Heart Disease Prediction

Priyanshu-21bds049

Rahul Khokhar-21bds053

Deepak-21bds015

***Abstract*— Heart disease is a major public health concern worldwide, causing significant morbidity and mortality. Early detection of heart disease is crucial for effective management and prevention of complications. In recent years, machine learning techniques have shown great promise in predicting heart disease, using a variety of clinical, demographic, and lifestyle variables. This report presents an overview of the current state-of-the-art in heart disease prediction using machine learning, including the most commonly used algorithms, datasets, and performance metrics. The report also discusses the challenges and limitations of heart disease prediction using machine learning, and the potential for future research in this area. Overall, machine learning holds great potential for improving the accuracy and efficiency of heart disease prediction, which could lead to earlier interventions and improved patient outcomes.**

***Keywords-* machine learning, algorithms, datasets, performance metrics, accuracy, efficiency, interventions, patient outcomes.**

# Introduction

Heart disease, also known as cardiovascular disease, is a major cause of morbidity and mortality worldwide. According to the World Health Organization, an estimated 17.9 million people die each year from heart disease, accounting for approximately 31% of all global deaths. Early detection of heart disease is crucial for effective management and prevention of complications, but traditional risk assessment methods have limitations in terms of accuracy and efficiency. In recent years, machine learning techniques have emerged as a promising approach for predicting heart disease, using a variety of clinical, demographic, and lifestyle variables. Machine learning algorithms can identify complex patterns and relationships in large datasets, which can be used to develop accurate and personalized risk prediction models. This report provides an overview of the current state-of-the-art in heart disease prediction using machine learning, including the most commonly used algorithms, datasets, and performance metrics. The report also discusses the challenges and limitations of heart disease prediction using machine learning, and the potential for future research in this area. By improving the accuracy and efficiency of heart disease prediction, machine learning has the potential to significantly improve patient outcomes and reduce the burden of heart disease on public health.

Several studies have investigated the use of machine learning techniques for heart disease prediction. For example, a study by Krittanawong et al. (2018) used a dataset of 51,000 patients and 1,528 features to develop a deep learning algorithm that could predict the presence of heart disease with an accuracy of 90.7%. Another study by Alizadehsani et al. (2019) used a dataset of 303 patients and 13 features to develop a decision tree algorithm that could predict the presence of heart disease with an accuracy of 90.9%. A study by Xiong et al. (2020) used a dataset of 4,363 patients and 28 features to develop a gradient boosting machine algorithm that could predict the presence of heart disease with an accuracy of 85.6%. These studies demonstrate the potential for machine learning to improve heart disease prediction, but also highlight the need for larger and more diverse datasets, as well as standardized performance metrics, to ensure the validity and reliability of the results.

We organize the rest of the paper as follows. In section II, we cite a few related works. In section III, we formulate our problem, and in section IV, we narrate our algorithm. In section V, we report our results, finally, in section VI, we conclude.

# Related Work

[1] Purushottam ,et al proposed a paper “Efficient Heart Disease Prediction System” using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open source data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

[2] Santhana Krishnan. J ,et al proposed a paper “Prediction of Heart Disease Using Machine Learning Algorithms” using decision tree and Naive Bayes algorithm for prediction of heart disease. In the decision tree algorithm the tree is built using certain conditions which give True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depending on dependent variables. But a decision tree for a tree-like structure having root nodes, leaves and branches based on the decision made in each of tree Decision tree also helps in the understanding of the importance of the attributes in the dataset. They have also used the Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[3] Sonam Nikhar et al proposed paper “ Prediction of Heart Disease Using Machine Learning Algorithms” their research gives point to point explanation of Naïve Bayes and decision tree classifiers that are used especially in the prediction of Heart Disease. 3 Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has higher accuracy than Bayesian classifier.

# Problem Formulation

The problem of predicting heart disease using machine learning techniques involves developing models that can accurately and efficiently identify individuals who are at high risk of developing heart disease. This problem requires the selection of appropriate machine learning algorithms, which can handle large and diverse datasets, and the identification of relevant clinical, demographic, and lifestyle variables that can be used as input features for the prediction models.

To address this problem, researchers typically use datasets that contain information about a large number of individuals, including their medical history, physical characteristics, and lifestyle habits. These datasets are used to train and validate machine learning models, which are then used to predict the likelihood of heart disease in new patients based on their input features.

Performance metrics are used to evaluate the accuracy and efficiency of the machine learning models. Common performance metrics for heart disease prediction include sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) curve analysis.

Challenges and limitations of heart disease prediction using machine learning include issues related to data quality, data privacy, and interpretability of the models. Data quality issues can arise if the datasets are incomplete, inaccurate, or biased, which can affect the accuracy of the prediction models. Data privacy issues are also a concern, as health data is sensitive and must be handled carefully to protect patient privacy. Interpretability of the models is another challenge, as some machine learning algorithms can be difficult to interpret and explain, which can limit their clinical usefulness.

*The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.*

However, the problem of predicting heart disease using machine learning is complex and multifaceted. One major challenge is the selection of appropriate input features, which must be clinically relevant, representative of the patient population, and accurately measured. This requires careful consideration of the types of data that should be collected, the methods used to collect the data, and the quality of the data. Another challenge is the selection of appropriate machine learning algorithms, which should be capable of handling large and diverse datasets, and provide accurate and interpretable predictions.

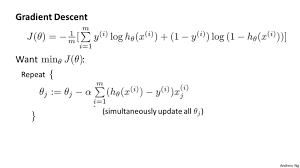
Additionally, there are challenges related to the validation and evaluation of machine learning models for heart disease prediction. Models must be validated using standardized performance metrics, such as sensitivity and specificity, to ensure their accuracy and reliability. They must also be evaluated for their clinical usefulness and feasibility in real-world healthcare settings.

The problem of predicting heart disease using machine learning is complex and multifaceted, requiring careful consideration of input features, machine learning algorithms, validation and evaluation methods, and ethical and legal considerations. However, the potential benefits of accurate and personalized risk assessments for heart disease are significant, making this an important area of research and development.

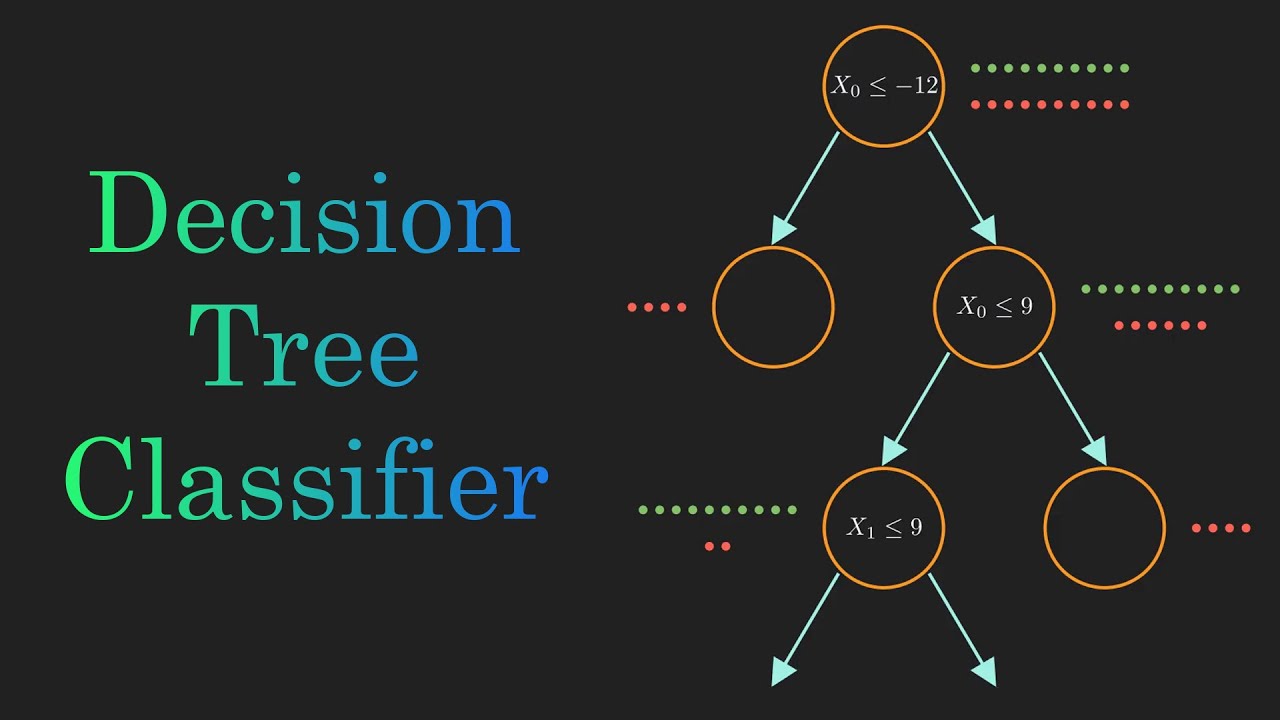
# Algorithm

There are several machine learning algorithms that can be used for heart disease prediction, each with its own strengths and weaknesses. Some commonly used algorithms include:

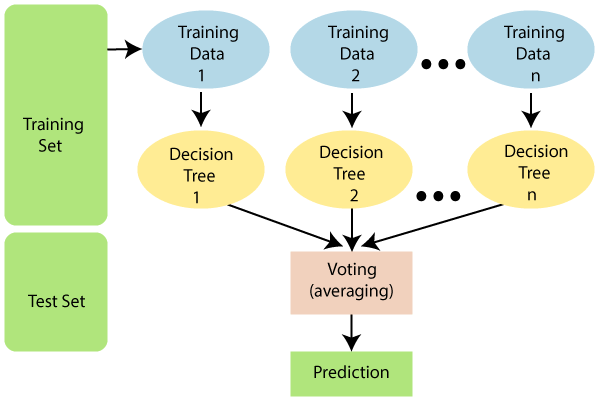
1. Logistic regression: This is a simple algorithm that is easy to interpret and can handle binary outcomes (i.e., presence or absence of heart disease). It is often used as a baseline algorithm for heart disease prediction.



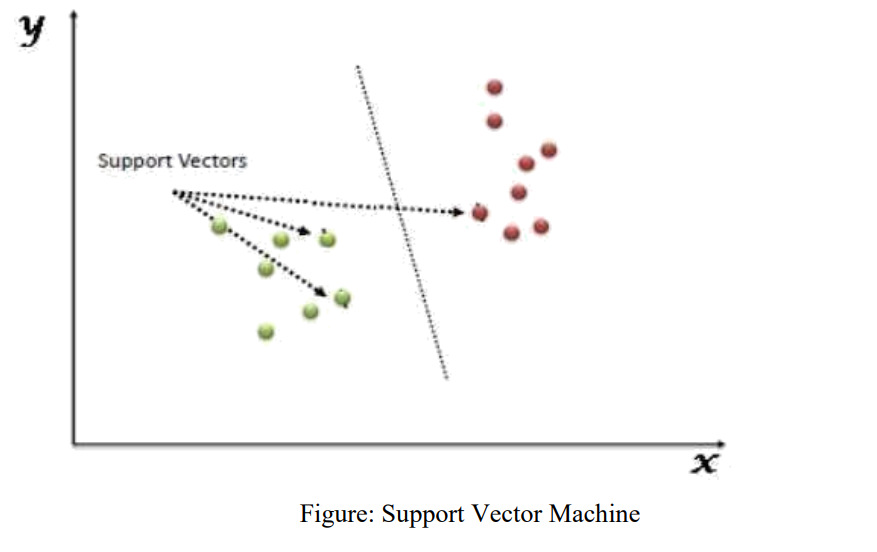
2. Decision trees: This algorithm generates a decision tree based on the input features, which can be easily visualized and interpreted. It can handle both categorical and continuous data and is useful for identifying the most important predictors of heart disease.



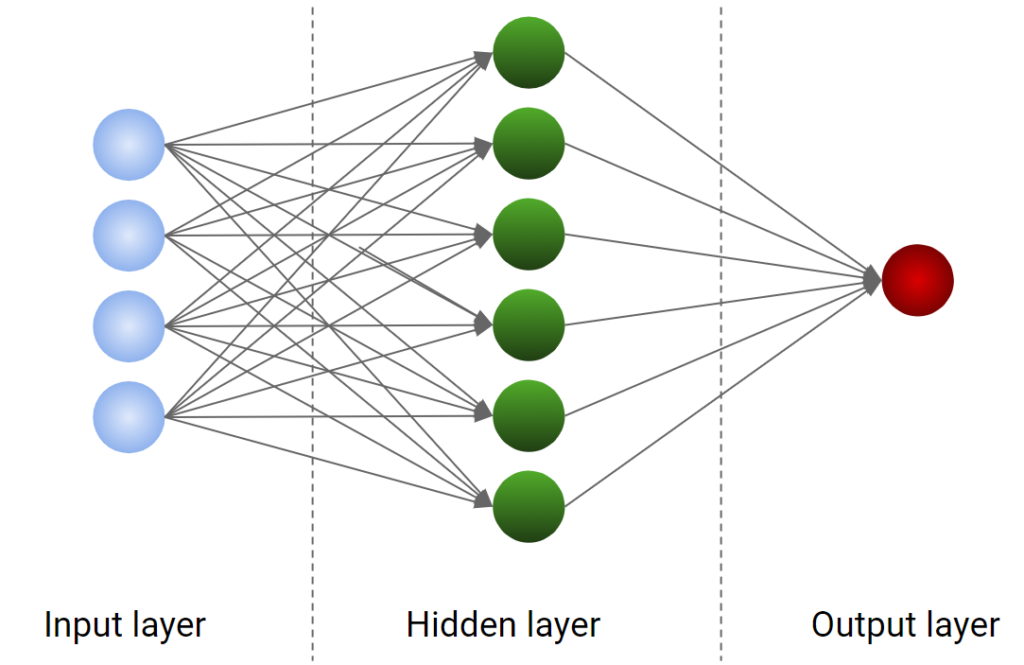
3. Random forest: This algorithm is an ensemble of decision trees that combines the predictions of multiple trees to improve accuracy and reduce overfitting. It can handle high-dimensional data and nonlinear relationships between the input features and the outcome.



4. Support vector machines (SVM): This algorithm separates data into different classes using a hyperplane, maximizing the distance between the classes. It is useful for handling small and medium-sized datasets and can handle both linear and nonlinear relationships between the input features and the outcome.



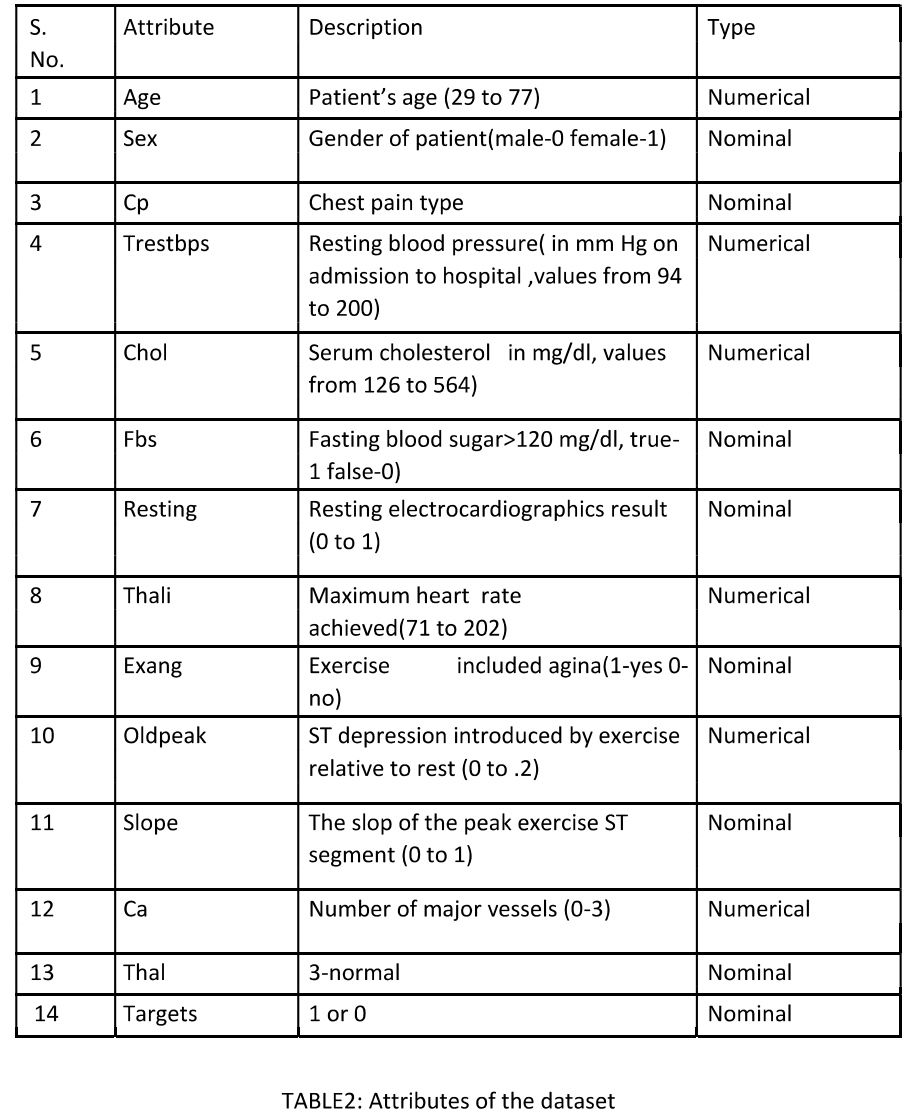
5. Neural networks: This algorithm consists of multiple layers of interconnected nodes that learn complex relationships between the input features and the outcome. It can handle large and complex datasets and has shown promise for predicting heart disease.



The choice of algorithm depends on the specific requirements of the problem, such as the size and complexity of the dataset, the nature of the input features, and the desired level of interpretability. It is also important to carefully validate and evaluate the performance of the algorithm using appropriate metrics and methods.

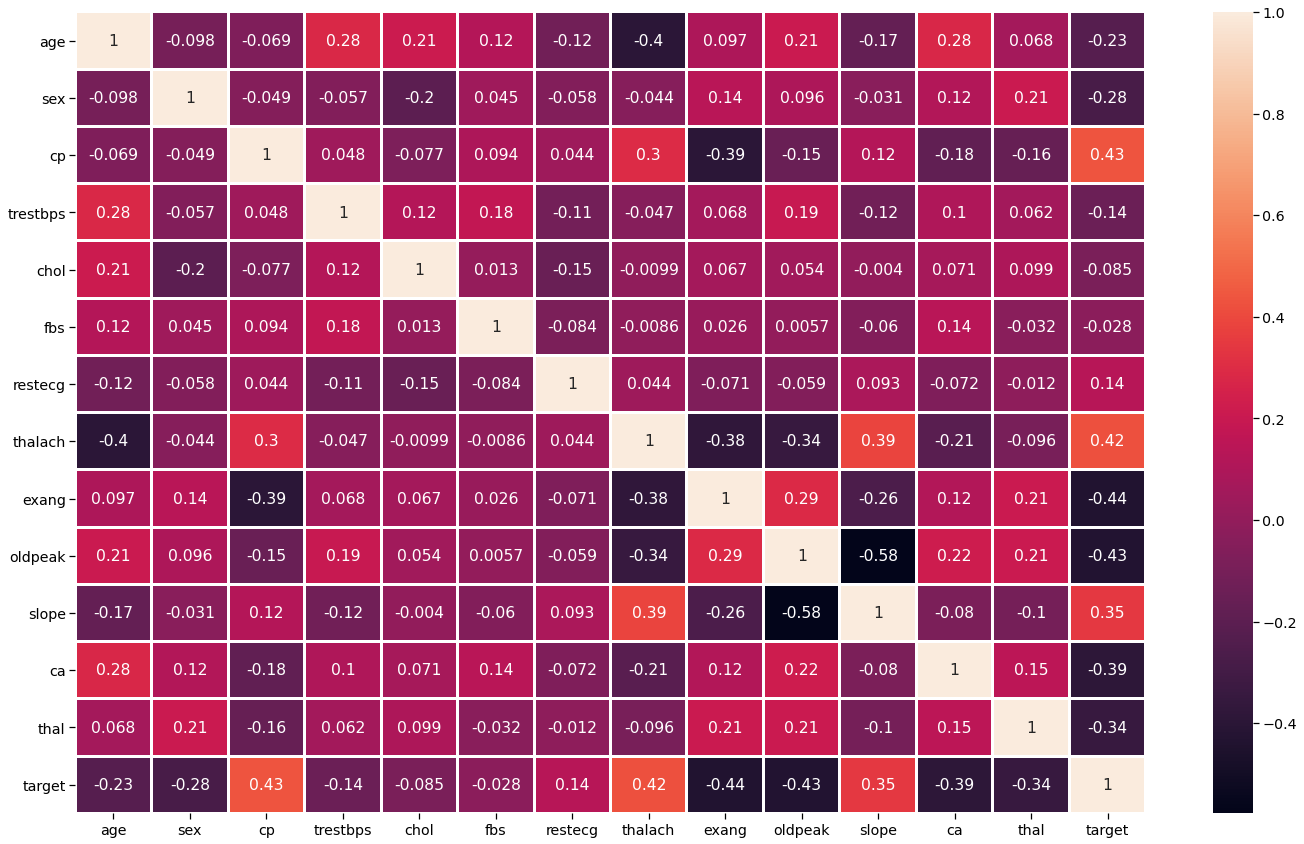
# Simulation Results

In this section, we describe our simulation scenario which we implemented to demonstrate the effectiveness of the proposed method. In the Table , we mentioned our simulation parameters.

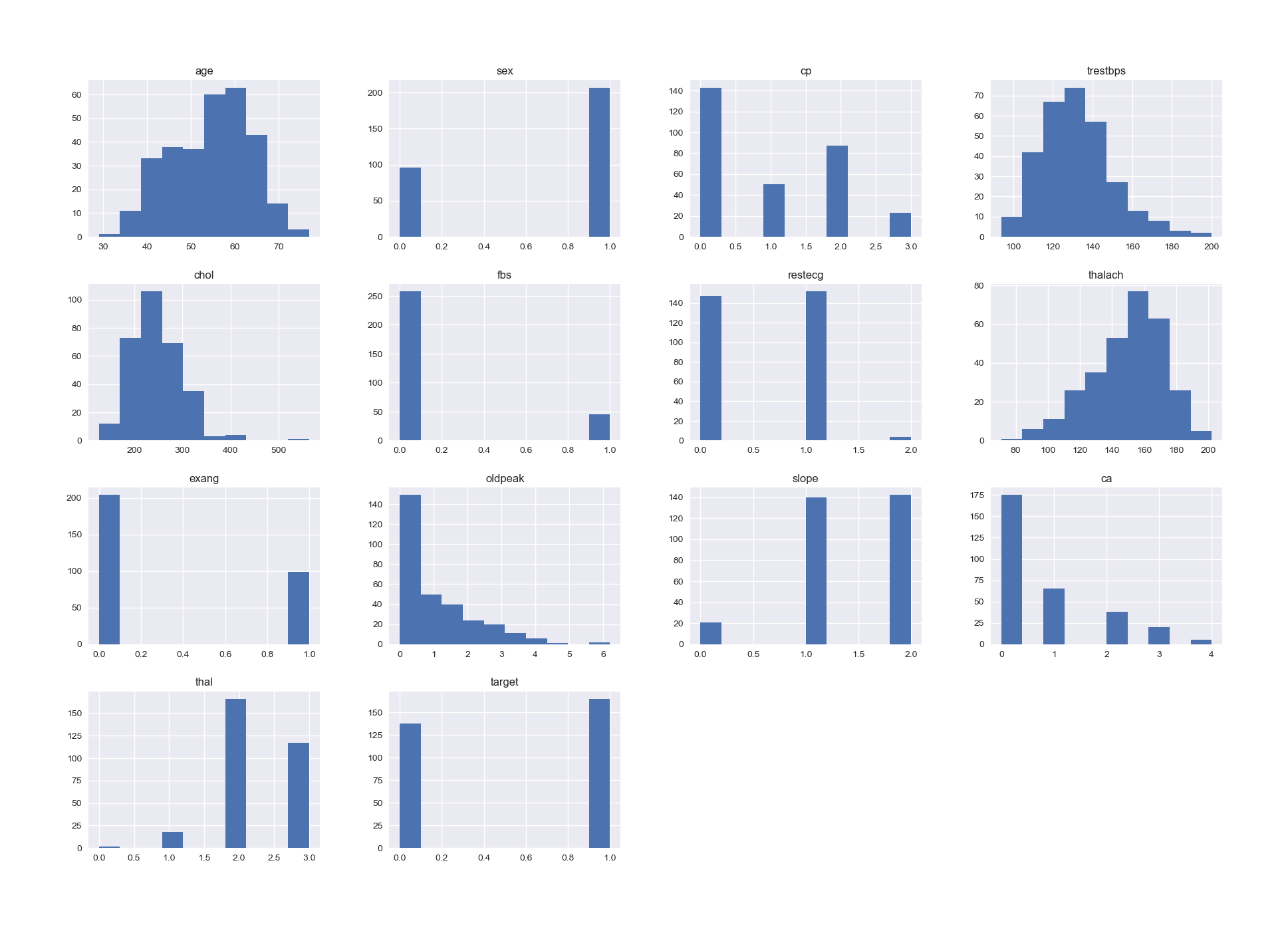


Initially, we collected a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 67% of training data is used and 33% of data is used for testing.

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model



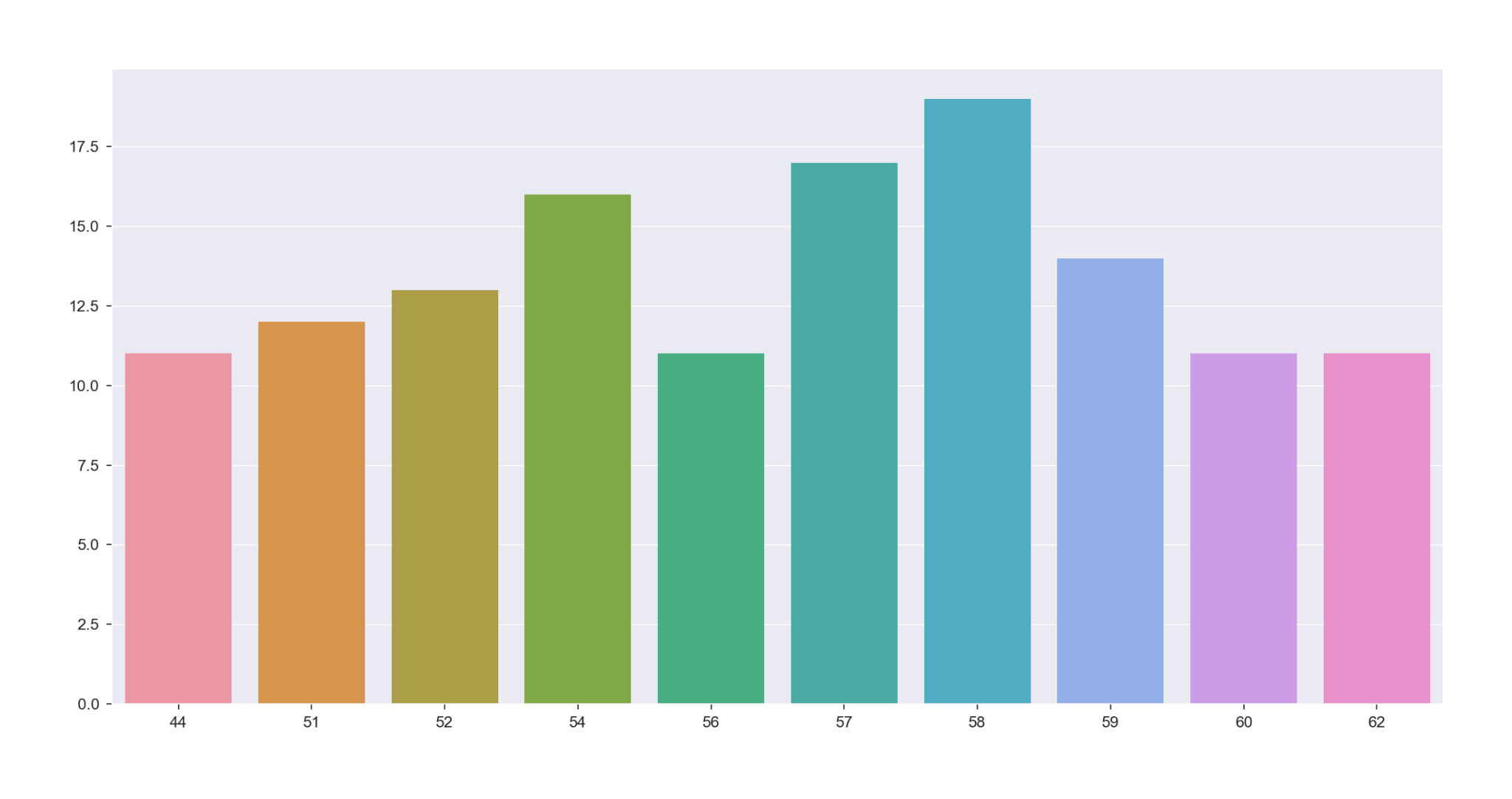
By far we have checked the correlation between the features but it is also a good practice to check the correlation of the target values.



It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning.

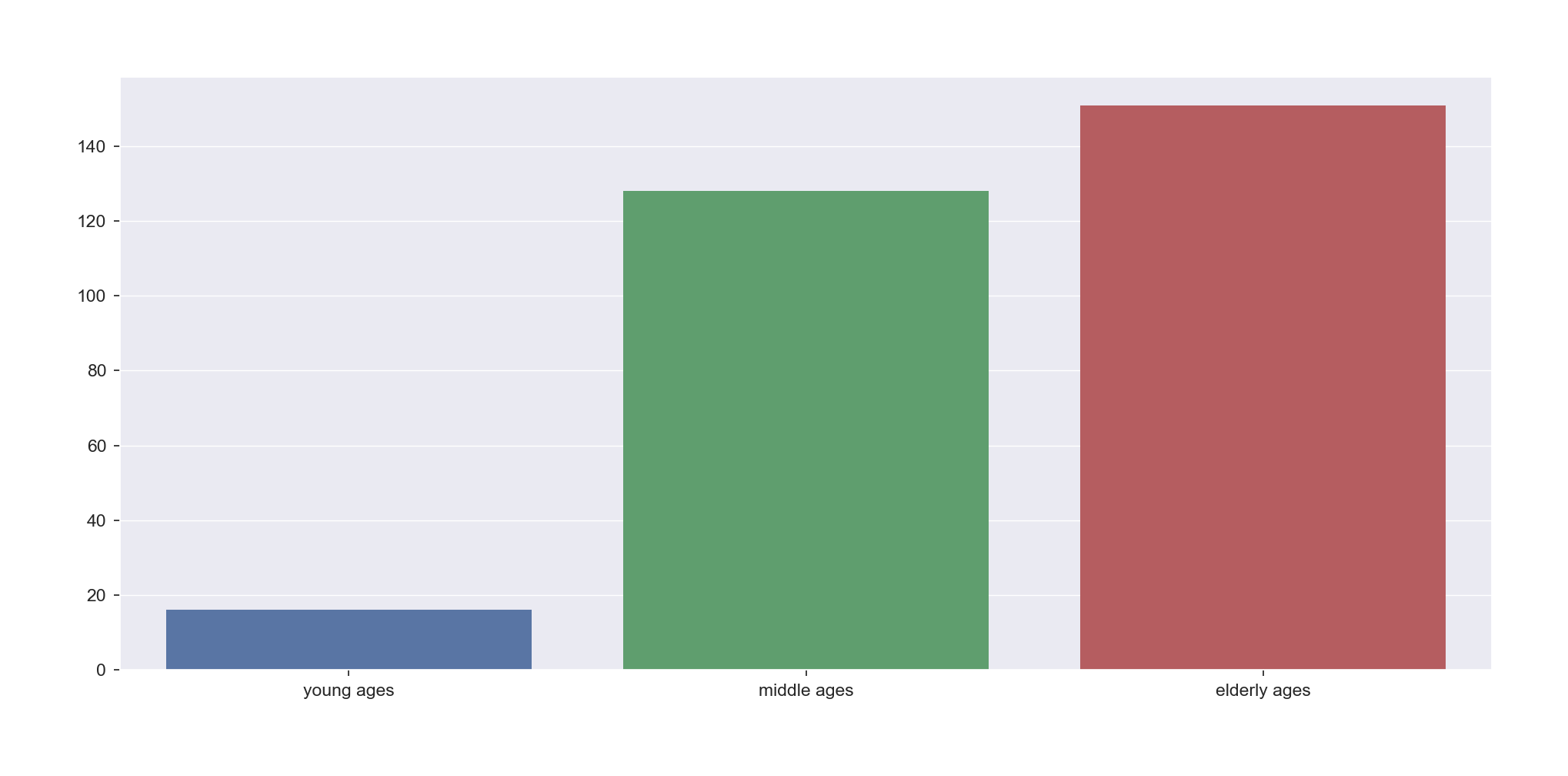
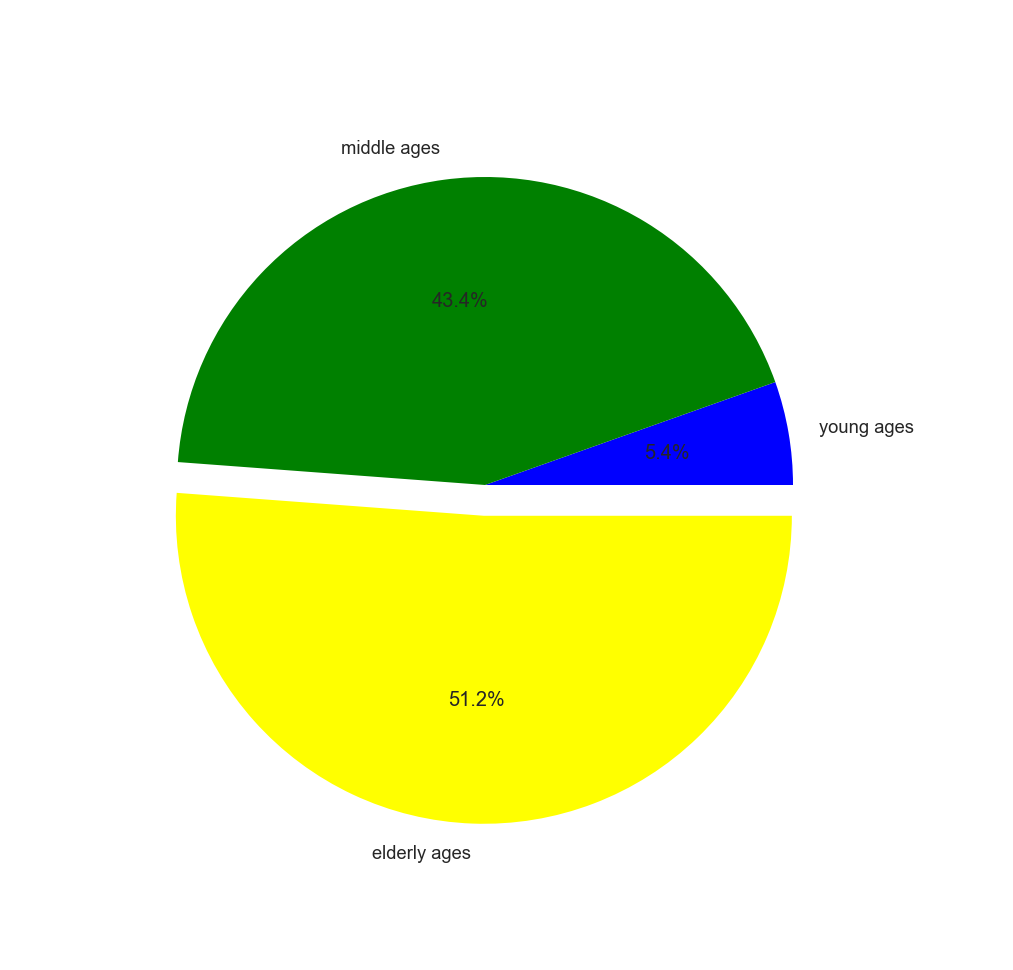
So, we have done enough collective analysis. Now let’s go for the analysis of the individual features which comprises both univariate and bivariate analysis.

Age(“age”) Analysis:- Here we will be checking the 10 ages and their counts.

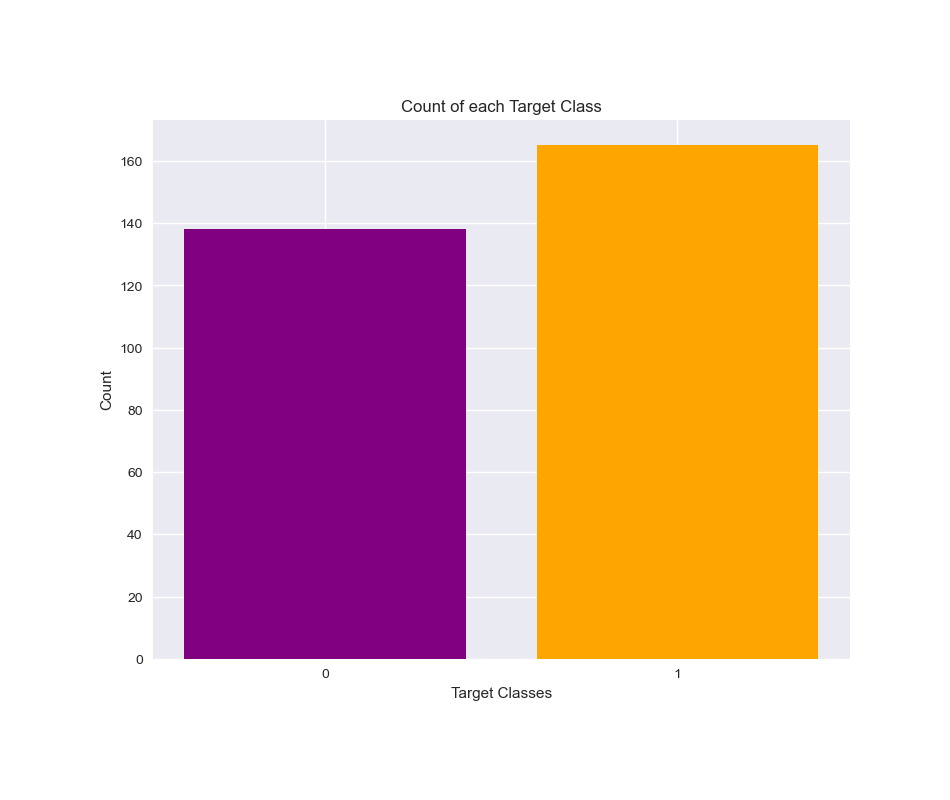


*Inference*-Here we can see that the 58 age column has the highest frequency.

We should divide the Age features into 3 parts- ‘Young’,’Middle’ and ‘Elder’.



Pre-processing of Data Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.



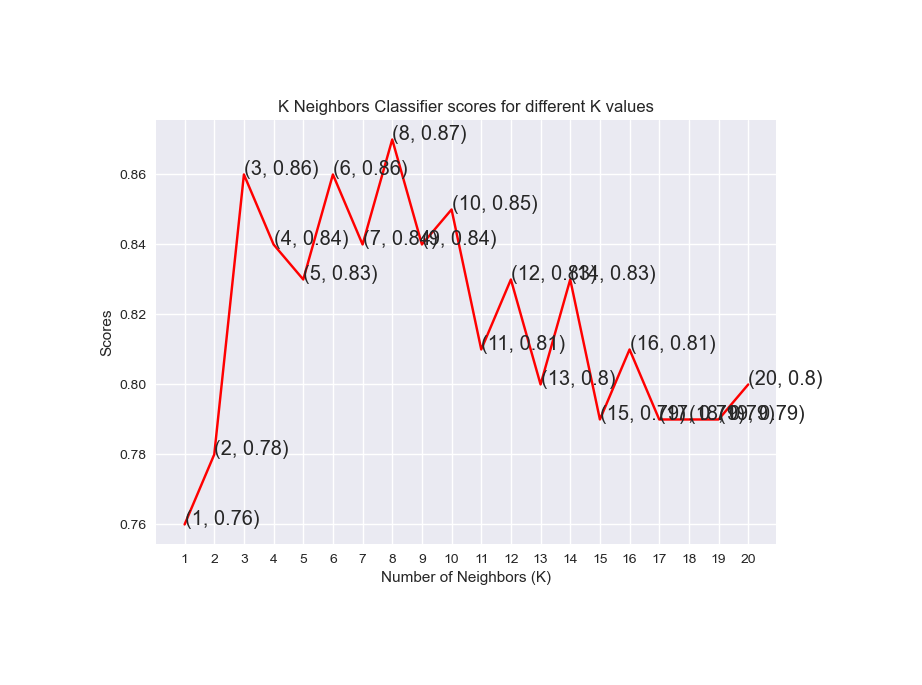
*Inference*: The ratio between 1 and 0 is much less than 1.5 which indicates that the target feature is not imbalanced. So for a balanced dataset, we can use accuracy\_score as evaluation metrics for our model.

Now we will do feature engineering to convert the categorical values in the dummies so that it will be easier for us to test the data in our machine learning algorithms.

Testing our ML algorithms in the model to check the score of our model to predict the patterns and make conclusions from them.

Now we will be using the standard scaler method to scale down the data so that it won’t raise the outliers. Also the dataset which is scaled to general units leads to having better accuracy.

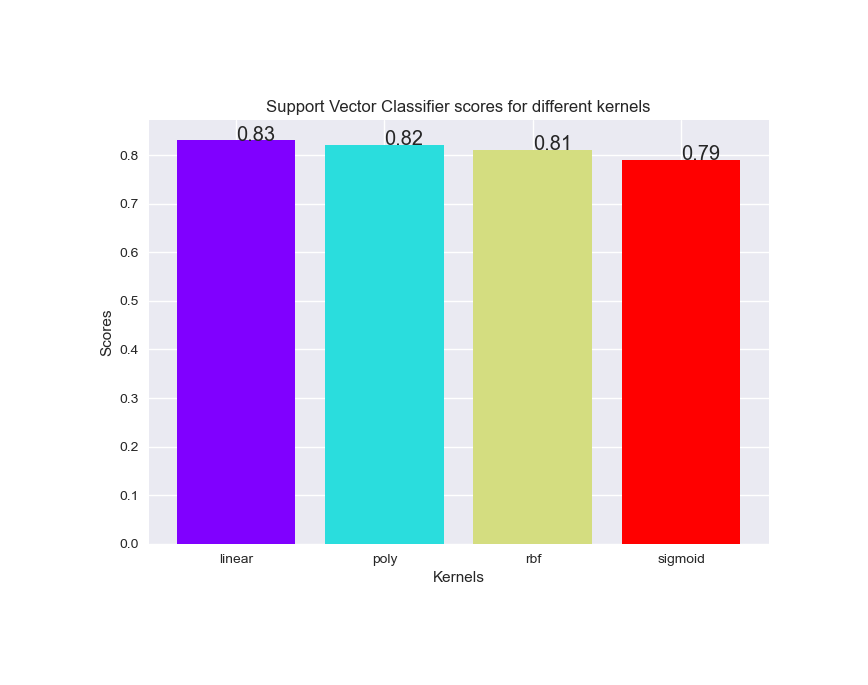
1- KNN:-This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied. I varied them from 1 to 20 neighbors and calculated the test score in each case.



***Inferenc*e**:-As you can see, we achieved the maximum score of 87% when the number of neighbors was chosen to be 8.

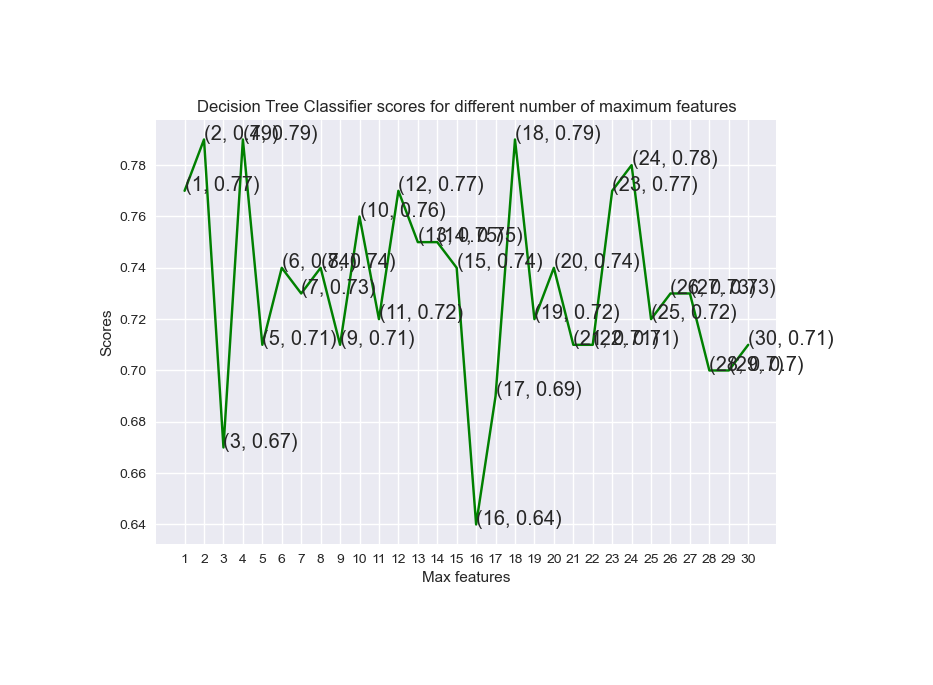
2- SVC:-This classifier aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. There are several kernels based on which the hyperplane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.

We use the rainbow method to color the different bars in the bar graph as you can see.



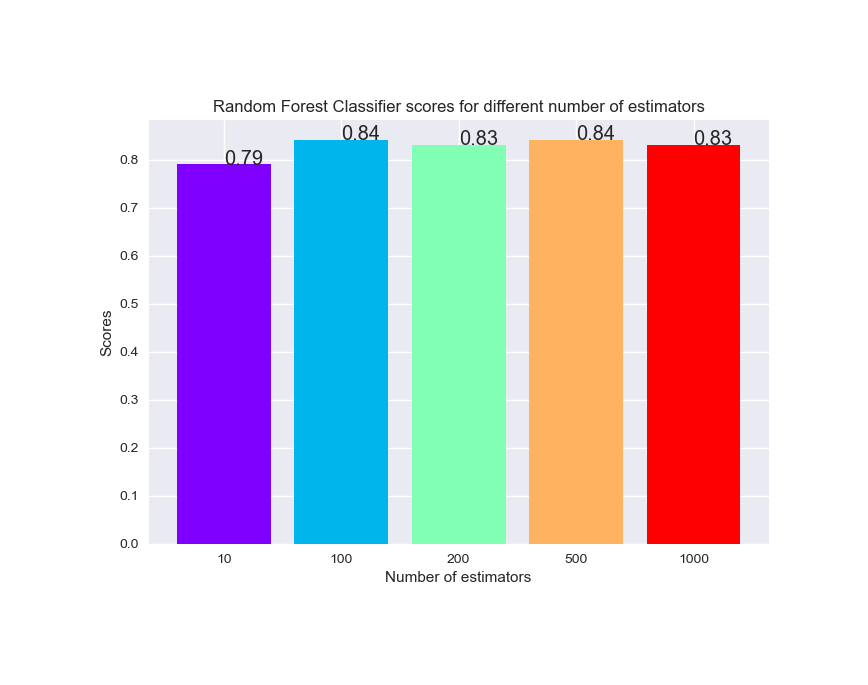
*Inference*:-As can be seen from the plot above, the linear kernel performed the best for this dataset and achieved a score of 83%.

3- Decision Tree Classifier:-This classifier creates a decision tree based on which it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model. I range features from 1 to 30 (the total features in the dataset after dummy columns were added).



*Inference*:-From the line graph above, we can clearly see that the maximum score is 79% and is achieved for maximum features being selected to be either 2, 4 or 18.

4- Random Forest Classifier:-This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features. Here, we can vary the number of trees that will be used to predict the class. I calculate test scores over 10, 100, 200, 500 and 1000 trees.



Next, I plot these scores across a bar graph to see which gave the best results. You may notice that I did not directly set the X values as the array [10, 100, 200, 500, 1000]. It will show a continuous plot from 10 to 1000, which would be impossible to decipher. So, to solve this issue, I first used the X values as [1, 2, 3, 4, 5]. Then, I renamed them using xticks.

*Inference*:-Taking a look at the bar graph, we can see that the maximum score of 84% was achieved for both 100 and 500 trees.

# Conclusion

The project involved analysis of the heart disease patient dataset with proper data processing. Then, 4 models were trained and tested with maximum scores as follows:

K Neighbors Classifier: 87%

Support Vector Classifier: 83%

Decision Tree Classifier: 79%

Random Forest Classifier: 84%

K Neighbors Classifier scored the best score of 87% with 8 neighbors.

##### References

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259- 64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757- 899X/1022/1/012072 9

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International MutliConference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE. International Journal of Management, Technology And Engineering Volume XI, Issue VI, JUNE/2021 ISSN NO : 2249-7455 Page No: 34

[9] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

[10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7.

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.