENT

PREDICTION OF CARDIOVASCULAR DISEASE WITH MACHINE LEARNING

GROUP 13

BY

FERANMI OYEDARE,

VADDE VIGNESH,

ADEBUSOYE OLUWAYIMIKA ELIZABETH

19TH OF MARCH, 2024

1.ABSTRACT

Cardiovascular disease (CVD) causes abnormal blood vessel and heart function, which frequently results in death or paralysis. Early and automated identification of CVD can therefore save a great deal of human life. Although several studies have been conducted to accomplish this goal, performance and dependability can yet be enhanced. This work is one additional step in that direction. The dataset given by the institution was used in this work to test four trustworthy machine learning models for CVD detection: Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier, and Logistic Regression. Eliminating outliers and characteristics with null values maximizes the performance of the models. The results of the analysis showed that, in contrast to the other models, the Random Forest Classifier model yields a better detection accuracy of 75% and an area-under-the-curve value of 82.41%. As a result, it was suggested to use the Random Forest Classifier model for automated CVD identification. Moreover, other common data sets can be used to evaluate the performance of the suggested model.

Keywords: cardiovascular disease, machine learning algorithms, Random Forest Classifier, Logistic Regression, Gradient Boosting Classifier, Decision Tree Classifier.

1.INTRODUCTION

As the primary cause of mortality worldwide, cardiovascular disease (CVD) has emerged as a major global public health concern. Patients, their families, and these nations' governments have all paid high socioeconomic costs as a result. Using risk stratification, prediction models can identify patients who are at high risk for cardiovascular disease (CVD). Following that, specific interventions for this population, such dietary modifications and statin usage, can help lower that risk and support primary CVD prevention (1). The most current estimates predict that by 2030, CVD would be the cause of mortality for around 23 million persons. Heart failure, atrial fibrillation, and myocardial infarction are examples of distinct forms of CVD (2, 3). The results of blood tests that assess variables like renal function, liver function, and cholesterol levels, as well as racial or ethnic background, age, gender, body mass index (BMI), height, and length of torso, can all have an impact on the incidence of cardiovascular disease (4, 5). Early diagnosis of cardiovascular diseases (CVD) stands as a paramount objective in medical practice, offering the potential to mitigate morbidity and mortality rates significantly. However, the conventional methodologies employed in CVD risk assessment exhibit limitations, primarily rooted in their simplistic linear modeling of risk factors. Such an approach fails to capture the intricate non-linear interactions inherent within the multifaceted landscape of CVD etiology. Consequently, there exists a pressing need to refine risk assessment models by integrating a comprehensive array of risk factors and elucidating the nuanced correlations between these factors and disease outcomes. In response to this imperative, the present study endeavors to explore the utility of machine learning (ML) algorithms in enhancing the accuracy of cardiovascular risk prediction within large population cohorts under primary care settings. A thorough examination of the body of research (WHO) indicates that ML-based models for the identification of CVD are becoming increasingly popular, highlighting the potential of sophisticated computational methods to transform prognostic efforts in the field of cardiovascular medicine.The investigation's methodology comprises a thorough analysis of several machine learning (ML) techniques, Random Forest Classifier, Gradient Boosting Classifier (6), Decision Tree Classifier, and Logistic Regression approaches, in order to determine their effectiveness in predicting CVD risk. Specifically, every algorithm is examined in terms of its advantages and disadvantages, which helps with well-informed optimization and selection procedures. Furthermore, the study encompasses a meticulous statistical analysis of input datasets, aimed at elucidating the impact of data range on CVD predictions. This analysis includes a comprehensive correlation study delineating the interplay between categorical and continuous patient features. Additionally, the utilization of data visualization techniques, such as scatter plots, serves to elucidate the significance of correlations among key features, thereby enhancing interpretability and insight generation. Optimally, the findings of this research endeavor hold profound implications for the advancement of cardiovascular risk assessment practices within primary care contexts. By leveraging the power of machine learning and robust statistical methodologies, this study seeks to furnish clinicians with invaluable tools for the early detection and proactive management of cardiovascular diseases, thereby effectuating tangible improvements in patient outcomes and public health at large.

2. MATERIALS AND STEPS

The goal of this study was to determine whether or not a patient would develop CVD if a set of clinical information is available. The confusion matrix of each technique was obtained, and out of 1066 occurrences in the data set, 852 (80%) were used to train the four models. To test the trained models, 214 instances were fed to know the class. The material required for CVD detection is the data of patients from publicly available standard CVD data as given to us by the facilitators. ThclassificatiolgorithmsusedareLogisticRegression,DecisionTreeClassifier,RandomForestClassifierGradientBoostingClassifier. Generally, the method comprises training of the proposed model via respective learning algorithms using relevant input test data of patients and then validating these models based on test data of patients. Finally, performance measurements are evaluated using accuracy score, confusion matrix, confusionmatrixdisplay and classification report. Then, they were compared to see the best performing model; RandomForestClassifier was the best model with an accuracy of 72%.

The following steps were carried out to predict CVD:

Step 1: Relevant CVD data set was first collected

Step 2: The necessary libraries were imported and the data was also imported to jupyter notebook.

Step 3: We gathered information from the data such as the shape of the data set, the number of null values present in each variable, the number of duplicate values, statistical information of the data and the presence of outliers

Step 4: The data samples were cleaned by filling up null values using the forwardfill method, duplicate values were dropped and outliers were removed using the data between the 10% percentile and 90% percentile.

Step 5: The data was preprocessed by checking for imbalanced data. We were able to resolve tat using Random oversampler from the imbalance library. The target variable was discovered to be a continuous numerical feature while we needed a binary feature. We were able to preprocess this by using our intuition.

Step 6: Exploratory Data Analysis was performed to identify patterns and trend in the data.

Step 7: Feature Engineering was performed using the age, resting_bp and chol features.

Step 8: Attributes with low predicting power were dropped as there were no strongly correlated features.

Step 9: The data was divided into features and target variables (X and y).

Step 10: The data was scaled using standardscaler from the sklearn library to reduce the complexity of the data.

Step 11: The data was divided into training and testing sets.

Step 12: Four effective ML algorithms were chosen to classify the selected features.

Step 13: Hyperparamater tuning was done to improve the performance of the four models.

3.DATA SOURCE

The CVD data set used for developing the detection models was provided and has been converted into a.csv comma-separated file. It contains 1592 samples and 14 attributes. Only 13 important test attributes (age, sex, chest pain (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic result (Restecg), maximum heart rate (thalach), exercise-induced angina (exang), ST depression (old peak), slope of peak ST segment (slope), number of major vessels (Ca), thallium stress result (thal)), and one target output (1 = patient having CVD, 0 = patient not having CVD) have been considered out of the 14 attributes to train and test the model. These are presented in Table A1. Our target value taken is whether a person has CVD (near to 1) or does not have CVD (close to 0). The data set was imbalanced as 62.412% of the patients had CVD and 37.5878% of the patients were normal

The data set contains both categorical and continuous features. The data set consists of patients between the ages of 29 and 77. Pandas, NumPy, sklearn, seaborn, and matplotlib python libraries were used to analyse and visualize the data. Two standard and reliable MLP and K-NN ML methods were employed for binary classification (CVD or no CVD). The below diagram (Figure 1) is the frequency of the data in each column.

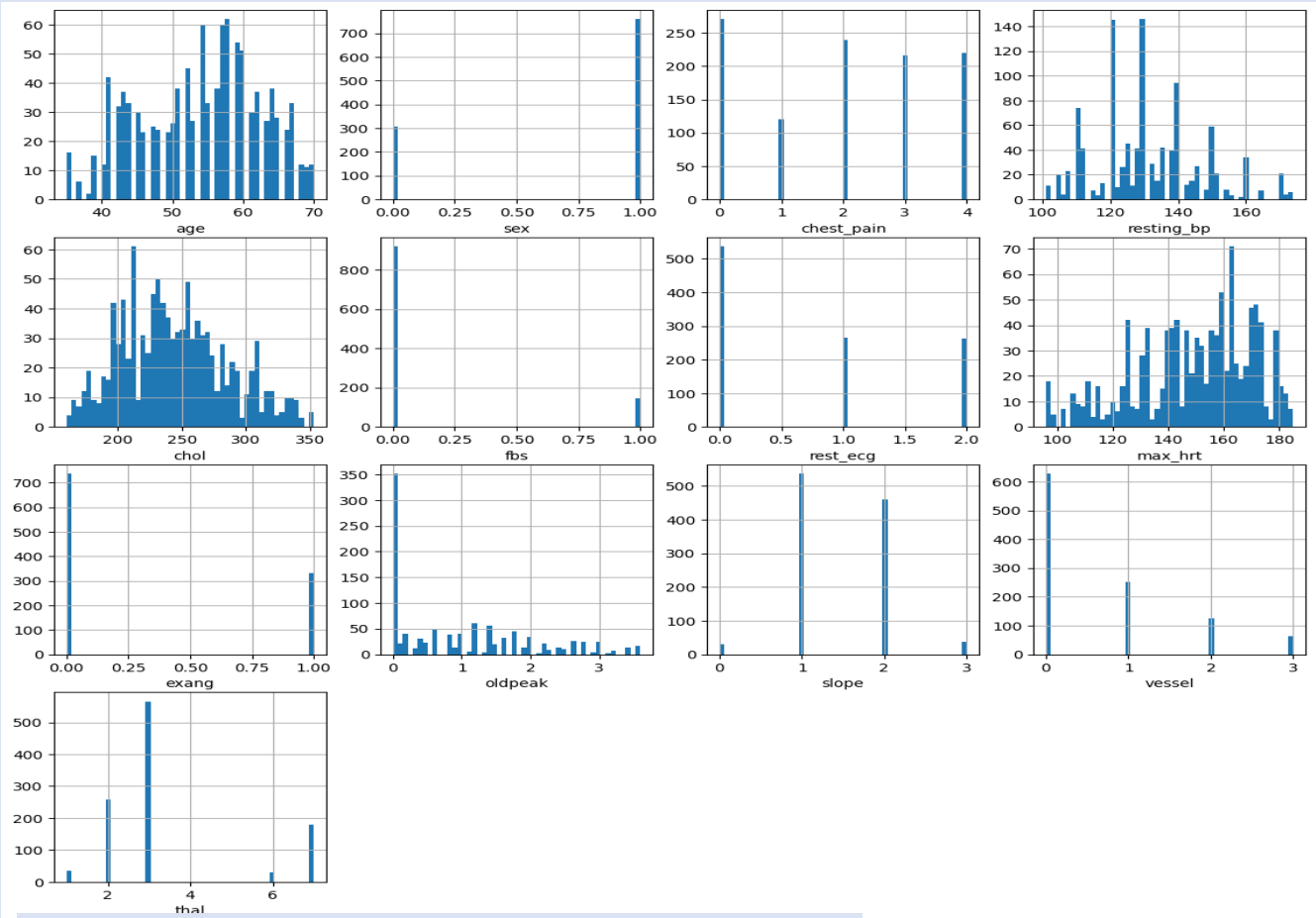

Figure 1

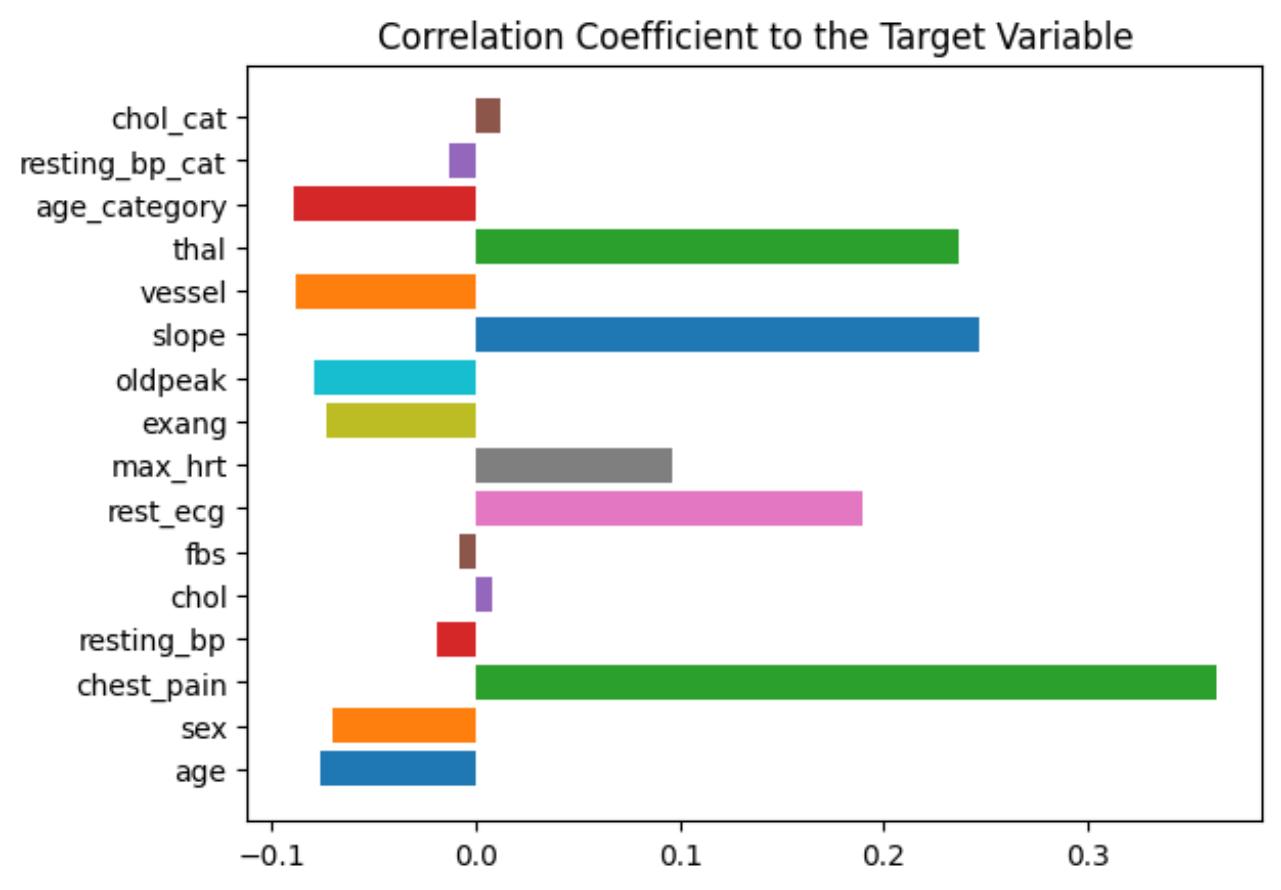

Figure 2

Correlation plot for CVD data set categorical features (cp: chest pain, fbs: fasting blood sugar, Restecg: resting electrocardiographic result, exang: exercise-induced angina, slope: slope of peak ST segment.

3.1. Pre-processing of CVD data

Public data sets contain a lot of noisy and missing data. Preprocessing of this data reduces skewed readings and improves forecast accuracy. Aggregation, standardization, and smoothing are all part of the pre-process phase. The correlation matrix indicates whether the features in the CVD data set are positively or negatively associated with one another and is used to determine the correlation between various features during the data pre-processing stage. Following the data set's pre-processing, dummy variables are created for a number of categorical variables, including cp, sex, chol, and trestbps. The resulting data are then scaled before being used to train machine learning models. Standard normal distribution was used to scale down each variable, and cross-validation was used.

3.1.1. A scatter diagram A mathematical figure in cartesian coordinates that displays the relationship between two variables of a particular data set is called a scatter plot, also known as a data visualization plot. It displays the connection between two numerical variables. Two variables are associated if they are located on a line or curve. As a result, it establishes the relationship between a given cause and effect objectively. Several scatter plots are used in this study to show possible CVD root causes.

3.2. ML

As an aspect of artificial intelligence (AI), machine learning (ML) refers to the process by which a model learns from prior experience without explicit programming. Various machine learning classifiers use medical data to forecast or classify diseases. The system learns using traits known as biomarkers of heart disease through supervised learning, which uses labeled input data for training. The machine receives patient data, and labeled results are produced. For creating data analysis models, several machine learning classifiers are available, including the Gradient Boosting Classifier, Random Forest Classifier, and Decision Tree Classifier. In general. Each classifier's primary goal is to create a model with outstanding disease detection power. DecisionTreeClassifier and GradientBoostingClassifier are the classifiers that were employed in this study. possess remarkable detecting potentials in general.

3.3. ML classifiers

3.3.1. Decision Tree Classifier:

Decision trees are a type of supervised learning algorithm used for both classification and regression tasks. The training process of a decision tree involves recursively partitioning the feature space into smaller regions based on the values of input features. This partitioning is done in a way that maximizes the homogeneity of the target variable within each resulting region. During training, the algorithm selects the best feature to split the data at each node based on criteria such as Gini impurity or information gain. This process continues until a stopping criterion is met, such as reaching a maximum depth or having nodes with a minimum number of samples. Decision trees are relatively interpretable, making them useful for understanding feature importance. However, they are prone to overfitting, especially with complex datasets.

3.3.2. Random Forest Classifier

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree in the random forest is trained on a subset of the original dataset, called bootstrap samples, and at each split, a random subset of features is considered. This randomness helps to decorrelate the individual trees, reducing overfitting and improving generalization performance. During classification, each tree's output is tallied, and the class with the most votes becomes the final prediction. Random forests are known for their robustness and ability to handle high-dimensional datasets with noisy features. They are less prone to overfitting compared to individual decision trees and often yield high accuracy across various types of data.

4.3.3. Gradient Boosting Classifier

Gradient Boosting is another ensemble learning technique that builds a strong predictive model by combining the predictions of multiple weak learners, typically decision trees. Unlike random forests, which build trees independently, gradient boosting builds trees sequentially, with each tree trained to correct the errors of its predecessors. During training, the algorithm starts with an initial model (usually a simple one like a single leaf), then iteratively adds new trees that minimize a loss function, such as mean squared error or log loss. Each new tree is fitted to the residual errors of the combined model, gradually reducing the overall error. Gradient boosting is highly effective

in producing accurate predictions and is robust to overfitting when appropriate hyperparameters are tuned carefully. However, it can be computationally expensive and sensitive to noisy data. Regularization techniques such as shrinkage (learning rate) and tree depth limitation are often applied to control model complexity and prevent overfitting.

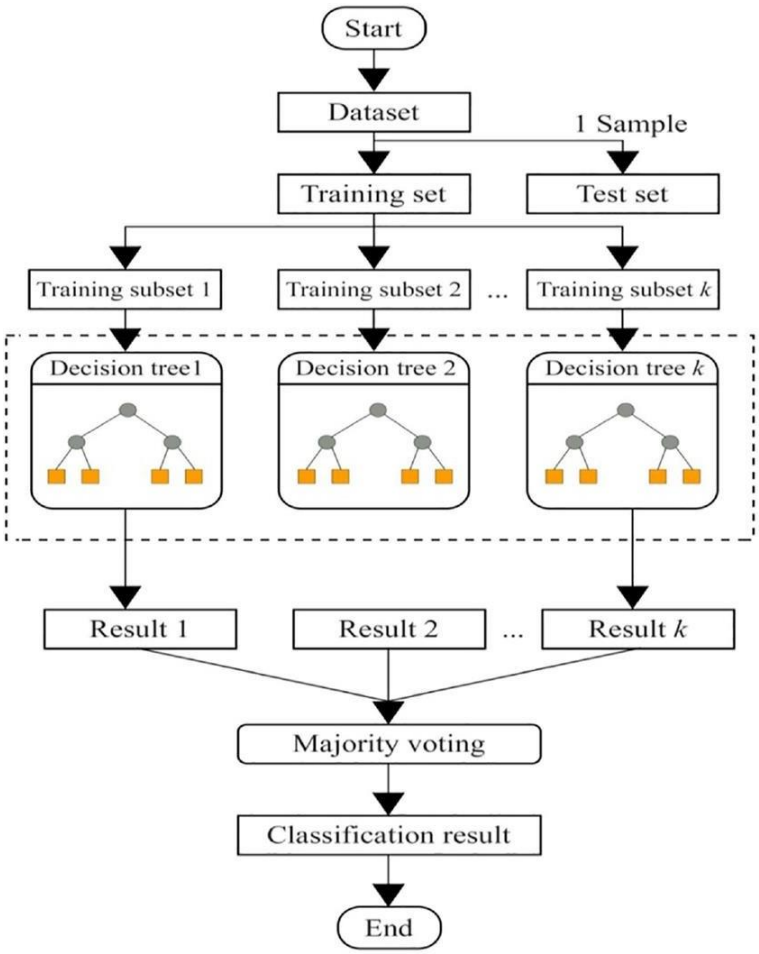

Figure 3

schematic diagram of random forest classifier in machine learning

4.RESULTS

Table 1

Definition of confusion matrix parameters

Confusion matrix parameters	Description
True positive	Instances where we predicted yes (patient has the CVD), and it turned out to be correct.
True negative	Instances where a patient does not have CVD and was predicted to not have CVD.
False positive	Instances where a patient does not have CVD but was predicted to have CVD.
False negative	Instances where a patient does not have CVD and was predicted to not have CVD.

Figures 4 and and 5 show the confusion matrix of Gradient Boosting Classifier and Random Forest Classifier ML models, respectively. Table 2 shows the combined confusion matrix, which reveals that between the two approaches, the Gradient Boosting Classifier and Random Forest Classifier model predicted TPs (78 vs 75), TNs (74 vs 72), FNs (28 vs 30), and FPs (34 vs 37) respectively.

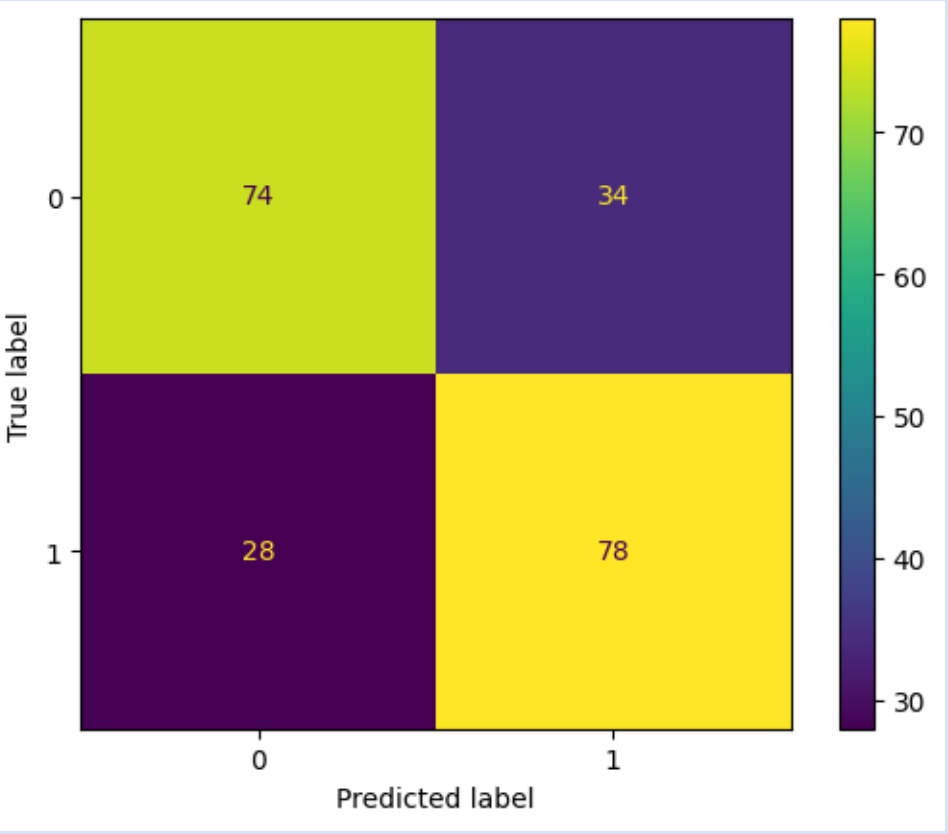

Figure 4

Confusion matrix for CVD prediction using the Gradient Boosting Classifier

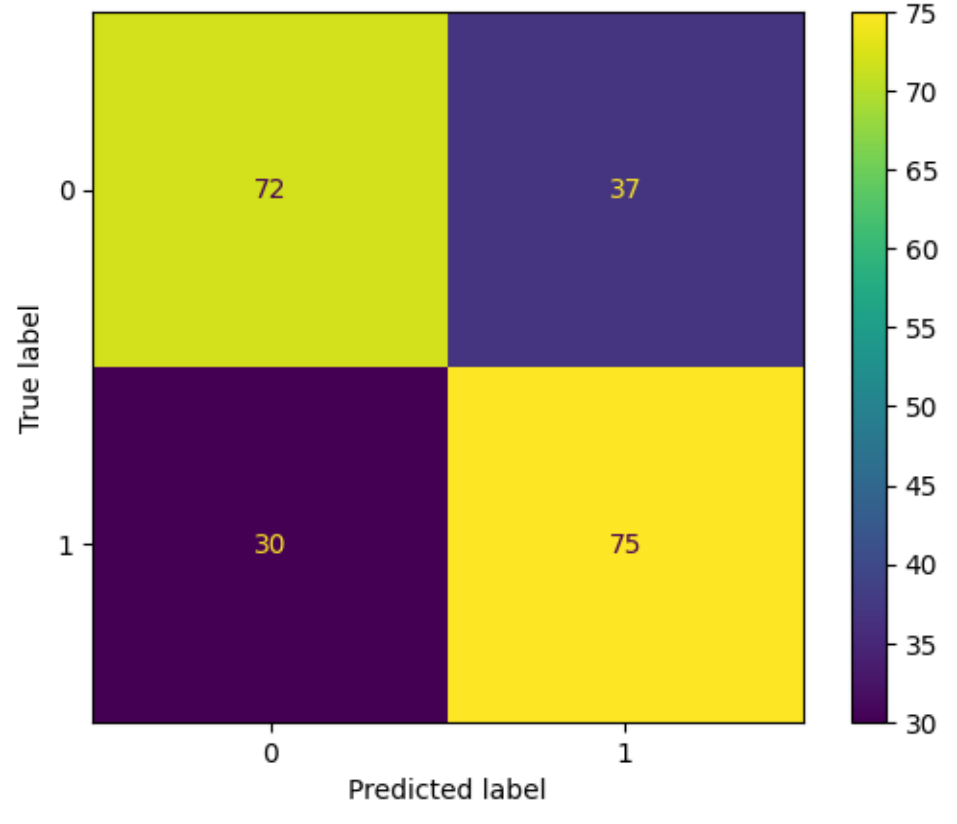

Figure 5

Confusion matrix for CVD prediction using the Random Forest Classifier.

Table 2

Confusion matrix results of ML models

Confusion matrix parameters	**ML algorithms**	
	GBC	**RFC**
TN	74	72
TP	78	75
FP	34	37
FN	28	30

4.1. Column Charts between features of CVD

The graph in Figure 6 shows people with normal resting electrocardiographic result are more than people having ST-T wave abnormality and left ventricular hypertrophy. Figure 7 shows that People with fast blood sugar < 120mg/dL are the ones who have more of the heart disease. In Figure 8 from our analysis, people with moderate cholesterol are the ones with the highest number of the heart disease . Figure 9 shows that people with elevated resting bloodpressure(120 - 139) are prone to have the heart disease . In Figure 10 People with asymptomatic chest pain tend to have the heart disease more as compared to typical angina, atypical angina, nonanginal chest pain. Figure 11 shows that senior_adult age category have the highest number of people with the disease . Figure 12 shows that males have the highest number of people with the heart disease. A data visualization plot showing the correlation between individual features is shown in Figure 13.

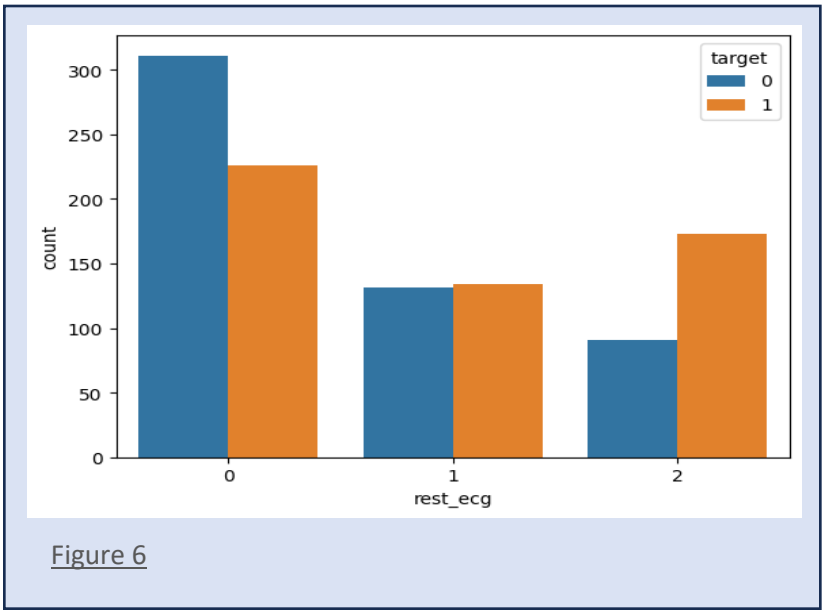

Figure 6

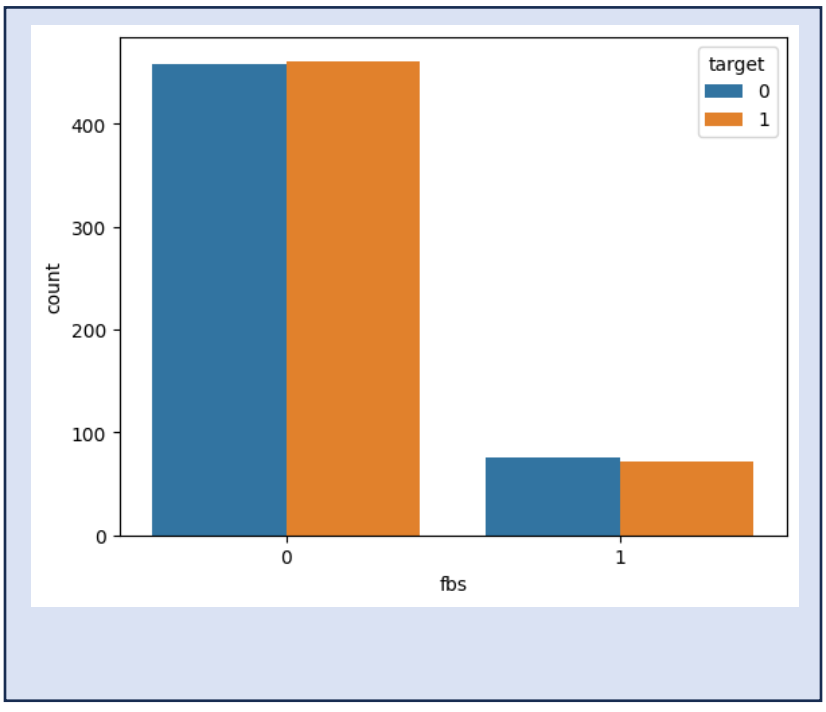

Figure 7

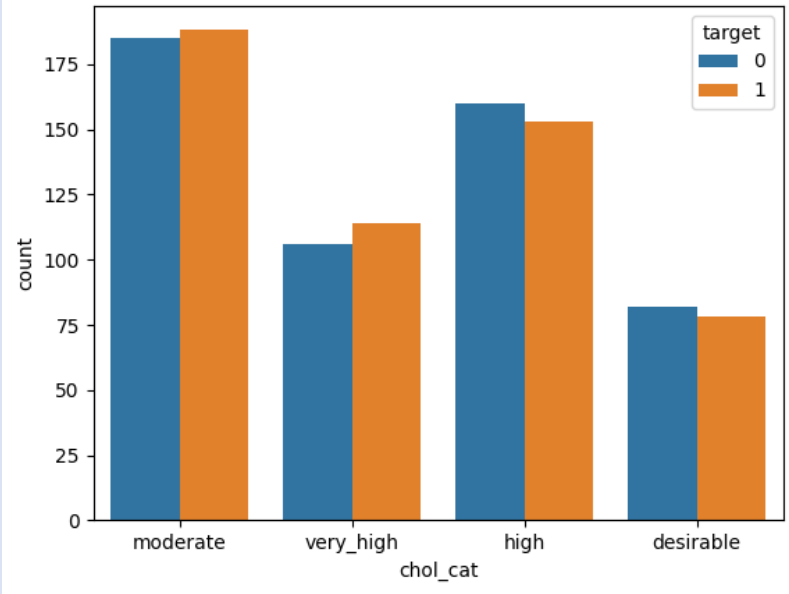

Figure 8

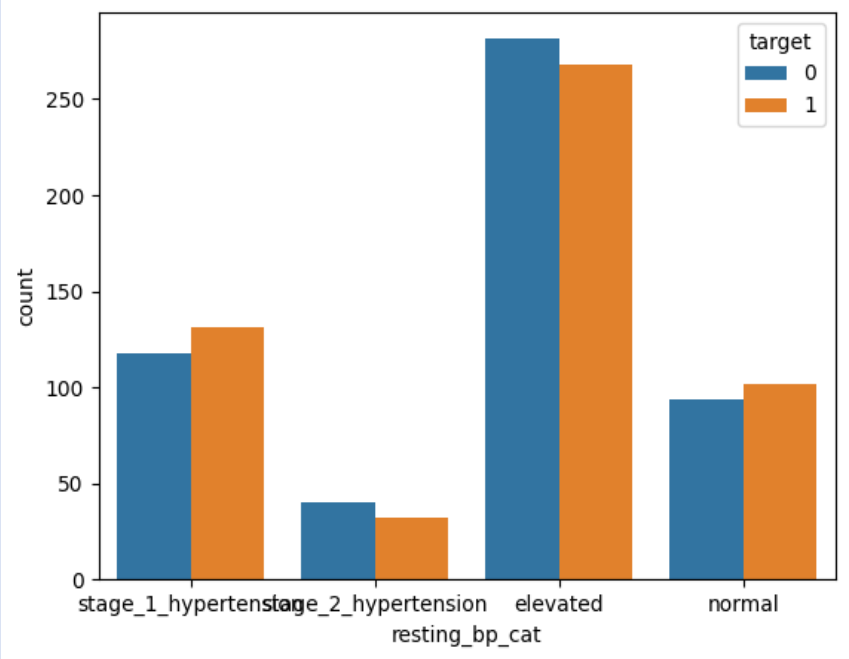

Figure 9

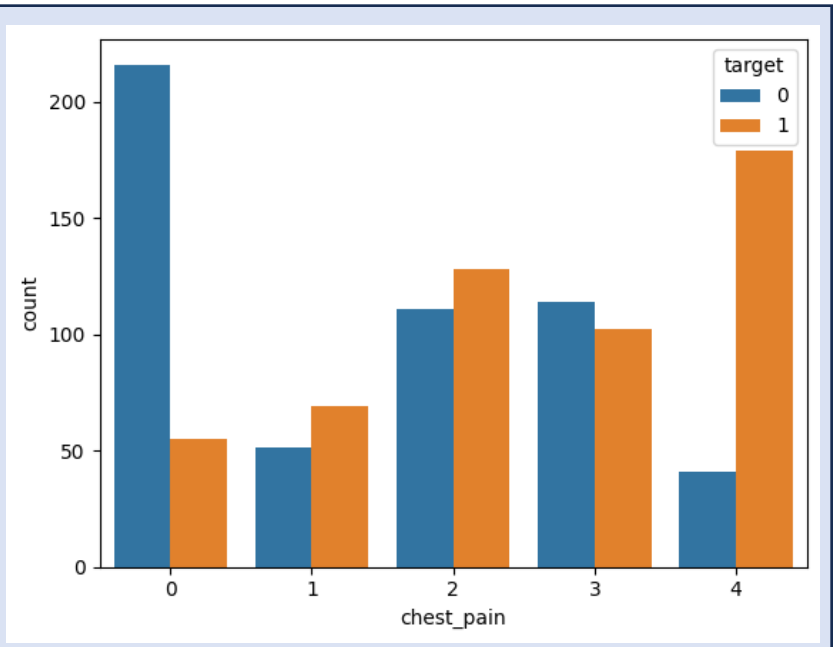

Figure 10

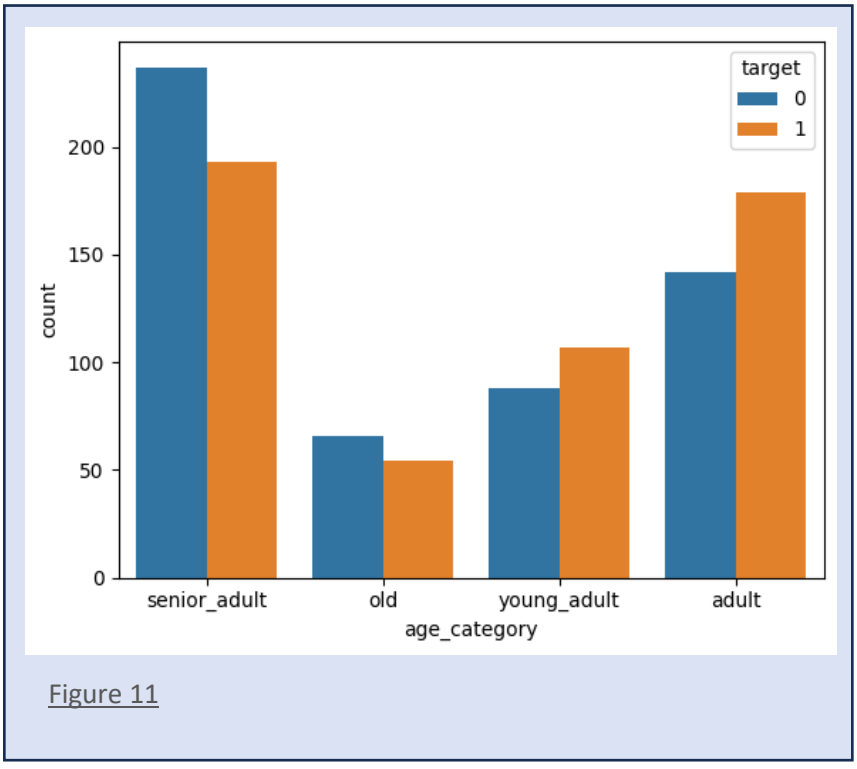

Figure 11

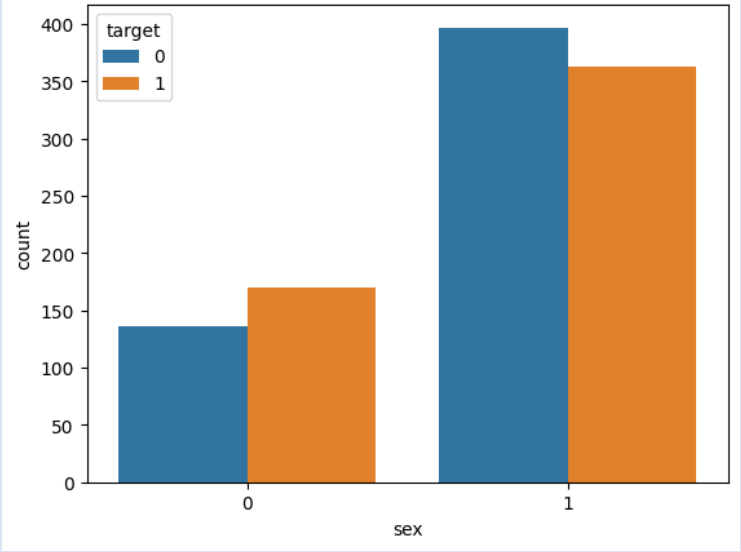

Figure 12

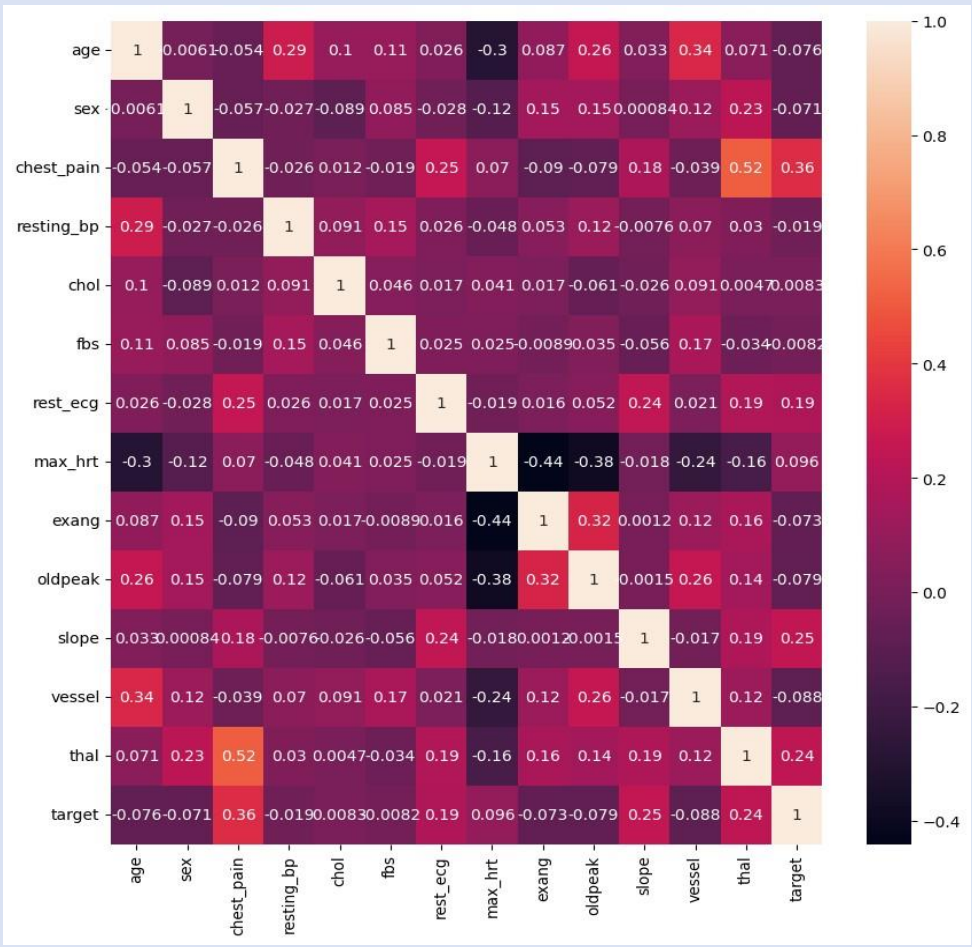

Figure 13

6.REFERENCES

1. Al Aref, SJ, Anchouche, K, Singh, G, Slomka, PJ, Kolli, KK, Kumar, A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. (2019) 40:1975–86. doi: 10.1093/eurheartj/ehy404 https://doi.org/10.1093/eurheartj/ehy404 site

2. 16. Juhola, M, Joutsijoki, H, Penttinen, K, and Aalto-Setälä, K. Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. *Sci Rep*. (2018) 8:1–10. doi: 10.1038/s41598-018-27695-5 https://doi.org/10.1038/s41598-018-27695-5 site

3. 17. Maheshwari, V, Mahmood, MR, Sravanthi, S, Arivazhagan, N, ParimalaGandhi, A, Srihari, K, et al. *Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control. J Nanomater*. (2021) 2021:1–8. doi: 10.1155/2021/3383146 https://doi.org/10.1155/2021/3383146 site

4. 18. Maini, E., Venkateswarlu, B., and Gupta, A. (2018). "Applying machine learning algorithms to develop a universal cardiovascular disease prediction system" in *International Conference on Intelligent Data Communication Technologies and Internet of Things*. Springer, Cham. 627–632. site

5. 19. Li, Q, Campan, A, Ren, A, and Eid, WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. *Int J Med Inform*. (2022) 163:104786. doi: 10.1016/j.ijmedinf.2022.104786https://doi.org/10.1016/j.ijmedinf.2022.104786

6. Cardiovascular Disease Prediction using Recursive Feature Elimination and Gradient Boosting Classification Techniques Prasannavenkatesan Theerthagiri* , Vidya J Department of Computer Science and Engineering, GITAM School of Technology, GITAM University Bengaluru, India. https://www.researchgate.net/publication/352479655_Cardiovascular_Disease_Prediction_using_Recursive_Feature_Elimination_and_Gradient_Boosting_Classification_Techniques