# PREDICTION OF HEART DISEASE USING MULTIPLE MACHINE LEARNING ALGORITHMS

**Saniya Malkan Ahmed [1], Asha Nair [2]**

Student, Coventry University, Coventry, United Kingdom [1][2]
13080547 [1] 12788754[2] [1] sanibaniahmed@gmail.com[1] ashanair9830@gmail.com[2]

***Abstract – This paper provides and discusses the results of applying different Machine Learning techniques to a heart diseases dataset. One of the key goals on this dataset is to make a prediction about whether a patient has heart disease or not based on the provided features, and another is an experimental task to identify and derive various insights from this dataset that may aid in better understanding the issue. In this paper three classification techniques are implemented and compared: a Decision tree, Gradient Booster and Random Forest machine algorithm. By presenting and discussing the expected classification accuracy on testing data, which is represented using confusion matrices, the performance of the classification techniques is compared.***

***Keywords – Support vector machine, Random Forest, Logistic Regression, Decision Tree***

## I. INTRODUCTION

UCI is a public research university located in Irvine, California, USA. It was founded in 1965 and is known for its strong programs in engineering, computer science, biology, and social sciences. UCI is a member of the Association of American Universities, a prestigious group of research-intensive institutions.

Heart disease refers to a range of conditions that affect the heart, such as coronary artery disease, heart attacks, and heart failure. It is a leading cause of death worldwide and can be caused by a variety of factors, including lifestyle factors like poor diet and lack of exercise, as well as genetic factors. Treatment for heart disease can vary depending on the specific condition and may include medication, surgery, and lifestyle changes.

According to WHO statistics, cardiovascular illnesses are the leading cause of death worldwide. Every year, it causes the deaths of seventeen million individuals, mostly from heart disease. Prevention is preferable to treatment. In other words, not only patients but also everyone may take action early to ward off sickness if we can assess the risk of every patient who is likely to have heart disease.

This dataset contains significant patient characteristics and is actual data. The model can then determine the weights of each characteristic, allowing us to quickly determine which feature is more important than others.

A standard method to assess the model's accuracy is to use a confusion matrix. From a medical perspective, recall rates are more significant than accuracy rates since nobody wants to be given the incorrect diagnosis when they genuinely have heart disease. Therefore, we shall assess the recall efficiency. Then, if the model is sufficient, we may explore the features while still evaluating it using the ROC curve.

## II. PROBLEM & THE DATASET

With the aid of a number of variables, including age, gender, blood pressure, and others, we will attempt to forecast a person's likelihood of getting heart disease in this kernel. For the aim of prediction, we'll employ a variety of classification models. We will go through a number of procedures, including model fitting, data cleansing, EDA, and evaluation. In the healthcare sector, this prediction is particularly helpful since it enables the doctors to provide the patient with early care.

This dataset is multivariate, which refers to multivariate numerical data analysis that provides or incorporates a number of distinct mathematical or statistical variables. Age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate reached, exercise-induced angina, oldpeak — ST depression brought on by exercise in comparison to rest, slope of the peak exercise ST segment, number of major vessels, and Thalassemia are the 14 attributes that make up this score. There are 16 attributes in this database, however only a subset of 14 of them are used in all published investigations.

TABLE-1: DATA SET AND DATA TYPES

| Attribute | Description | Data Type |
|---|---|---|
| Age | The person's age in years | float64 |
| Sex | The person's sex (1 = male, 0 = female) | float64 |
| CP | The chest pain experienced (Value 1: typical, 2: atypical,3: non-anginal pain, 4: asymptomatic) | float64 |
| TRESTBPS | The person's resting blood pressure (mm Hg on admission to the hospital) | float64 |
| Chol | The person's cholesterol measurement in mg/dl | float64 |
| FBS | The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false) | float64 |
| Restecg | Resting electrocardiographic measurement (0 = normal, 1 = abnormality, 2 = showing probable a) | float64 |
| Thalach | The person's maximum heart rate achieved | float64 |
| Exang | Exercise induced angina (1 = yes; 0 = no) | float64 |
| Oldpeak | ST depression induced by exercise relative to rest | float64 |
| Slope | the slope of the peak exercise ST segment (Value 1: upsloping, 2: flat, 3: downsloping) | float64 |
| Ca | The number of major vessels (0-3) | Object |
| Thal | A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect) | Object |
| Target | Heart disease (0 = no, 1 = yes) | Int64 |

Researchers studying machine learning have only ever used the Cleveland database. One of the main tasks on this dataset is to predict, based on the patient's given attributes, whether or not that patient has heart disease. Another major task on this dataset is the experimental task, which is to identify and discover various insights from this dataset that may help in better understanding the issue.

## III. CLASSIFICATION TECHNIQUES

### A. DECISION TREE

Typically, a tree has roots, branches, and leaves. Decision Tree adheres to the same format. Along with branches and leaf nodes, it has a root node. Testing an attribute occurs on each internal node, the test's result appears on the branch, and the class label as a result appears on the leaf node. As its name implies, a root node is the topmost node in a tree and is the parent of all other nodes.

Each node in a decision tree reflects an attribute, every relation indicates a decision, so each leaf represents the conclusion. One of the potent techniques frequently employed in a variety of domains, including machine learning, image processing, and pattern recognition, are decision trees. A set of fundamental tests are successfully and cogently combined by the decision tree, where each test compares a numerical property to a threshold value.

### B. LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to predict the outcome of a binary dependent variable based on one or more independent variables. It is a type of generalized linear model that uses a logistic function to model the probability of a binary outcome. The logistic function takes the form of the S-shaped curve, which allows the model to predict the probability that an example belongs to a certain class. Logistic regression is used in a wide range of applications, including medical diagnosis, credit risk assessment, and natural language processing.

### C. RANDOM FOREST

Leo Breiman created the Random Forest, which is a collection of unpruned classification or regression trees created from a random sample of training data. The features chosen during the induction procedure are at random. The ensemble's predictions are combined (by majority vote for classification, or by average for regression) to produce the prediction. Generally speaking, Random Forest performs significantly better than a single tree classifier. Using a set of potential trees with K random features at each node, a random tree is one that is created at random. In this context, "at random" means that there is an equal probability of sampling each tree in the set of trees. The distribution of trees might also be described as "uniform." It is efficient to produce random trees, and combining several such random trees typically results in models that are correct. In the field of machine learning, there has been a great deal of research done recently on random trees.

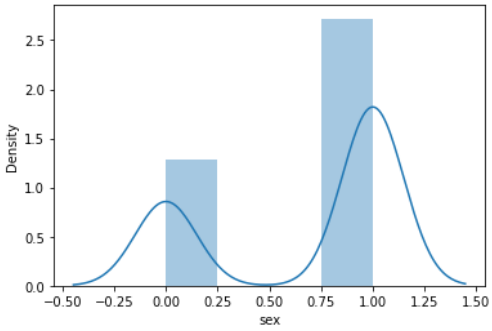
### D. SUPPORT VECTOR MACHINE

Support vector machine (SVM) is a type of supervised learning algorithm that can be used for both classification and regression tasks. It is based on the idea of finding a hyperplane in an N-dimensional space that maximally separates the two classes. The data points closest to the hyperplane, known as support vectors, are the most influential in defining the position of the hyperplane. SVM models are effective in high-dimensional spaces and can be used to classify complex, non-linear relationships. They are also well-suited to problems with a large number of features, such as text classification and image recognition.

## IV. EXPERIMENT RESULTS

The various activities are categorized using the prior outlined classification approaches. The 14 separate activities are categorized using the classification methods previously outlined. There are a total of 920 instances that can be used. For the heart disease

dataset, we started with the exploratory analysis, in EDA first we dropped the ID column since it is not of more importance. Then we separated the numerical and categorical values and plotted using bar charts for categorical values and Scatter plots for




Numerical Values as univariate analysis.

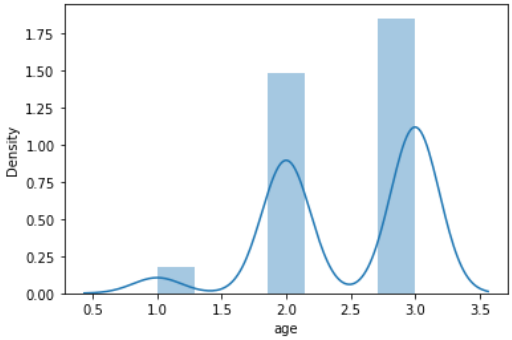



*Figure-1 Categorical values of Density with Sex*

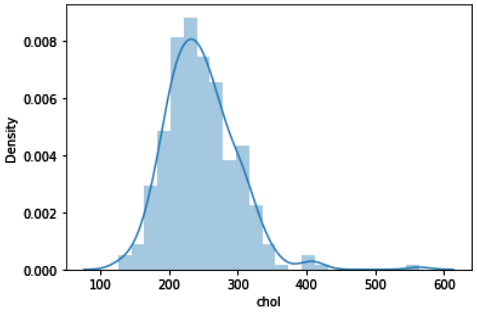



*Figure-2 Categorical values of Density with Age*

*Figure-3 Categorical values of Density with cholesterol*

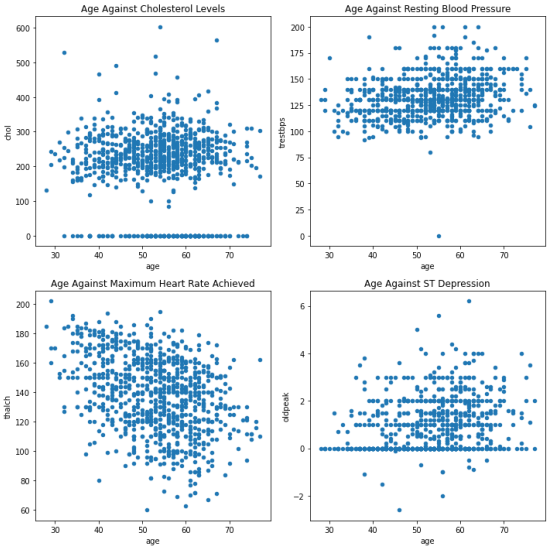



*Figure-4 Scatter plots.*

Then in the next step we checked the data description by analysing data and calculated the average cholesterol level based on target variable and chest Pain type, average cholesterol level based on target variable and Patient Gender and average cholesterol level based on target variable and Cardiographic Results which is represented in the figure.

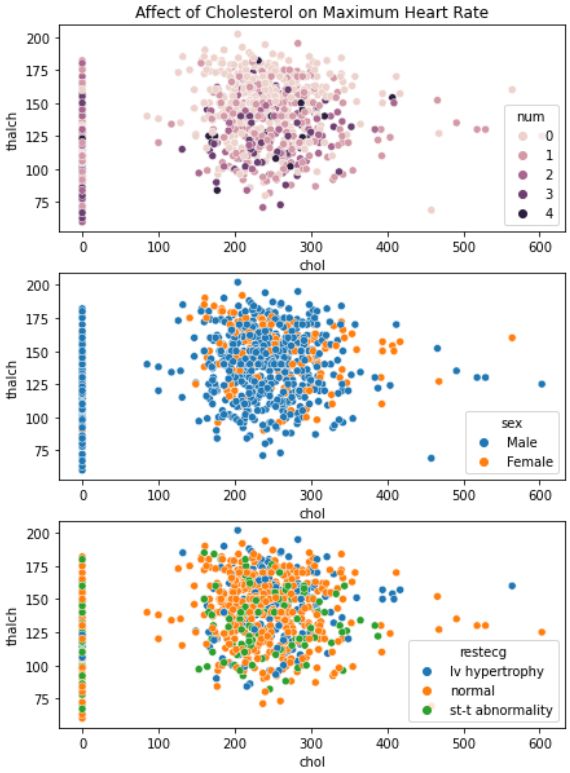



*Figure-5 Affect of Cholesterol on Maximum Heart Rate*

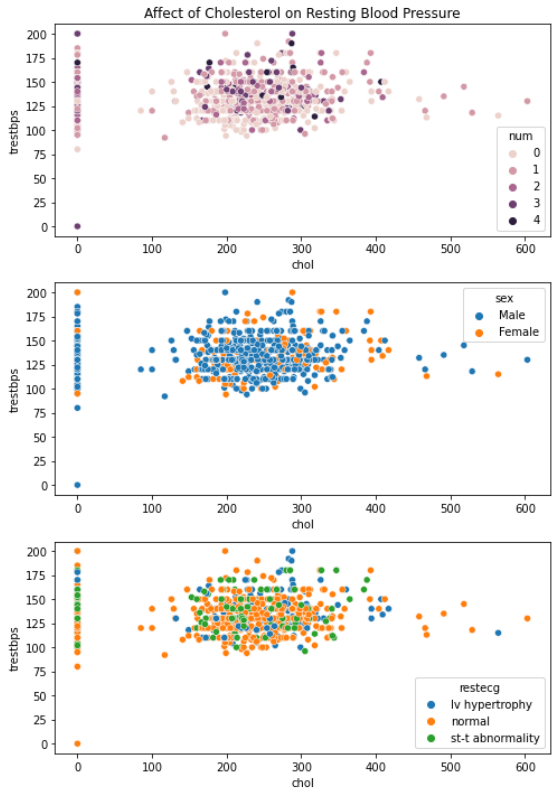



*Figure -6 Affect of Cholesterol on Resting Blood Pressure*

In the next step we created the heat map for our heart disease dataset by creating correlation matrices.

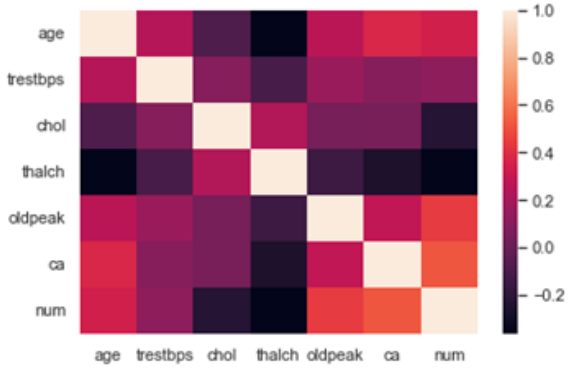



*Figure-7 Correlation Matrix*

In the next step we were checking whether there is any outliers present in the dataset, in the fig() we plotted it using boxplots. In this dataset we found out outliers but removing them doesn't mean sense in our dataset since this the medical dataset we choose there will be variation the cholesterol level, chest pain or the blood pressure level of a patient. Next is to check for the null values that are available in the dataset, we found null values in Cholesterol level, Maximum heartbeat, Resting blood pressure and old peak attributes. In order to remove those null values, we performed imputation method for this attribute, we made use of mean and median imputations.

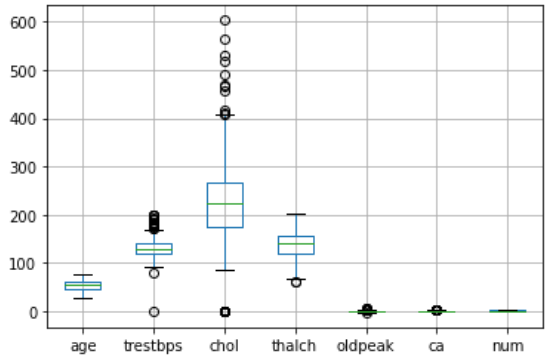



*Figure-8 Box-Plot for outliers*

Once all the data pre-processing steps are done step is to split the data, we did Principal component Analysis (PCA), which is explained in the figure () and ()

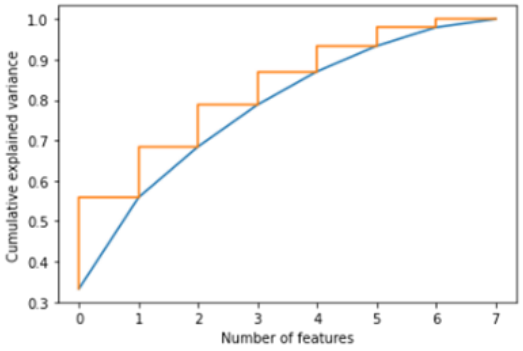



*Figure-9 Cumulative explained variance with No. of features*

The next step is to create classification algorithms for heart disease data in our data we made use of Random Forest, Logistic Regression, Support Vector machine and Decision tree., among all these Support Vector Machine (SVM) gave us better accuracy than any other machine learning classification algorithm.

## V. EXPERIMENT RESULTS

The original data had 920 features, hence the data needed dimensionality reduction to decrease the computational complexity of building models using it. Therefore, once the data was standardized, Principal Component Analysis (PCA) was used for dimensionality reduction. Fig 6 shows the cumulative explained variance of the dataset for the first principal components retrospectively. Fig 7 shows the scatter plot for the first and second principal content.

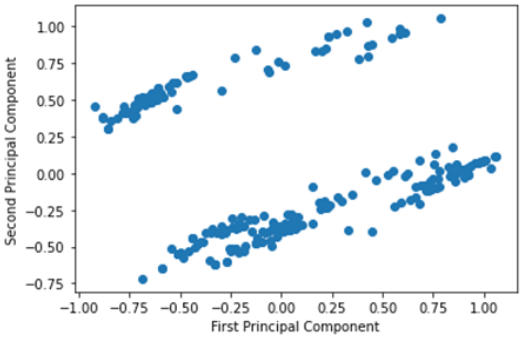



*Figure-10 Second with first Principal components*

This is the outcome that was achieved by using a variety of machine learning techniques, such as logistic Regression, KNN, Gradient Boosting, SVC, Random Forest Classifier, Decision Tree, approaches to the dataset. For our data we made use of Support vector machine, Logistic Regression, Random Forest, and Decision tree which gave the accuracy of 0.83, 0.81, 0.75, 0.77 respectively. Where in SVM has the better performance when compared to all other algorithms. Here, SVM has the maximum accuracy, where the model will predict the test data and give the score of that data which is our accuracy percentage. We created a confusion matrix which shows the actual and the predicted value of the heart disease dataset which is shown in the figure ()

## VI. DISCUSSION AND CONCLUSIONS

Either the SVM or Logistic Regression is probably the best model for this particular categorization issue. The SVM (linear) model has a somewhat greater percentage of correctly classifying objects, but training it takes a lot longer than for the other models. In order to analyse the data, we made use of mainly Support vector machine, Logistic Regression, Random Forest and Decision Tree among all these we found out SVM and Logistic gives better accuracy when compared to Random Forest and Decision tree, but while comparing SVM and Logistic we found out SVM is even better when compared to Logistic Regression model. Decision trees are prone to overfitting, so if there isn't a lot of pruning done, it should be the least preferred model. However, it is crucial to remember that by using PCA to apply such a considerable dimensionality reduction, the risk of overfitting would have been greatly diminished.

TABLE-2 : TECHNIQUE AND ACCURACY

| Technique | Accuracy (.) | Accuracy (%) |
|---|---|---|
| Support Vector Machine | 0.83 | 83% |
| Random Forest Classifier | 0.75 | 75% |
| Decision Tree Classifier | 0.77 | 77% |
| Logistic Regression | 0.81 | 81% |

As we can see in the table (1) the Support Vector machine is the better performing algorithm when compared to all other algorithms which gave the accuracy of 83%. In the fig () explains the confusion matrix of the predicted and the actual results of the algorithms we made use of.

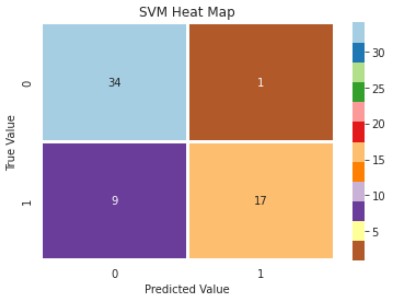



*Figure-11 Confusion matrix of SVM*

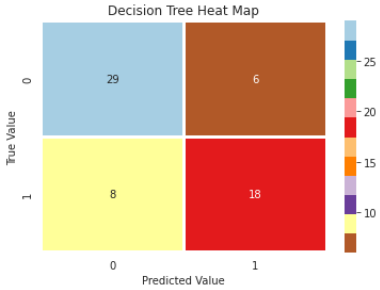



*Figure-12 Confusion matrix of Decision Tree*

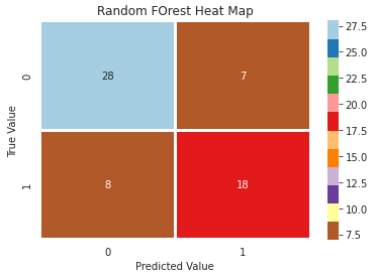



*Figure-13 Confusion matrix of Random Forest*

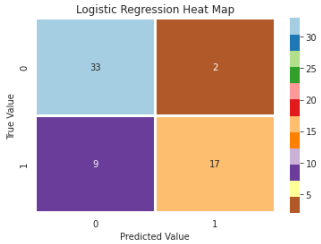



*Figure-14 Confusion matrix of Logistic Regression*

The use of multiple Machine Learning classification approaches on the same data set has been addressed and compared in this research. By comparing the correct classification rates of different strategies and the length of time needed to train and test the model, we assessed the effectiveness of these techniques when applied to this data set. This study demonstrates that machine learning approaches can categorise heart disease datasets successfully even though the different categories are equivalent to standing in diverse settings. Due to the vast array of potential applications, there are many different research avenues that can be investigated. The information obtained might be compared with a sizable collection of training data to determine whether the person will eventually develop heart disease or not. However, access to such information is still only available to individuals who are prepared to consent to the use and analysis of their personal information.

## VII. APPENDIX

The following Google Drive link includes the original data as well as the Python code used to pre-process the data and apply the categorization techniques:
https://colab.research.google.com/drive/1CP3tNOhrx23B6Kf9rfkZxYDy_e6wg7mr#scrollTo=ecEce9oRsP99

## VIII. REFERENCE


- Sony, M.R.K. (2020) UCI Heart Disease Data, Kaggle. Available at: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data Applying various machine learning algorithms, the collected normal dataset is utilized directly for classification. (Accessed: December 14, 2022)
- *(PDF) random forests and decision trees - researchgate* (no date). Available at: https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees (Accessed: December 15, 2022).
- *Classification based on Decision Tree Algorithm for Machine Learning* (no date). Available at: https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning (Accessed: December 15, 2022).
- *Study and analysis of decision tree based classification algorithms* (no date). Available at: https://www.researchgate.net/publication/330138092_Study_and_Analysis_of_Decision_Tree_Based_Classification_Algorithms (Accessed: December 15, 2022).
- *Support Vector Machines: Theory and applications - researchgate* (no date). Available at: https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications (Accessed: December 15, 2022).