EasyChair Preprint

№ 10698

An Experiment on Various Classification Methods for Predicting Cardio Vascular Disease

Maya B Dhone, Swathi Voddi and Swarna Kamalam Vaddi

August 15, 2023

An Experiment on Various Classification Methods for Predicting Cardio Vascular Disease

[1]Maya B. Dhone
Assistant Professor, Department of IT, Maturi Venkata Subba Rao Engineering College, Telangana , India

[2]Swathi Voddi
Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Andhra Pradesh, India.

[3]Swarna Kamalam Vaddi
Assistant Professor, Department of IT, Maturi Venkata Subba Rao Engineering College, Telangana , India

Abstract:

The heart is regarded as one of the body's most significant and complex organs, as is generally accepted. Without it, most life forms will disappear. Because the heart's primary function is to pump blood to all of the body's organs and distribute it there, it plays a crucial role. Because of this, heart-related conditions are extremely delicate and require extreme caution. It is often said that prevention is better than cure. Most cardiac disorders tend to be identified after they have actually transpired. However, research has demonstrated that, with advance warning, approximately 90% of cardiovascular diseases can be prevented. We can determine the relationship between age, blood pressure, gender, and other variables from the findings of this study. Through their investigation, we will be able to gain a better understanding how these variables affect cardiovascular health. Along with other classification models, we simulate KNN, Gaussian Naive Bias, Decision Trees, Logistic Regression, SVM, Random Forests and others. Predicting whether a heart condition is present early on is critical. It provides medical foresight and aids in numerous potentially fatal situations. This paper provides an overview of the existing work as well as an insight into the existing algorithm.

Keywords:

Cardio Vascular Disease (CVD), Logistic Regression, Decision tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest.

I INTRODUCTION

At present according to WHO reports cardio vascular diseases accounted for 17.9 Million (32.1 percent) deaths, up from 12.3 million (25.8 percent) in 1990. Ninety percent of CVDs can be avoided. We will investigate numerous risk factors for heart diseases. The main goal of this study is to develop a model that can forecast the probability of cardiovascular disease in light of the numerous characteristics (hazard factors) that characterize the condition. Standard execution metrics, such as exactness, will be used to actualize and evaluate distinct AI arrangement methods [1, 2].

The sections of this paper are as follows:

- A thorough analysis of the existing research is provided in Section II.
- The resources and methods are thoroughly explained in Section III.
- Results and discussion come next,followed by a conclusion andrecommendations for upcoming work.
- The references for this paper are provided at the conclusion.

II LITERATURE SURVEY

Classification techniques in machine learning have been the subject of an investigation in a research paper. This paper's goal is to look into the results of several categorization techniques for a dataset on heart disease. Linear Regression, Random Forest, Decision tree, SVM algorithm are the classification algorithms tested in this work [3-7]. Research has looked at the sound and wiped-out viewpoints that cause heart disease in both men and women. The natural database associated with the UCI Cleveland dataset has been proposed to be rather close to the Predictive Apriori, and Tertius age computations [8–10]. To differentiate these segments, a computational information conduct entitled affiliation rule mining is employed. Multiple forms of heart disease, comprising coronary course illness, coronary heart disease, ischemic heart disease, heart disappointment, congenital heart

disease, and cardiovascular disease, has been used to base a survey of research on heart infection [10,11].

III MATERIALS AND METHODOLOGY

The general approach and procedure taken during the experimental study is shown using Figure1. This begins with data analysis and study, then preparation for heart disease prediction using various ML classification algorithms, and finally conclusion.

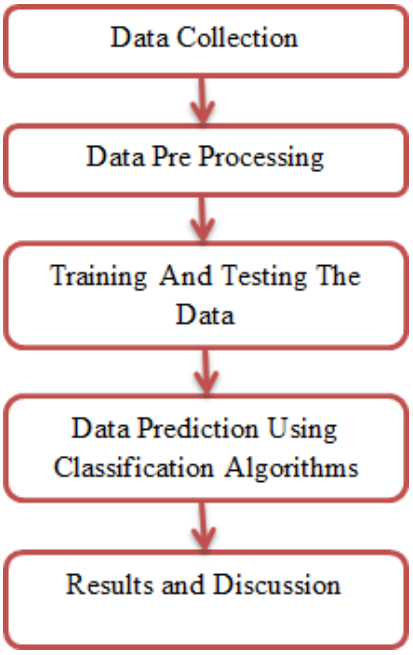

Figure 1: Principal components of the Planned Approach

A. Data Collection

In this research, we use the "Heart Disease Prediction Data Set" from the data archive of the University of California, Irvine. In this dataset there are 270 patients' data and there are 13 independent predictive variables or column attributes. There are 76 attributes in this database. However, upon closer examination, it can be determined that the most accurate outputs can be obtained by using by using a subset of 14 of them. These values will be applied when training our model using different techniques.

B. Information about the Data

The prior subsection made clear that, only 14features in this dataset are relevant to the study, and their specifics are listed in table I. Any dataset's feature variables can be divided into three categories: categorical, ordinal, and continuous features.

Therefore,

- Based on the dataset and its accompanying values, sex and chest discomfort can be categorized into qualitative characteristics.
- The ordinal aspects will include fasting blood sugar, electrocardiogram, provoked angina, slope, number of vessels, and diagnosis.
- Continual features include age, hypertension, and cholesterol levels in the blood, peak heart rate, and ST depression.

Fig. 2 provides a visualization of whether or not this dataset can predict CVD.

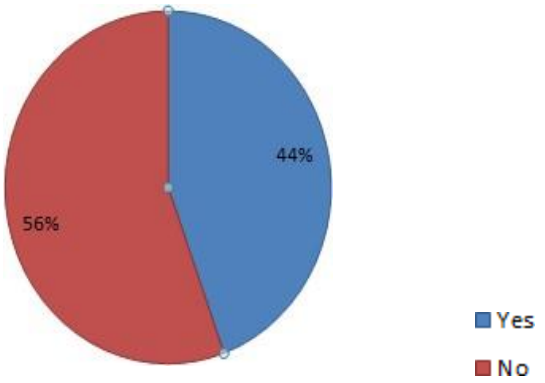

Figure2: Sorting of the dataset according to whether the disease has been predicted or not

From the dataset, we can predict the cardio vascular disease among the male and female patients. As the visualization of Figure2 shows the results of patients having heart disease percentage is shown as 44% and the patients having no heart disease percentage is shown as 56%. Thus thisdataset helps in predicting the heart disease by applying various classification algorithms. In addition to that Accuracy is also defined as a performance metrics for all the classification techniques in Machine learning to identify which algorithm achieves better accuracy.

TABLE I: FEATURE ATTRIBUTES AND THEIR POSSIBLE RESULT

Feature Attribute	**Outcome**
Age	Numeric value
Sex	Male-1 Female-0
Chest discomfort	Typical angina-1; Atypical angina-2; Non - anginal pain 3; Asymptomatic-4
BP	Blood pressure at rest (in mm Hg)
Cholesterol	Cholesterol in mg/dl
FBS over 120	During Fasting>120 mg/dlTrue-1; False-0
EKG results	While Resting Normal-0; HavingST-T-1; Hypertrophy-2
Peak Heart Rate	Numerical Heart Rate possible
Exercise angina	resulting from exercising Yes-1; No-0
ST depression	Value ranges from 0 to 6.2
Slope of ST	The ST segment's slope inclining upwards-1, downwards-3, and flat-2
Number ofbloodvessels	loccosopytal-derived coloring –(0-3);
Thallium	Normal-3; Fixed Defect- 6; Reversable Defect- 7
Heart Disease(Prediction)	Angiographic Status-0for < 50% Diameter Narrowing Angiographic Status-1for> 50%

The pair plot and heat map for the key important aspects of this dataset, shown in Figures 3 and 4, show the interdependence of all the variables on one another. The relations that can generate helpful predictions have been

illustrated here using the various graphical features supplied.

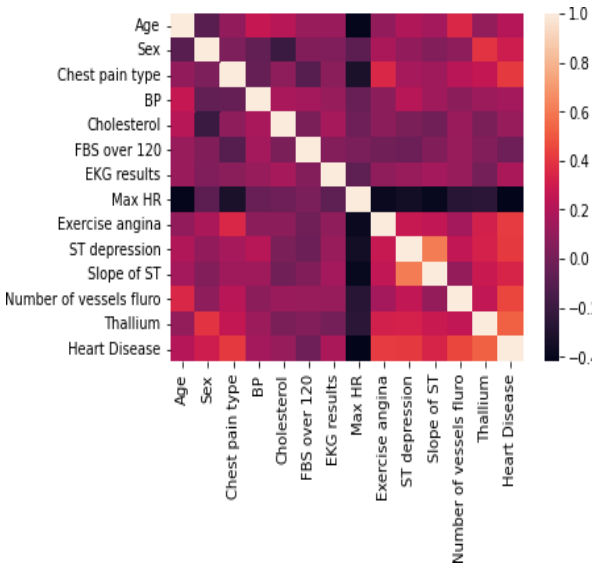

Figure 3: Visualization of Heat map for 14features mentioned in Table I

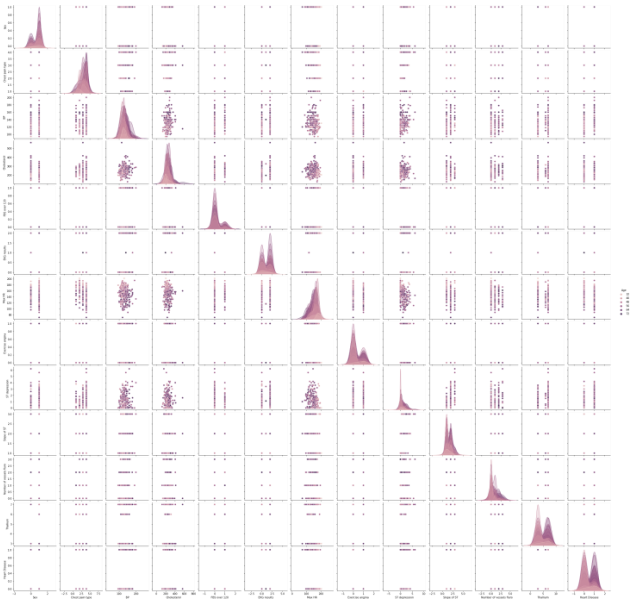

Figure 4: Visualization of Pairplot

We can observe that individuals without aheart disease typically have higher heart ratesand is shown in Figure 5.

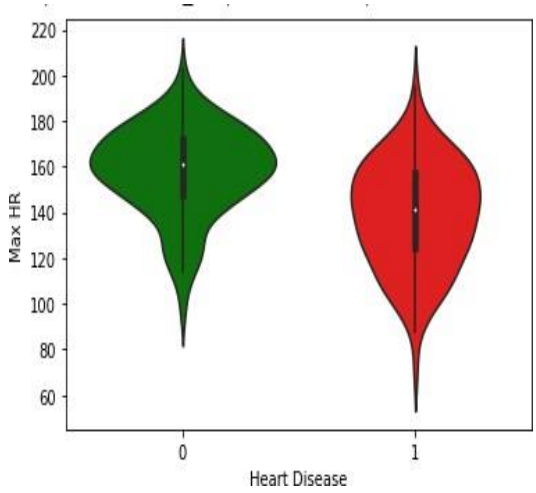

Figure 5: Relation between Heart Diseaseand Max Heart Rate (Max HR)

From Figure6, We can see that the majority of people between the ages of 50 and 65 are most likely to receive a heart disease.

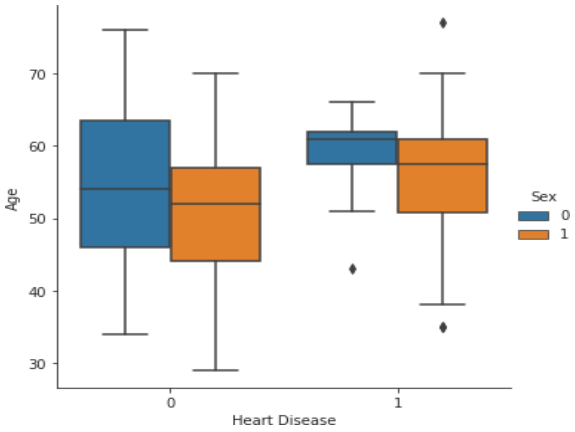

Figure 6: Relation between Heart Diseaseand Age

From Figure 7 We are able to determine that men are more prone to develop heart disease than women. Someone who has exercise-induced angina may have a risk of heart disease that is even three times higher.

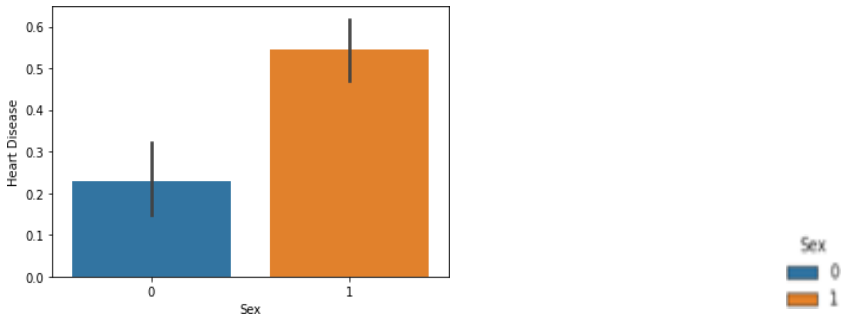

Figure 7: Relation between Sex and HeartDisease

C. Algorithms Employed

On the given dataset, we use a variety of classification algorithms, including KNN, Decision Trees, Logistic Regression, Navie Bayes, SVM, and Random Forests to determine whether a person has CVD based on Heart Related Features.

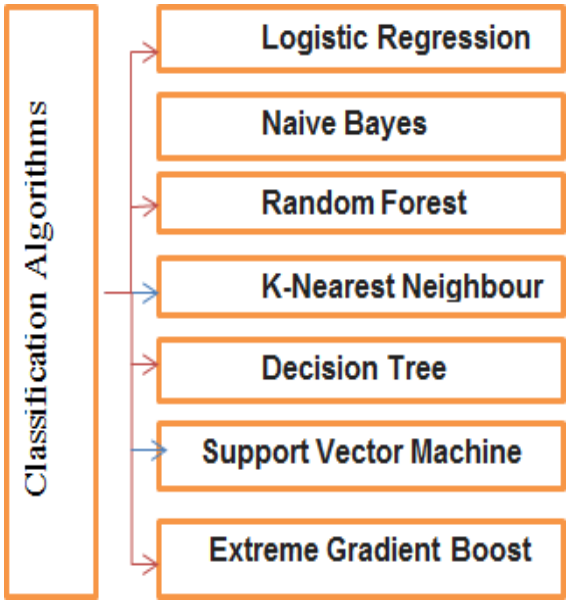

Figure 8: Classification algorithms Used

1. LOGISTIC REGRESSION:

To analyze data based on a number of independent variables, a statistical method known as the classification algorithm for logistic regression is used. It produces a single result by working with a variety of distinct dataset parameters. The dataset's outcome is measured in binary terms, such as true orfalse, or 0 or 1. Therefore, the tuple will be encoded as 1, if the given condition is met; otherwise, it will be encoded as 0.

2. NAIVE BAYES

Naive Bayes is an explanation algorithm for challenges involving double (two classes) and more types of structures. Utilizing input esteems that are parallel orcrystal clear makes the method easiest to understand. It is known as naïve Bayes or idiot Bayes because the probabilities for each hypothesis are condensed to make their calculating tractable. Given the goal-directed value rather than attempting to calculate the predictions of each quality worth P (d1, d2, d3|h), P(d1|h)*P(d2|H, etc.) is considered to be restrictively independent. This is a sound premise that is often improbable in real facts, like the notion that the attributes don't complement one another. However, even when this assumption is incorrect, the approach still performs reasonably well with the data.

3. RANDOM FOREST:

A technique for supervised classification is the random forest algorithm. As the name implies, this algorithm grows the forest by a significant amount. In general, a forest's strength correlates with its number of trees. Similar to the previous statement, we may claim that when using the random forest classifier, the accuracy of the outcomes rises as the numb er of trees in the forest does. Because they are mostly used to make decisions, these treesare also called decision trees.

4. KNN ALGORITHM:

The K Nearest Neighbor Algorithm, or KNN, works by forming groups of valuesthat are more similar to one another. The data are rapidly sorted into various categories relative to which value shares the most attributes with each group. So, the information is arranged to correspond to some recognized pattern. The classes are numbered 0, 1, 2, and so on, depending on the number of items that need to be categorized. For the objective of this experimental study, K=1 was chosen.

5. DECISION TREES:

Decision Trees, as the name signifies, are used to make decisions and have the shape of a tree. The aforementioned

method uses supervised machine learning, and a specific parameter signals that the data is permanently separated. The tree can be defined through its decision nodes and leaves. The leaves stand in for the decisions or outcomes at the end. At each node, a choice is made. In the present layout, a choice is made, leading to the final leaf.

6. SVM:

To address categorization and relapsing challenges, the "Support Vector Machine" (SVM) supervised machine learning approach could possibly be utilized. However, it is frequently utilized to classification-related problems. The estimation of each component is the estimation of a particular organization in the framework of SVM, and every dataset is plotted as a point in n-dimensional space (where n is the quantity of highlights you have). The hyper-plane that distinctly divides the two classes is then located using classification.

7. EXTREME GRADIENT BOOST

A subset of ensemble machine learning algorithms known as gradient boosting can be applied to classification and regression predictive modeling issues. Decision tree models are used to build ensembles. To correct the prediction errors made by previous models, trees are added to the ensemble one at a time and fitted. Boosting is a type of ensemble machine learning model like this.
The gradient descent optimization algorithm and any arbitrary differentiable loss function are used to fit models. Gradient boosting is the name given to the method because, similar to a neural network, the loss gradient is reduced as the model is fitted.

D. Methodology

a. First and foremost, we have chosen the dataset and researched the various parameters extensively.
b. In this study, we followed the usual steps for studying any data using machine learning: beginning with comprehension, cleaning, visualizing, and modeling.
c. We begin by acquiring the data,understanding it, and then cleaning it.
d. The dataset is then reviewed in order to identify patterns, goals, and numerical values. Multiple relationships between numerous elements have been found.
e. Then, in order to make our dataset compatible with Sci-Kit, we address all of the missing data. Learn how to use library machine learning algorithms.
f. Finally, to find the classification algorithm that offers the highest level of accuracy for this dataset, we employed machine learning to model the different classification algorithms and train them on the dataset.

E. Software Employed

After applying each of the fore mentioned algorithms to the provided dataset, accuracy for each algorithm is determined. The platform of Python has verified the accuracy of all of these machine learning algorithms. Jupyter Notebook, supported by Python 3.6.5, was utilized in this study. A tool called Scikit Learn is used for data mining and analysis. Scikit learn was used to put the algorithms into action. Matplotlib version 3.1.0 was used to create the graphs.

IV.RESULT DISCUSSION

Table II shows the ccuracy of all classification algorithms.

Algorithms Used	Accuracy (%)
Logistic Regression	77.77
Navie Bayes	74.07
Random Forest	85.18
KNN Classifier	81.48
Decision Tree	75.92
Support Vector Machine	74.07
Extreme Gradient Boost	72.22

Table II: Accuracy of Classification Algorithms

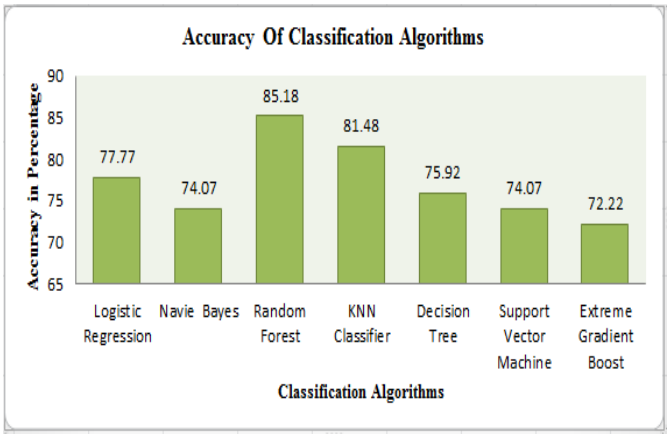

Figure 9: Visualization of Accuracy ofAlgorithms

V CONCLUSION

The objectives of this research set out to identify the best algorithm for categorizing the variables that affect whether or not heart disease is diagnosed. Newer approaches to preventing CVDs before they occur will benefit from this study. Table II shows the various algorithms accuracy rates, which have beendetermined.

Based on the accuracy scores of the algorithms employed in this experiment, we can say that Random Forest Algorithms are the most efficient algorithm. A precise algorithm for devices that can determine whether an individual is susceptible to cardiovascular disease (CVD) and, if so, at what point in life can be created using the trials and tests from this research. Since the majority of cardiovascular diseases can be avoided if caught early, this will be extremely beneficial.

REFERENCES

[1] National Health Council,'Heart HealthScreenings',2017.[Online]Avai lable:http://www.heart.org/HEARTO RG/Conditions/HeartHealthScreenings_UCM_428687_Article.jsp#.WnsOAeeYPIV

[2] Ahmed Fawzi Otoom, Emad E. Abdallah , Yousef Kilani ,Ahmed Kefaye and Mohammad Ashour(2015)Effective Diagnosis and Monitoring of Heart Disease ISSN: 1738-9984 IJSEIA

[3] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A HeartDisease Prediction Model usingSVM-Decision Trees- Logistic Regression (SDL)", InternationalJournalof Computer Applications, vol. 68, 16 April 2013.

[4] Machine learning based decision support systems (DSS) for heartdisease diagnosis:a review. Online: 25 March 2017 DOI: 10.1007/s10462-01.

[5]M.A.Jabbar,"Heart DiseasePrediction System using Associative Classification and Genetic Algorithm", ICECIT, pp183-192, Elsevier, vol 1(2012)

[6] AliveKor,[Online]Available:https://www.alivecor.com/how-it- works.

[7] Prerana T H M, Shivaprakash N Cet al "Prediction of Heart Disease Using Machine Learning Algorithms- Naive Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS",Vo l3, PP: 90- 99©IJSE, 2015

[8] Nahar, Jesmin, et al., "Associationrule mining to detect factors whichcontribute to heart disease in males and females", Expert Systems with Applications, Vol. 40, No. 4, pp.1086-1093, 2013.

[9] Gayathri, P., and N. Jaisankar, "Comprehensive study of heart disease diagnosis using data mining and soft computing techniques", 2013.

[10] Vijiyarani, S., and S. Sudha, "Anefficient classification tree techniquefor heart disease prediction",International Conference on Research Trends in Computer Technologies(ICRTCT-2013) Proceedingspublished in International Journal ofComputer Applications (IJCA) (0975– 8887). Vol. 201, 2013.