**PREDICTION**

**OF CARDIO VASCULAR DISEASE**

***A***

***project***

***report submitted to***

***MALLAREDDY***

***UNIVERSITY***

***in***

***partial***

***fulfillment***

***of***

***the***

***requirements***

***for***

***the***

***award***

***of degree***

***of***

**BACHELOR**

**OF**

**TECHNOLGY**

**in**

**COMPUTER**

**SCIENCE**

**&**

**ENGINEERING**

**AI**

**(**

**&**

**ML)**

**Submitted**

**by**

**Y.Cholavi**

**:**

**2011CS020231**

**M.Srinath**

**:**

**2011CS020232**

**M.Srinithish : 2011CS020233**

**M.Suma :**

**2011CS020234**

***Under***

***the***

***Guidance***

***of***

**Prof**

**.**

**SabyaSachi**

**Chakraborty**

**Assistant**

**Professor**

**DEPARTMENT**

**OF**

**COMPUTER**

**SCIENCE**

**&**

**ENGINEERING**

**(**

**AI**

**&**

**ML)**

2023



|  |
| --- |
| **COLLEGE CERTIFICATE**        This is to certify that this is the bonafide record of the application development entitled, **Prediction of Cardio Vascular Disease** Submitted by Y. Cholavi (2011CS020231),M. Srinath (2011CS020232), M. Srinithish (2011CS020233),M. Suma (2011CS020234), B. Tech III year I semester, Department of CSE(AI&ML) during the year 2022-23. The results embodied in the report have not been submitted to any other university or institute for the award of any degree or diploma          **INTERNAL GUIDE HEAD OF THE DEPARTMENT**    **Prof .SabyaSachi Dr. Thayyaba Khatoon**  **Assistant Professor CSE(AIML)**  **EXTERNAL EXAMINER** |

|  |  |  |
| --- | --- | --- |
|  | **CONTENTS** |  |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| 1. | INTRODUCTION  1.1 Problem definition | 1 |
| 2. | 1.2 Objective of project 1.3 Limitations of project ANALYSIS: | 1  2 |
|  | 2.1 Introduction Problem definition | 2 |
|  | 2.2 Software requirement specification | 2 |
|  | 2.2.1 Software requirement | 2 |
|  | 2.2.2 Hardware requirement | 2 |
|  | 2.3 Existing System | 3 |
|  | 2.4 Proposed Systems | 3 |
|  | 2.5 Modules | 3 |
|  | 2.6 Architecture | 6 |
| 3. | DESIGN:  3.1 Introduction |  |
|  | 3.2 Data Set Descriptions | 7 |
|  | 3.3 Data Preprocessing Techniques | 8 |
|  | 3.4 Methods & Algorithms | 8 |
|  | 3.5 Building a Model | 11 |
|  | 3.6 Evaluation | 14 |
| 4. | DEPLOYMENT AND RESULTS  4.1 Source Code | 14 |
|  | 4.2 Final Results | 22 |
| 5. | CONCLUSION:  5.1 Project conclusion | 23 |
|  | 5.2 Future Scope | 24 |

# ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all the efforts with success. We are very grateful to our project mentor Sabyasachi, for the guidance, inspiration and constructive suggestions that helped us in the development of this application.

We are much obliged to A. Shiva Kumar (Application Development Incharge) for encouraging and supporting us immensely by giving many useful inputs with respect to the topic chosen by us, throughout the development of the application ensuring that our project is a success.

We also express our heartfelt gratitude to Dr. Thayyaba Khatoon(Head of Department), for

giving all of us such a wonderful opportunity to explore ourselves and the outside world to work on the real-life scenarios where the machine learning is being used nowadays.

We also thank our parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

# ABSTRACT

Heart disease is one of the most significant problem that is arising in the world today.

Cardiovascular disease prediction is a critical challenge in the area of clinical data analysis. Hybrid Machine learning (ML) has been showing an effective assistance in making decisions and predictions from the large quantity of data produced by the healthcare industries and hospitals. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT).

Various studies give only a glimpse in predicting heart disease with ML techniques.

In this paper, we propose a narrative method that aims at finding significant features by applying machine learning techniques that results in improving the accuracy in the prediction of cardiovascular disease. The prediction model is proposed with combinations of different features and several classification techniques. We produce an enhanced performance level with an accuracy level of 92% through the prediction model for heart disease with the hybrid random forest with a linear model

Heart Disease Prediction is a critical challenge in the area of clinical Data Analysis. Machine

learning and Artificial Intelligence are being used in the health field. Machine learning (ML) especially, has been shown to be effective in making decisions and predictions from the large quantity of data produced by the Healthcare Industry. Based on different attributes such as blood pressure, heart rate, and other characteristic attributes, heart diseases can be predicted. This project intends to point the most relevant risk factors of heart disease as well as predict the overall risk by using various Machine Learning techniques.

# Chapter 1

## INTRODUCTION 1.1 PROBLEM DEFINITION

Heart disease can be managed effectively with a combination of lifestyle changes, medicine

and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive

# 1.2 OBJECTIVE OF PROJECT

The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level,etc. A data set is selected from the UCI repository with patients medical history and attributes

# 1.3 LIMITATIONS OF PROJECT

1. Prediction of cardiovascular disease results is not accurate
2. Data mining techniques does not help to provide effective decision Making
3. it cannot handle time complexiety of Doctor
4. Cannot handle enormous data sets for patient records.

**Chapter 2 ANALYSIS**

# 2.1 INTRODUCTION

The heart is one of the main organs of the human body. It pumps blood trough the bloodvessels of the circulatory system.The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body.Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death.Predicting and diagnosing heart disease is the biggest challenge in the medical industry and relies on factors such as the physical examination, symptoms and signs of the patient. Machine learning algorithms play an essential and precise role in the prediction of heart disease. Heart disease can be predicted based on various symptoms such as age, gender, heart rate, etc, and reduces the death rate of heart patients**.**

## 2.2 SOFTWARE REQUIREMENT SPECIFICATION

### 2.2.1 Software Requirement

* Jupyter Notebook

* Python

* Google Chrome or Microsoft edge of Latest Version

* Libraries : Python (pandas, Numpy , Matplotlib , Seaborn,Sklearn)

### 2.2.2 Hardware requirement

* OS: Windows 10 or Higher

* Processor: Intel i5 processor or Higher

* Ram: Minimum 8 GB or Higher

* Hard Drive: Minimum 256 GB or Higher

## 2.3 EXISTING SYSTEM

In this system, the input details are obtained from the patient. Then from the user inputs, using ML techniques heart disease is analyzed. Now, the obtained results are compared with the results of existing models within the same domain and found to be improved. The data of heart disease patients collected from the laboratory is used to discover patterns with Support Vector machines SVM. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 87% for F-measure, competing with the other existing methods

## 2.4 PROPOSED SYSTEMS

Data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This project proposes a prediction model to predict whether a people have a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

## 2.5 MODULES

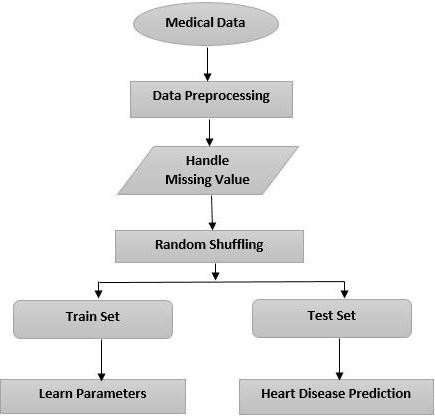


Fig: Modules of Heart disease prediction

### 1) Collection of Data and Preprocessing

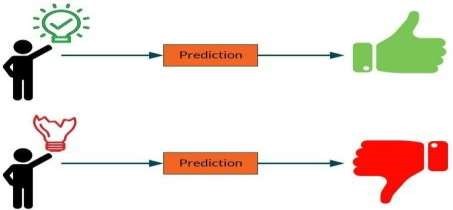
At first, we are going to collect image datasets that consist of fish, sea turtles, andother aquatic animals, and we are going to do preprocessing in which we are going to do data integration, data transformation, data reduction, and data cleaning. From those images,we are going to send them for training and testing.

**2) Handling Missing Values**

Data handling is necessary as it removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this partof approach, the Null (NaN) values are removed from the dataset to make it more useful asthese values decrease the productivity of the algorithm. At the data handling stage, the dataset is also normalized to not have any ambiguity after cleaning

### 3) Prediction of Disease

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives thehighest accuracy is used for heart disease prediction.



**2.6**

**ARCHITECTURE**

Fig:

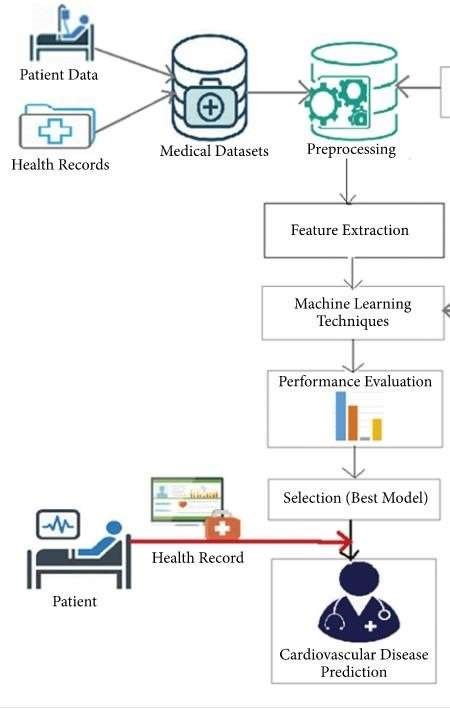
Architecture

of

Heart

disease

prediction



Multi

-

filter



NAVIE

BAYES

KNN

**Chapter 3**

**DESIGN**

# 3.1 Introduction

These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. In this paper, a brief review and future prospect of the vast applications of machine learning algorithms has been made.

# 3.2 Dataset Description



This is our dataset which is consisting of 918 rows and 12 columns.

The attributes used are:

* AGE: Age of the person.
* SEX: Whether the person is male or female. If female give 0 and male give 1. - BPLEVEL: (Blood Pressure) Here the user has to specify their bp level.
* CHOL: (Cholesterol) The user is requested to mention their cholesterol level.
* FBS: (Fasting Blood Sugar) Elevated levels are associated with diabetes and insulin resistance.
* RESTECG: (electrocardiogram) An ECG is a simple test that can be used to check your heart’s rhythm and electrical activity.
* THALACH: It is the person’s maximum heartrate received.
* OLDPEAK: Asymptomatic chest pain
* SLOPE: The ST segment shift relative to exercise-induced increments in heart rate. - CA: The calcium present in your body.

# 3.3 Data Preprocessing Techniques

**Process Structuring:**

There are five steps in Process Structuring, means gaining insight into the way your organisation operates. It introduces a layering in the way of working that gives you and your employees insight into which processes there are and how these relate to each other. Which method to choose for the structuring of your processes depends on your organisation. They are:

Data Collection.

Pre-Processing.

Data Analysis.

Application of Algorithms.

Evaluating the Models.

* DATA COLLECTION:

Data collection is the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

* PRE PROSESSING:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data preprocessing task.

* DATA ANALYSIS:

Data Analysis is a process of inspecting cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

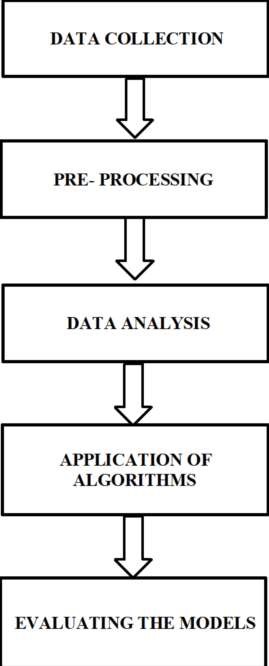
* APPLICATION OF ALGORITHMS:

An Algorithm in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the

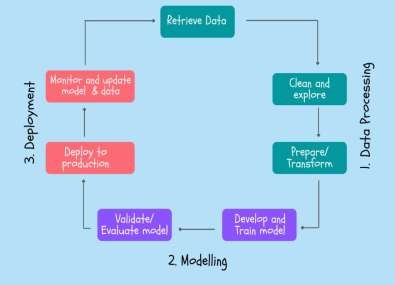
results of this analysis over many iterations to find the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.

* EVALUATING THE MODELS:

Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring. It is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and overfitted models.



# 3.4 Methods and Algorithms



Process Of Data Modelling:

1.Data Processing:

Here in this step first we will be accessing the data and processing in such a way that it is in a structured format. Here we use the EDA process. There are 4 steps involved in EDA.

* + Knowing initial values
  + Removing unwanted data
  + Converting into required format
  + Plotting graphs

1. Data Modelling:

here to develop a model firstly we have divided the data into 2 groups i.e., train and test parts so we have accessed the algorithm from sklearn learn model. Later we have fit the data into the model. With the help of accuracy score we have validated and evaluated. We developed a predictive system here.

1. Deploy and Production:

publishing the application into the market. After deploying, we keep on updating our application constantly. While updating, if any data is required to add to the application, then we perform the above-mentioned process again. Hence, it is a cyclic process.

By following the above-mentioned procedure of data modelling, we have developed the following code **Algorithms :**

**Decision Tree** is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

**Support Vector Machine** or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.It chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

**Random Forest** is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**K-Nearest Neighbour** is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. That stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. This can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

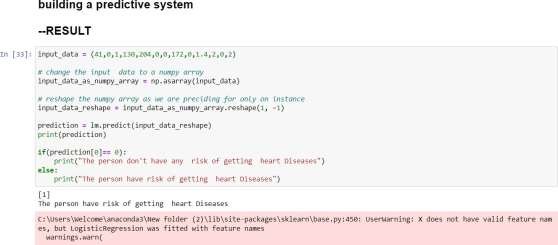
These are the Accuracy of all ML Algorithms chosen for the purpose of Classification are given below:

|  |  |
| --- | --- |
| Naive Bayes | 86.96% |
| Decision Tree | 82.61% |
| SVM | 86.41% |
| Logistic Regression | 85.87% |
| KNN | 88.04% |

# 3.5 Building a Model

Here we can chosen K-NN Algorithm to build the Model for our dataset collected, Since it got Good Accuracy and gives better results.

# 3.6 Evaluation



Using K-nn as per the input values it predicts if the person is having Heart Disease or not. Here we got 1 in so “The person have any risk of getting heart diseases.



Here we got 0 in so “The person don’t have any risk of getting heart disease

**Chapter 4**

**DEPLOYMENT AND RESULTS**

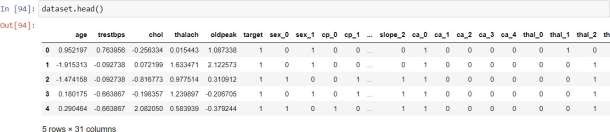
# 4.1 Source Code

**JUPYTER NOTEBOOK CODE:**

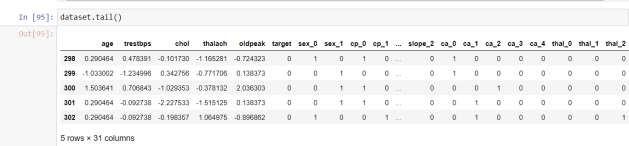


Here the various packages are being imported. They are: pandas, matplotlib.pyplot, numpy, seaborn and sklearn.

Using the read function and pandas package we are loading the data



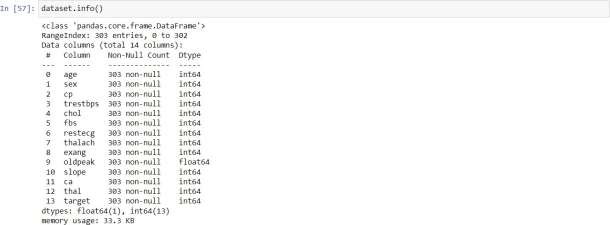
Here using the head function we get the first 5 rows of the dataset.



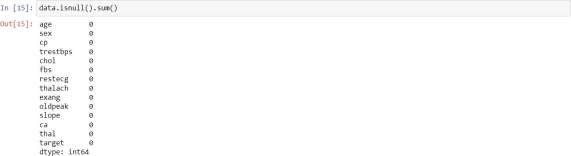
Using tail function we can view the last 5 rows of the dataset.



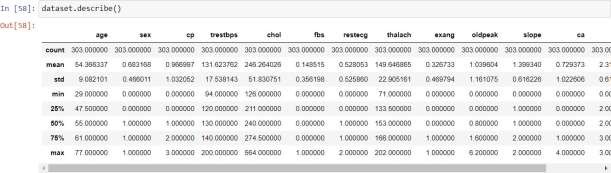
Using shape we can get dimensions of the dataset.



Here, we get the complete information about the attributes i.e., the columns present along with their datatype.

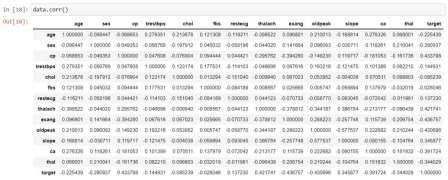


With the help of isnull function we can find if any null values present present in the dataset and sum gives the sum of them.

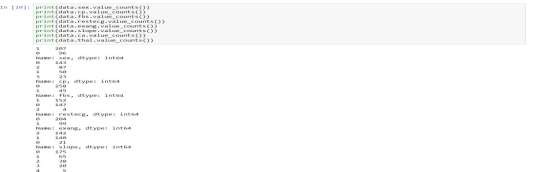


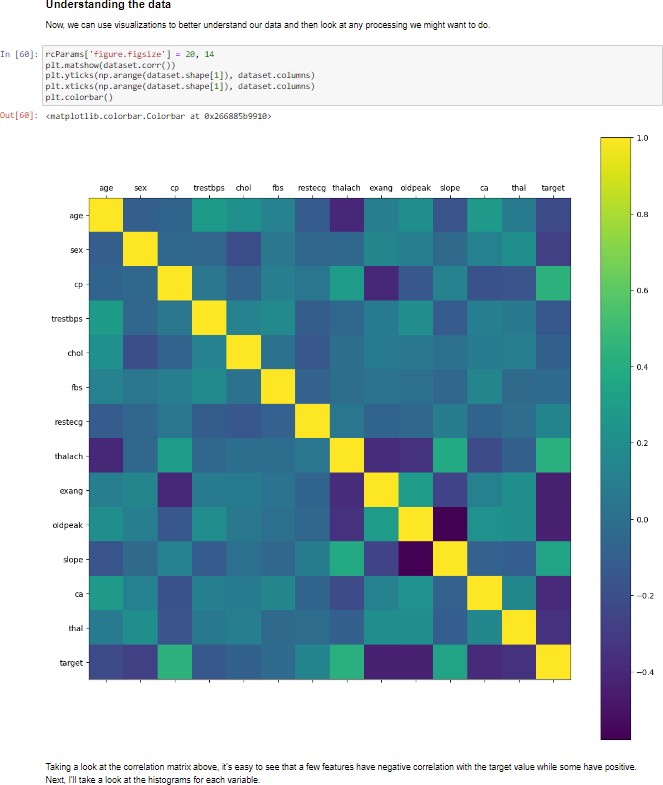
Here using describe() function, we have got values of some Statistical information like count, mean, standard deviation, minimum, maximum, 25%, 50% and 75% of all attributes like age,sex,cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca.

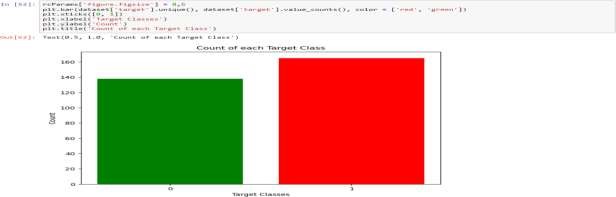
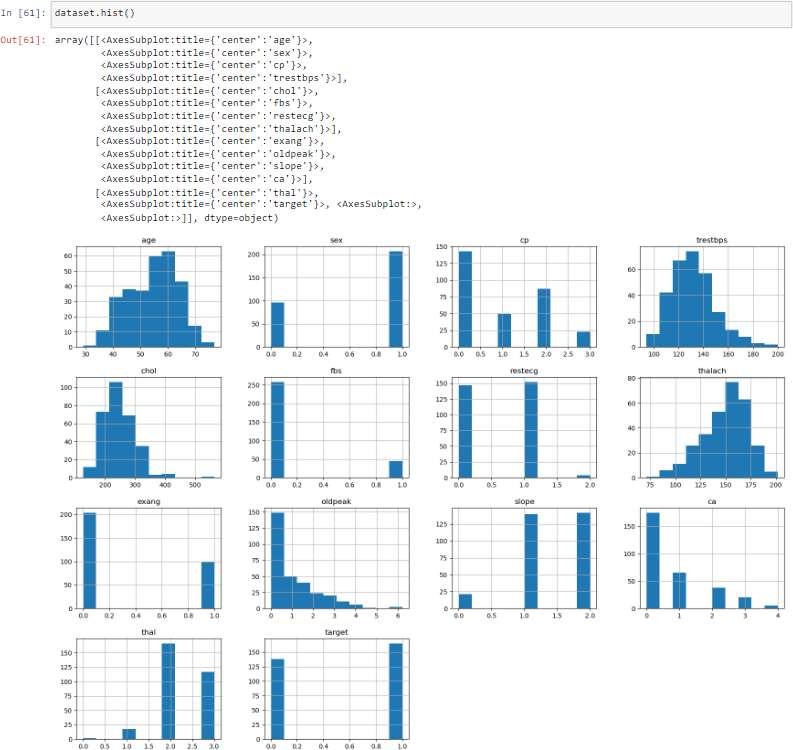




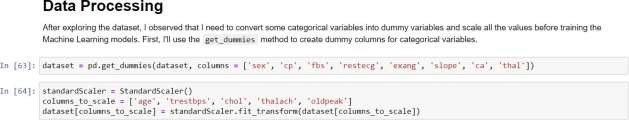
Here using corr() function we can get the relation between each Attribute with the other Attribute using Numeric value.



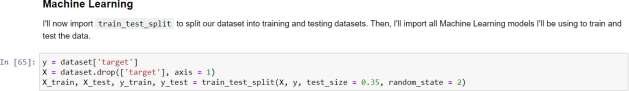




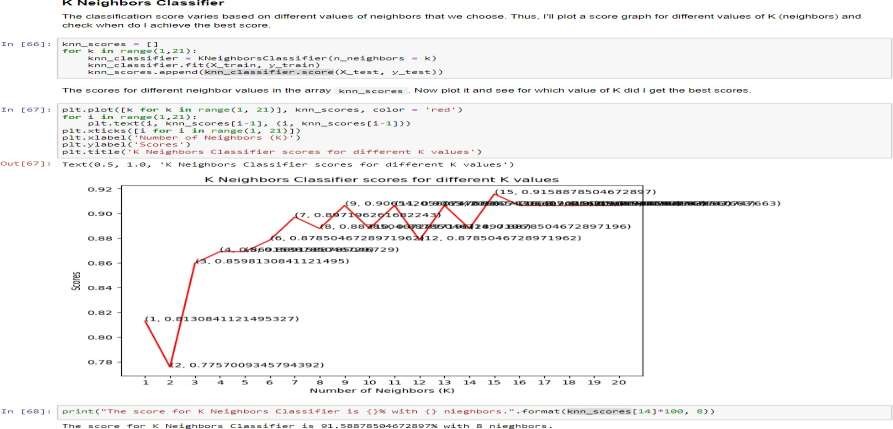
This bar plot shows categorical data as rectangular bars with the height of bars proportional to the value they represent.



Here we do the data preprocessing.



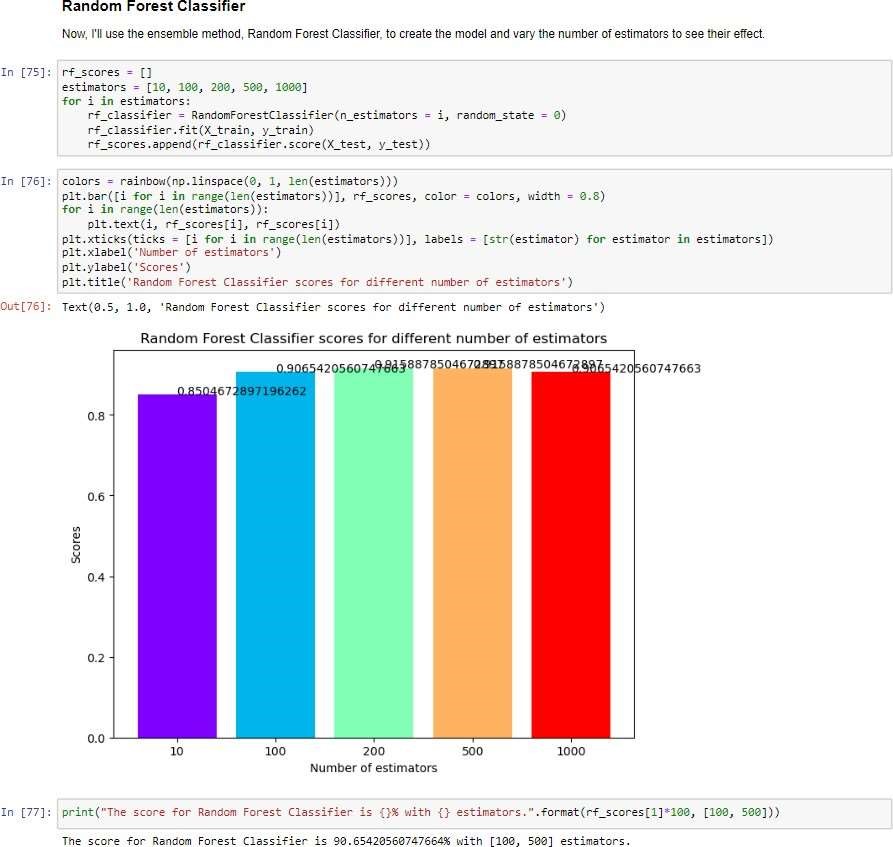
We have Trained and Tested the Data into 3.5:6.5 Ratio.



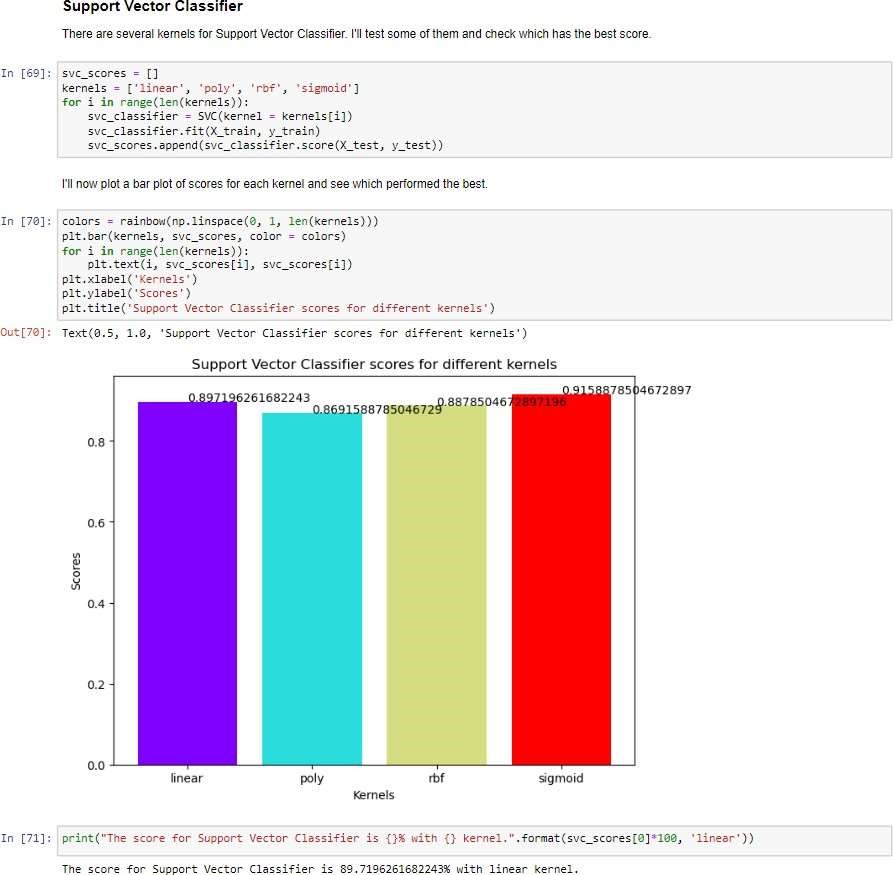
Here we had fit the data into K-nn by X\_train and Y\_train parameters.

After training the model we used the accuracy\_score function to check our model accuracy. We can also use score() function which gives the accuracy of the model for this model.

By fitting the Model into K-nn, we acquired the Accuracy of 88%.

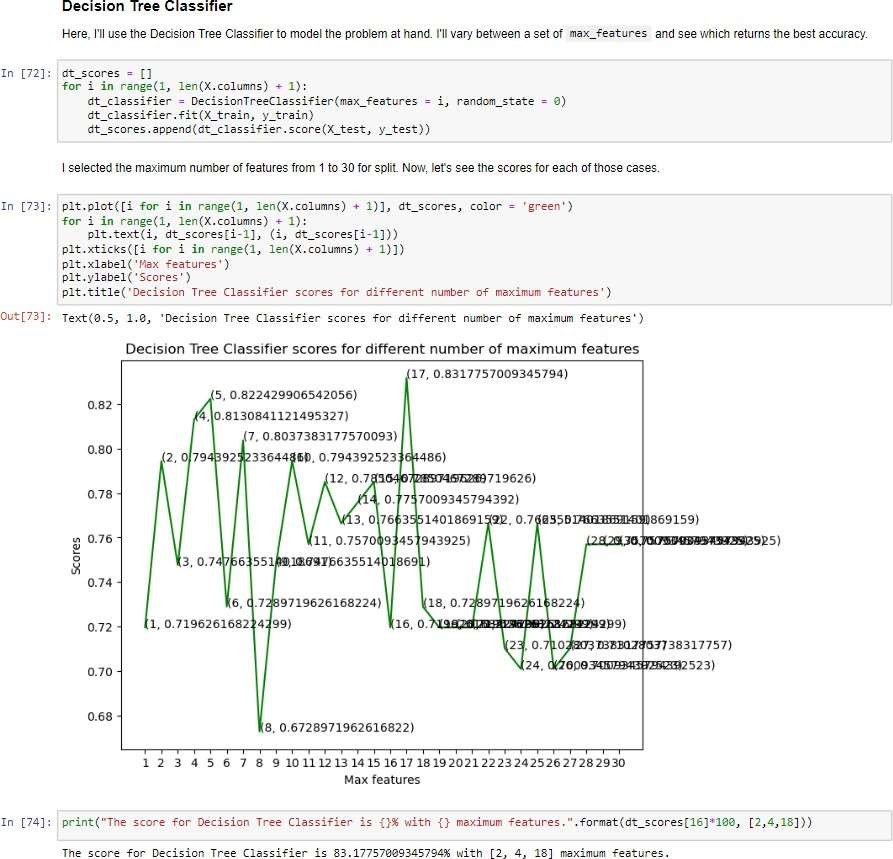


Random Forest is a powerful and versatile **supervised machine learning algorithm** that grows and combines multiple decision trees to create a “forest.” Here we got 90.65% Accuracy.



SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

Here we got 85.9% Accuracy.



# 4.2 Final Results

Accuracy of ML Algorithms

|  |  |
| --- | --- |
| Naive Bayes | 86.96% |
| Decision Tree | 82.61% |
| SVM | 86.41% |
| Logistic Regression | 85.87% |
| KNN | 88.04% |

# Chapter 5

# CONCLUSION AND FUTURE SCOPE

**5.1 Project Conclusion:**

The identification of raw healthcare data of heart information processing would aid in the long-term saving of human lives and the early detection of defects in heart conditions. Based on the project's findings, it can be inferred that machine learning algorithms have a lot of potential in the medical field. In the medical world, predicting cardiovascular disease is both difficult and crucial. . However, the death rate can be drastically controlled if the disease is detected at the early stages and preventative measures can be adopted as soon as possible. It would be ideal if this research could be expanded to include real-world datasets rather than only theoretical methods and simulations. The proposed KNN algorithm solution has been shown to be very effective in terms of heart disease prediction.

**5.2 Future Scope:**

* The future scope of this project can be performed with diverse mixtures of machine learning techniques to better prediction techniques.
* Furthermore, new feature-selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.
* Presently, we are just predicting whether a person will be affected with any heart disease or not, but later we will be working on if the person if prone to get a heart disease, which disease he/she would be affected with
* Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting.

# REFERENCES

1. T.M. Mitchell, “Machine Learning”, McGraw-Hill, 1997.

1. Machine Learning, Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Pearson, 2019.

[3] Ethern Alpaydin, “Introduction to Machine Learning”, MIT Press, 2004.

1. Stephen Marsland, “Machine Learning -An Algorithmic Perspective”, Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

1. Andreas C. Müller and Sarah Guido “Introduction to Machine Learning with Python: A Guide for Data Scientists”, Oreilly.