

PREDICTION OF HEART BASED ILLNESS USING MULTIPLE MACHINE LEARNING ALGORITHMS

UNIVERSITY OF HERTFORDSHIRE

School of Computer Science

Antony Praveen Raj Irudayaraj : 20021575

02 December 2022

**Course:** MSc Project for Artificial Intelligence and Robotics 7COM1039: Final Project Report

**Student Name:** Antony Praveen Raj Irudayaraj

**Student Number:** 20021575

**Supervised by:** Chidinma Chiejina

# **Abstract**

The World Health Organization estimates that 18 million people worldwide die each year from cardiovascular diseases, making them the leading cause of death globally. Because an early diagnosis and detection of diseases may hold the key to a cure, they are essential. Since early disease detection is the main challenge, most scientists and researchers are focusing on machine learning algorithms that can precisely detect massive amounts of complex data. This research aims to answer can AI be useful in detecting heart problems using machine learning algorithms and what is the maximum accuracy machine learning algorithms can give? In order to detect heart diseases early and stop their consequences, this study employs a variety of machine learning algorithms, including KNN, Decision Tree, Logistic Regression, Random Forest, and Naïve bayes. to identify the heart illnesses and prevent their repercussions. I have selected a heart failure records dataset and from the dataset I am using serum\_creatinine and ejection\_fraction to train the machine learning models to determine heart problems. After conducting the research the accuracy for the different algorithms are as follows: Logistic regression accuracy - 84%, K - Nearest Neighbors accuracy - 79%, Naive bayes accuracy - 73%, Decision tree accuracy - 78% and Random forest accuracy - 82%. Among these methods, logistic regression has the highest accuracy and can be used to detect heart diseases. But it is good practice to use multiple methods to determine the presence of heart disease to eliminate false positive. As per the stated result it was found that machine learning algorithms are very much useful in predicting the heart based illness with the good amount of accuracy than other methods and I used five machine learning algorithms in this research method. I found that using multiple machine learning algorithms, I happened to get the better accuracy to predict the heart-based illness. By using these five machine learning algorithms, it helps us to get better accuracy than the existing models. Further research can be done by using natural language processing for the self-learning process in this project.

# Acknowledgements

Antony Praveen Raj Irudayaraj, my final project report is finished. Thanks to the assistance and direction of numerous people, for making this possible. I’d like to take this opportunity to thank everyone who helped me along the way and made this possible.

I want to express my gratitude and a special thank you to my supervisor Chidinma Chiejina, for all of her support, encouragement, and excellent mentoring throughout the project. She inspired me to learn and put new skills into practice by making this project exciting. Finally, I want to sincerely express my gratitude to my parents, without whose unwavering support and sustained help I could not have finished my project.

# MSc Final Project Declaration

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Artificial Intelligence and Robotics at the University of Hertfordshire (UH). It is my own work except were indicated in the report. I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged.

**Table of Contents**

[Abstract 2](#_Toc131256349)

[Acknowledgements 3](#_Toc131256350)

[MSc Final Project Declaration 3](#_Toc131256351)

[1. Introduction 6](#_Toc131256352)

[2. Literature Review 8](#_Toc131256353)

[3. Methodology 10](#_Toc131256354)

[3.1. Implementation Strategy and Issues dealt 11](#_Toc131256355)

[3.2. Data Collection 11](#_Toc131256356)

[3.3. Data Analyzation 12](#_Toc131256357)

[3.4. Machine Learning Algorithm 20](#_Toc131256358)

[3.5. Supervised Learning 20](#_Toc131256359)

[3.6. Machine Learning Algorithms Used 22](#_Toc131256360)

[3.7. Data pre-processing 28](#_Toc131256361)

[3.8. Fitting the machine learning algorithms to prepare the models for training 28](#_Toc131256362)

[3.9. Legal, Ethical, Social And Professional Implication 31](#_Toc131256363)

[3.10. Commercial Considerations 31](#_Toc131256364)

[4. Results 32](#_Toc131256365)

[4.1. Predicting the test values and finding the accuracy 32](#_Toc131256366)

[5. Discussion, Evaluation and Conclusion 36](#_Toc131256367)

[5.1. Discussion 36](#_Toc131256368)

[5.2. Evaluation and Conclusion 36](#_Toc131256369)

[References 38](#_Toc131256370)

[Appendix 44](#_Toc131256371)

[Link to dataset 44](#_Toc131256372)

[Code used in the research 44](#_Toc131256373)

**List of Figures**

[Figure(1): Age differences in the dataset. 13](#_Toc131157058)

[Figure(2): Anaemia condition in the patient in dataset. 13](#_Toc131157059)

[Figure(3): Creatinine phosphokinase differences in the dataset. 14](#_Toc131157060)

[Figure(4): Diabetes condition in the patient in dataset. 14](#_Toc131157061)

[Figure(5): Ejection fraction differences in the dataset. 15](#_Toc131157062)

[Figure(6): High blood pressure condition in the patient in dataset. 15](#_Toc131157063)

[Figure(7): Platelets differences in the dataset. 16](#_Toc131157064)

[Figure(8): Serum creatinine differences in the dataset. 16](#_Toc131157065)

[Figure(9): Serum sodium differences in the dataset. 17](#_Toc131157066)

[Figure(10): Whether the people smoke or not in dataset. 17](#_Toc131157067)

[Figure(11): How many percentages of men and women present in dataset. 18](#_Toc131157068)

[Figure(12): Whether the people have heart problem or not in dataset. 18](#_Toc131157069)

[Figure (13): Represents people having heart problem when the ejection fraction is at these level in the dataset. 19](#_Toc131157070)

[Figure (14): Represents people having heart problem when the serum creatinine is at these level in the dataset. 19](#_Toc131157071)

[Figure (15): Represents whether the people having heart problem or not in each age in the dataset. 20](#_Toc131157072)

[Figure (16) Flow chart representing process of training and implementing the machine learning models. 28](#_Toc131157073)

[Figure (17): Training result for the selected variables. Purple dots are who doesn’t have heart problem and green dots are who have the heart problem. 29](#_Toc131157074)

[Figure(18): Training result for the selected variables 29](#_Toc131157075)

[Figure (19): Training result for the selected variables 30](#_Toc131157076)

[Figure (20): Training result for the selected variables 30](#_Toc131157077)

[Figure (21): Training result for the selected variables 31](#_Toc131157078)

[Figure (22): Test set result of dataset using logistic regression. 33](#_Toc131157079)

[Figure(23): Test set result of dataset using KNN Algorithm 33](#_Toc131157080)

[Figure (24): Test set result of dataset using naïve bayes 34](#_Toc131157081)

[Figure (25): Test set result of dataset using decision tree. 34](#_Toc131157082)

[Figure (26): Test set result of dataset using random forest. 35](#_Toc131157083)

# **Introduction**

In Machine Learning, the usage and creation of computer systems that can adapt and learn without being given explicit instructions by analysing data patterns and making conclusions using statistical models and algorithms.

Heart affliction prediction is a technique created with the use of machine learning algorithms to forecast heart-based illnesses by analysing the patient's medical history to monitor their state of health and by feeding the patient data into the algorithm. By changing the patient's eating habits and lifestyle choices, gathering this type of data will be more beneficial in a number of ways, such as in saving the patient's life before he develops a heart condition.

My interest in machine learning led me to choose this subject because early detection of heart disease and better prevention can be achieved with the aid of machine learning algorithms like KNN Decision Tree, Logistic Regression, naïve bayes, and Random Forest.

A number of related projects have been developed in the present, but they are all based on an age range starting at fifty. These days, existing models contain only two machine learning algorithms to obtain accuracy but I have improvised that model based on five machine learning algorithms to get better accuracy. Additionally, I created a project based on the random age concept that predicts heart disease at all ages. In the modern era, everyone consumes more fast food, and many young people are addicted to alcohol and smoking, both of which may cause heart-related illnesses. My main idea is to prevent heart illnesses before they become a permanent heart disease by predicting using data from their daily activities and medical history with the aid of machine learning algorithms.

Worldwide, coronary disease is among the leading causes of death, and it can be treated. An accurate and timely diagnosis of the disease is crucial for treatment and recovery time. In their separate domains, research professionals are making intense efforts to lower the death rate. Even though there has been extensive research in this area, there is still room to improve forecast accuracy.

The main goal of this project’s planned effort is to create a machine learning based heart affiliation prediction model which will be used to predict all heart-based diseases with help of the patient’s medical data. With the help of the five machine learning algorithms, I can perfect its prediction by using Logistic Regression, KNN, Decision Tree, Naive Bayes and Random Forest.

* **Logistic Regression:**

The logistic regression model is one of the best statistical models for estimating the probability of a certain class or event, such as success or failure. The logistical regression uses a lot of predicted variables, which could be digital or class-based. Logistic regression is also known as Logit or the general study of entropy. supervised machine learning methods for "classification" issues include regression in logistics. Logistics regression is a faster and more efficient fix for binary and linear classification problems that excel with linear separation layers. By analysing the link between one or more independent variables and the data value based on the historical observation of a data collection, the logistics regression model forecasts a dependent data attribute (Ahdal et al., 2021).

* **Random Forest:**

The Random Forest algorithm, one of the finest for classification, is used for both classification and regression. The Random Forest approach, as its name suggests, consists largely of a significant number of separate decision trees that work together as a group. Each tree with the same distribution differs from the others in some way. The prediction is averaged using bootstrap aggregation and random feature selections (Yousef & Khaled Batiha, 2021).

* **KNN:**

KNN methodology is an easy-to-use but powerful classification technique. It frequently applies to classification issues and makes no simplifying assumptions, even when there is little to no prior knowledge about the distribution of the data. With this approach, the target value is replaced with the average value of the k data points in the training set that are closest to the data point where it is missing (Singh, Gupta and Sidhu, 2021).

* **Decision Tree:**

The Decision Tree approach is demonstrated with a structure that looks like a tree. To solve classification problems, this strategy makes use of supervised learning. It divides the data into smaller subsets, with the inner subsets (node) representing the dataset characteristics and the outer subsets (branch) representing the outcome. This approach chooses the property with the highest information gain after calculating the sample homogeneity and information gain using entropy. Decision Tree is a rapid, reliable, and easy-to-use tool (Ahdal et al., 2021).

* **Naive Bayes:**

The Naive Bayes method which is based on the Bayes theorem, is a straightforward and effective supervised learning technique. The existence class in NB is independent from other classes, and since training in NB only requires a limited amount of data, this independence is crucial when classifying objects. The Naive Bayes technique is a condensed form of predictive modelling, and it typically uses training datasets with several dimensions (Yousef & Khaled Batiha, 2021)**.**

These are the methodologies which I have been using in this project.

The main research questions that this project is trying to answer are:

RQ 1: Can AI be useful in detecting heart problems using machine learning algorithms?

RQ 2: What is the maximum accuracy machine learning algorithms can give?

The plan for the project is as follows:

1. Find an appropriate dataset
2. Clean the dataset to make it more manageable.
3. Identify the characteristics required for training the machine learning models.
4. Train the machine learning models using KNN, Logistic Regression, Naive Bayes, Decision Tree and Random Forest.
5. Evaluate the results of the trained machine learning models.

# **Literature Review**

As there have been an increase in the availability of data in various fields, one of the fields that can benefits a lot from using this data is the healthcare field. This data can be used to train machine learning models for various use cases in healthcare. There are a number of industry and research initiatives that aim to apply machine learning technology in the healthcare realm in order to improve the lives of patients around the world (Taghizadeh, 2019).

One of these is a research facility run by Stanford University in Palo Alto, California. Dr. Nigam Shah is the director of The Shah Lab, which is a division of the Centre for Research in biomedical informatics at Stanford. To "enable the learning health system”, the Shah Lab researchers apply machine learning and data mining in medical ontologies. It's crucial to highlight that this programme analyses unstructured data as the researchers look at Electronic Health Records that contain longitudinal patient care data (Bhardwaj et al., 2017). According to their research, machine learning is a very practical method for analysing vast volumes of data in order to identify patterns and draw conclusions that would not be obvious otherwise. Additionally, it demonstrates the value of machine learning in the healthcare sector by assisting physicians in making more informed diagnoses and supporting them in differentiating between available treatments. All of these contribute to the Shah Lab's mission, which is to "assist clinical decision making at the point of care." (Bhardwaj et al., 2017)

Founded in 2008, Pathway Genomics gives doctors and their patients precise genetic information to enhance or maintain health and wellness. The business's mobile health applications combine deep learning and artificial intelligence with user genetic data to offer individualised health and wellness advice (CBI Insights, 2018). AliveCor, a different startup in this field and the manufacturer of the portable ECG equipment Kardia Mobile in Silicon Valley, is now wagering that artificial intelligence can aid doctors in monitoring patients' heart issues. With the use of its machine learning algorithms, aberrant ECGs will be automatically flagged, enabling early recognition of typical heart rhythms and assisting in the prevention of strokes (Bhardwaj et al., 2017).

Machine learning can also be used for detecting various medical conditions such as sleep stress level detection (Batabyal et al., 2022), diabetes (Kumar and Pranavi, 2017), bladder pressure detection (Geramipour, Makki and Erfanian, 2015), Alzheimer's Disease detection (Shetty et al., 2022) cancer (Sruthi, 2022) etc.

Sakthimohan et al., 2022 proposes a solution to detect heart problems using parameters such as sex, age, and weight, as well as tests such as cholesterol, blood pressure, and diabetes. This study uses logistic regression to train the machine learning model.

Ouyang, 2022 proposes the study for predicting the heart disease problem using the parameters cholesterol, blood pressure, maximum heart beat, vessels condition, ECG, ST depression, ST segment. This study uses logistic regression to train the machine learning model. Decision tree random forest and support vector machine algorithms.

Khan et al., 2020 proposes the study for finding the heart disease problem using the machine learning and artificial internet of things with parameter using ECG records.

Gavhane et al, 2022 proposes that machine learning algorithms are evolved a lot and hence they use Multi Layered Perceptron (MLP) to detect the heart disease problems.

One of the drawbacks of machine learning technic is over fitting and dataset size (Shen et al., 2020). Overfitting mainly occurs due to using the large number of parameters while training the machine learning model. I try to reduce the impact of over fitting by using less parameters for training.

All the previous studies that I observed do not use more than 2-3 machine learning techniques to predict heart diseases. The parameters used to train machine learning models in the studies available are using different parameters in each of the studies. Some of the studies require additional tests to collect more data for training of the machine learning models. The parameters used in this research are different form the other studies and the reasoning will be explained in the next chapter.

Chicco & Jurman, 2020 analyses a dataset of about 300 patients with heart failures in 2015. This research uses a number of machine learning classifiers to rank the features that correspond to the most significant risk factors as well as predict the patients' survival. Additionally, using conventional biostatistical tests, the researchers perform a different feature ranking analysis and contrast the outcomes with those produced by machine learning algorithms. Both the feature ranking algorithms identify serum creatinine and ejection fraction as the two most relevant features.

Based on all the research papers reviewed, I have decided to train the machine learning model using serum creatinine and ejection fraction. Most of the research papers use features which may not directly affect the results of the machine learning models. By using just two of the features, it makes training of the machine learning models easier.

Most of the research papers reviewed use two or three machine learning models to determine the validity of a prediction. While using three techniques may give a better result, I will be using more machine learning models trained using different techniques to decrease the chance of false positives or false negatives.

# **Methodology**

In this project, the heart disease forecasting is a method developed with the aid of machine learning algorithms that analyses the patient's medical history to track their state of health and feeds the information into the algorithm. Data has been collected using specific criteria, such as age, anaemia, creatinine, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, and smoking. The method of data collection may also differ depending on the gender.

This Research has been carried out by using the medical dataset which is the core of this project and the loaded data will be analysed accordingly. Then, the data will be divided as per the requirement and it will be fetched in the machine learning according to the data fitting process, the data will be separated as per the test set and training set, it will be processed by using the machine learning algorithm to test and find the accuracy of the heart illness prediction method.

Our main objective is to build a machine learning-based heart affiliation prediction model that will be utilized to forecast all heart-related ailments using patient medical records. I can improve its prediction using Logistic Regression, KNN Decision Tree, Naive Bayes, and Random Forest, five machine learning algorithms. This research will be carried out based on supervised learning.

One of the strongest statistical models for determining the likelihood of a particular group or event, such as success or failure, is the logistic regression model (Ahdal et al., 2021). Many predicted variables, which could be digital or class-based, are used in the logistical regression. Logit or the comprehensive study of entropy are other names for logistic regression. Regression in logistics is one of the supervised machine learning techniques for "classification" problems. For binary and linear classification issues that benefit from using linear separation layers, logistics regression is a quicker and more effective solution. The logistics regression model predicts a dependent data attribute by examining the relationship between one or more independent variables and the data value based on historical observation of a data collection.

Both classification and regression are accomplished using the Random Forest algorithm, one of the best for classification (Singh et al., 2022). The Random Forest technique, as its name suggests, primarily comprises a sizable number of distinct decision trees that collaborate. Despite having the same distribution, every tree is unique from the others in some way.

KNN algorithm is a straightforward but effective classification method. Even when there is little to no prior knowledge about the distribution of the data, it frequently applies to classification problems and involves no simplifying assumptions. With this method, the target value is changed to the average value of the k training data points that are closest to the missing data point.

A structure that resembles a tree is used to illustrate the Decision Tree method. Using supervised learning, this technique addresses classification-related issues. Smaller subsets of the data are created, with the inner subsets (node) indicating the dataset's features and the outer subsets (branch) representing the result. After determining the sample homogeneity and information gain using entropy, this method selects the property with the maximum information gain. Decision Tree is a quick, dependable, and user-friendly tool.

The Bayes theorem-based Naive Bayes method is an easy-to-use and powerful supervised learning tool. Since training in NB only needs a small quantity of data, the existence class in NB is independent from other classes. This independence is essential for classifying objects. A compressed version of predictive modelling, the Naive Bayes technique frequently employs training datasets with several dimensions. I have been applying these approaches for this project.

## Implementation Strategy and Issues dealt

The first step in this project's methodology was to get the medical data set from the open source that was accessible on Kaggle. Python was already something I was familiar with, but I've never created a machine learning project in real time. Then, I started getting ready with numpy, pandas, and matplotlib by practicing daily to get knowledge in machine learning, which was challenging when I first started. After 6 months, I was familiar with Logistic Regression, KNN, Naive Bayes, Decision Trees, and Random Forest. Additionally, I have practiced knowledge that helped me complete this project's machine learning component without any issues. For the creation of algorithms, PyCharm and an online course which I studied were used to implement Logistic Regression, Naive Bayes, Decision Trees, and Random Forest. The feature extraction functions were also put into place and evaluated for functionality using debugging.

Particularly, when applied to medical records, machine learning can be a useful tool for predicting the prognosis of each patient exhibiting heart failure symptoms as well as identifying the key clinical traits or risk factors that may contribute to heart failure. Machine learning can be used by scientists for feature rating in addition to clinical prediction. In particular, computational intelligence demonstrates its predictive value when integrated with imaging or applied to medical records.

## Data Collection

In this project, acquiring data from a primary or secondary source is referred to as data collection, and this is a crucial procedure. Medical data set was taken from the secondary open source for my project purpose with the help of Kaggle. And that data source contains the data set of heart based illness records. Age, anemia, creatinine, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, and smoking are only a few of the particular criteria that were used to obtain the data. Depending on the gender, the method of data gathering may also change.

* **Age:**

The age factor is used to calculate the ratio of the health of the patient according to their medical record which will be useful to predict their heart based illness. Likewise there are various ways to predict heart based illness.

* **Anemia:**

When you have anemia, your body doesn't produce enough healthy red blood cells to supply your tissues with enough oxygen. Being anaemic, or having low haemoglobin, can make you feel exhausted and frail. Anemia can have many different forms, each with a unique etiology. This is one of the factors which will be used to predict heart based disease.

* **Creatinine:**

A chemical by-product of creatine is creatinine. The chemical creatine is produced by the body and mostly used to fuel muscles. This is one of the variables that will be taken into account when predicting heart disease.

* **Diabetes:**

A condition where the body's capacity to create or react to the hormone insulin is compromised, leading to improper carbohydrate metabolism and increased blood glucose levels.

* **Ejection Fraction:**

Ejection fraction measures how much blood leaves your heart with each contraction . It is only one of many tests that your doctor might use to find out how well your heart functions.

* **High Blood Pressure:**

Hypertension, another name for high blood pressure, is elevated blood pressure. Depending on your activity, your blood pressure changes throughout the day. A diagnosis of high blood pressure may be made if blood pressure readings are frequently higher than normal (or hypertension).

* **Platelets:**

a significant portion of blood contains a tiny, colourless, disc-shaped cell fragment without a nucleus that is involved in clotting. This is one of the factors that will be taken into account when predicting heart disease.

* **Serum Creatinine:**

A waste product from your muscles, creatinine is found in your blood. Your urine is filtered by healthy kidneys to remove creatinine from your blood. Based on a blood test that gauges the amount of creatinine in your blood, your serum creatinine level is determined.

* **Serum Sodium:**

In contrast to total body sodium content, serum sodium concentration measures the amount of sodium in relation to the amount of water in the blood. Patients with hyponatremia or hypernatremia conditions may have lower, normal, or higher total body sodium levels.

* **Smoking:**

Smoking is a habit, which is one of the things that makes the blood poisonous and causes lung and heart problems.

* **Gender:**

One of the most important factors used to forecast heart-related illnesses is the health difference between genders, which is calculated using this factor.

## Data Analyzation

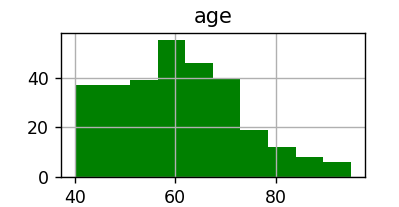
Data Analyzation is processed using the following data such as age, anemia, creatinine, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking and death event.

Above stated factors will be explained in detail accordingly.

* **Age:**

The patient's age is taken into account when calculating their health status according to their medical history, which can be used to anticipate heart-related illnesses. Similar to that, there are numerous approaches to forecast heart-related sickness.

Using a histogram shown in Figure(1), the age factor has been examined. According to the data, this age component was differentiated by age from 0 to 100, with the results showing that 20% of people experience heart disease by the time they are 50 years old, and 40% or more experience heart disease by the time they are 65 to 75% of their life.



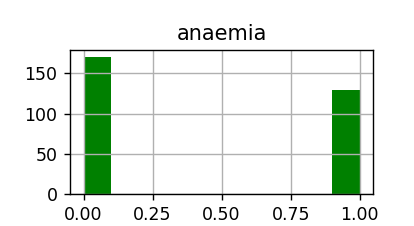
Figure(1): Age differences in the dataset.

* **Anemia:**

Anemia is a condition in which your body produces insufficient healthy red blood cells to adequately oxygenate your tissues.

You might feel worn out and weak if you have low haemoglobin, or are anaemic. Different types of anemia can exist, and each one has a distinct origin. One of the variables that will be taken into account when predicting heart-related disease is this.

According to the data in the histogram shown in Figure(2), the percentage of anemia-conditioned people was recorded slightly above to one hundred people and the percentage of non-anaemics was recorded as more than one hundred and fifty people or higher.

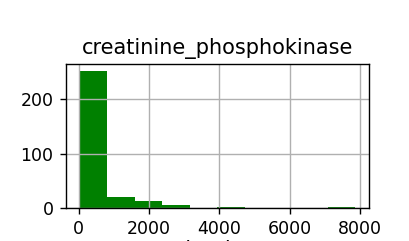


Figure(2): Anaemia condition in the patient in dataset.

* **Creatinine:**

Creatinine is one of creatine's chemical by products. Creatine is a substance that the body produces and mostly uses as a fuel for muscles. When forecasting heart disease, this is one of the variables that will be considered.

As per the histogram shown in Figure(3), the data of creatinine was displayed at its peak for 200 people between zero to thousand and less than 50 people having creatinine phosphokinase over 1500.

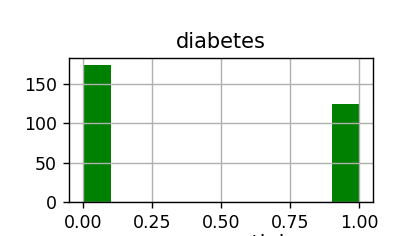


Figure(3): Creatinine phosphokinase differences in the dataset.

* **Diabetes:**

A condition in which the body's ability to produce or respond to the hormone insulin is hampered, causing improper carbohydrate metabolism and elevated blood glucose levels.

In the histogram shown in Figure(4), person affected was raging between 100 to 130 people and the non-affected people were slightly above 150 people as the recorded data.



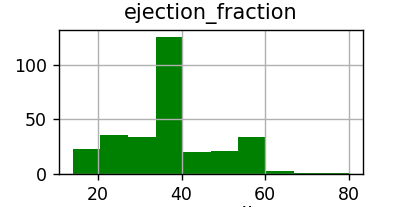
Figure(4): Diabetes condition in the patient in dataset.

* **Ejection Fraction:**

The amount of blood that leaves your heart with each contraction is measured by the ejection fraction.

Your doctor may perform a variety of tests to determine how well your heart is working, and this is just one of them.

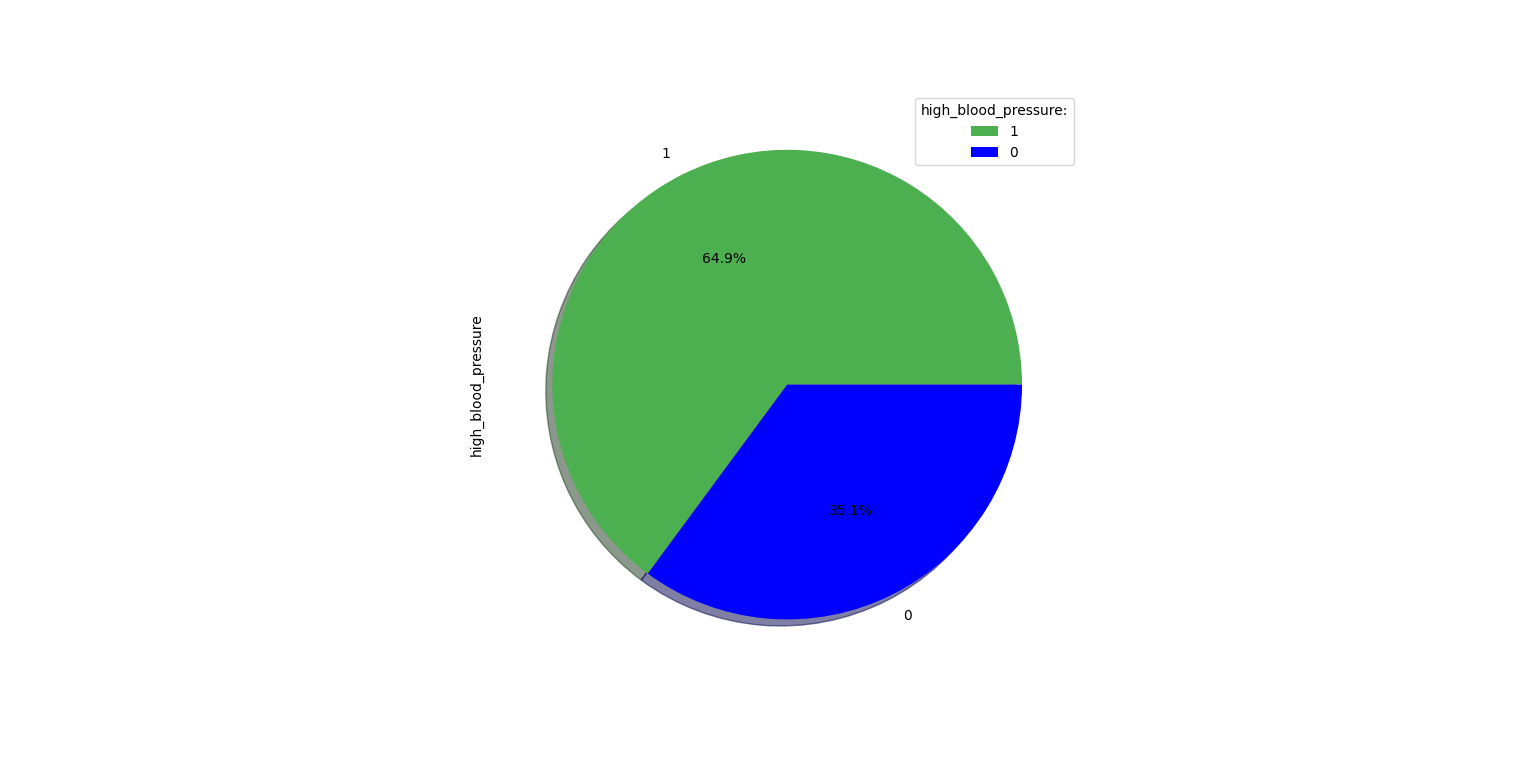
As per the data set of ejection fraction shown in Figure(5), more than 100 people out of 300 people have the ejection fraction up to the level of 30 to 40 and the rest of the people levels are differentiated but they are below 40 people in all range



Figure(5): Ejection fraction differences in the dataset.

* **High Blood Pressure:**

Elevated blood pressure is known as hypertension, or high blood pressure. Your blood pressure varies throughout the day based on your activity. If blood pressure readings are frequently higher than normal, high blood pressure may be diagnosed. In this pie chart shown in Figure(6), a person who got high blood pressure is stated as 64.9% of people are having the high blood pressure and the 35.1% of people are not having the high blood pressure as in data set.



Figure(6): High blood pressure condition in the patient in dataset.

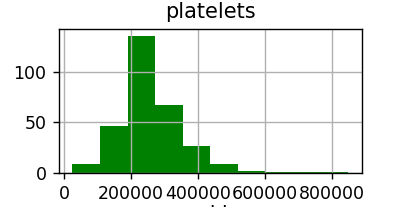
* **Platelets:**

The tiny, colourless, disc-shaped cell fragment without a nucleus that is involved in clotting is present in significant amounts in blood.

This is one of the variables that will be considered when making a heart disease prediction.

Graph in Figure(7) represents the platelet structure according to the medical data set; the platelet count has been up to its peak when there is more than 300000 count of platelets in the histogram.

This shows the recorded data’s peak performance in predicting the heart based illness using platelet count.



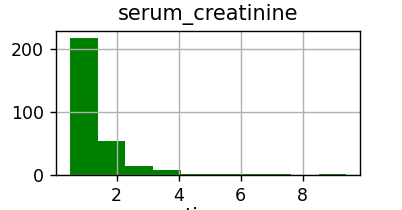
Figure(7): Platelets differences in the dataset.

* **Serum Creatinine:**

Your blood contains creatinine, a by-product of your muscles. Healthy kidneys filter your urine to remove creatinine from your blood.

Your serum creatinine level is computed using a blood test that measures the amount of creatinine in your blood.

According to the graph in Figure(8), the serum creatinine level is elevated above 200 people when it is less than 2 units.



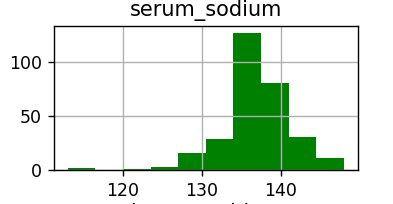
Figure(8): Serum creatinine differences in the dataset.

* **Serum Sodium:**

Serum sodium concentration, as opposed to total body sodium content, gauges how much sodium there is in relation to how much water is present in the blood.

Total body sodium levels may be lower, normal, or higher in patients with hyponatremia or hypernatremia conditions.

With the aid of visualization, the histogram in Figure(9) indicates that the peak level of serum sodium in the human body is between 130 and 140, the other levels are from minor group of people less than 50 people.

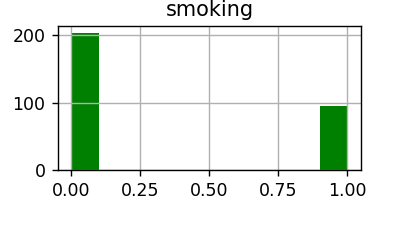


Figure(9): Serum sodium differences in the dataset.

* **Smoking:**

One of the habits that contributes to the blood becoming poisonous and to lung and heart issues is smoking.

Bar Graph in Figure(10) states the recorded data about smoking which shows that 200 percent of the people are not smoking and people who are smoking is less than 100.

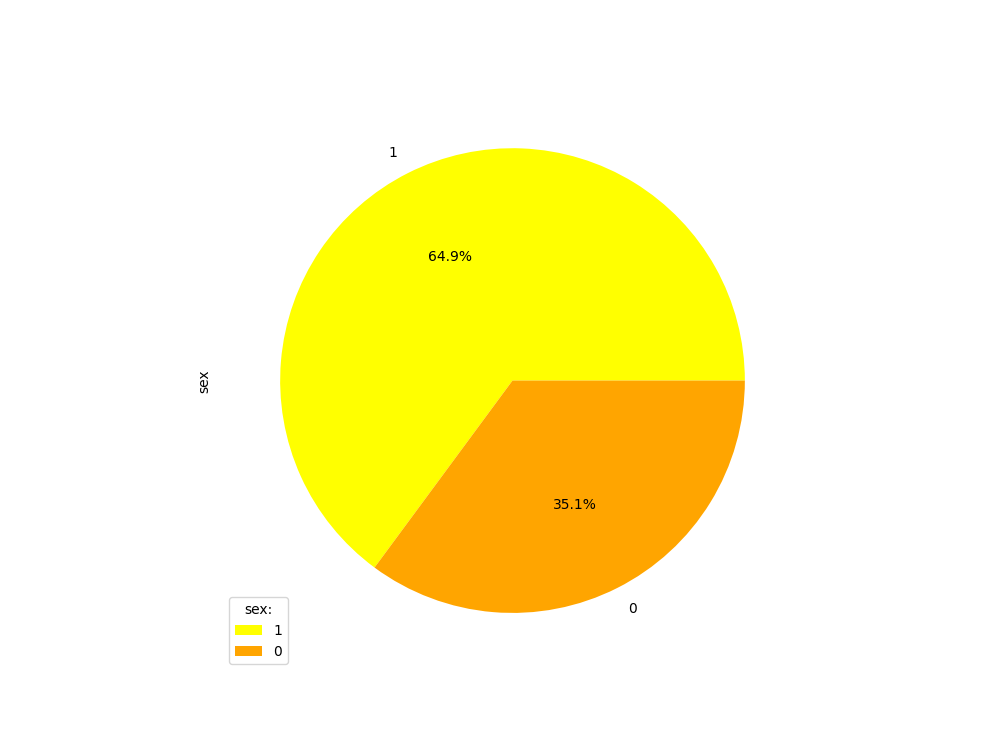


Figure(10): Whether the people smoke or not in dataset.

* **Gender:**

The gender-based health disparity, which is calculated using this factor, is one of the most significant variables used to predict heart-related illnesses.

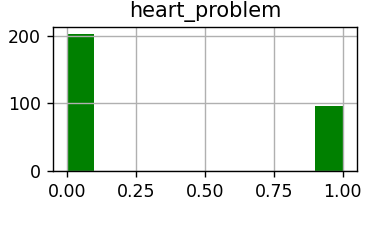
This pie chart in Figure(11) shows that the 64.9% men are affected by the heart based illness and the women 35.1% are affected which is less percentage than men as in this dataset.



Figure(11): How many percentages of men and women present in dataset.

* **Heart Problem :**

This graph in Figure(12) shows data on those who develop heart-related illnesses, showing that out of 300 percent of people, 100 percent are at risk.



Figure(12): Whether the people have heart problem or not in dataset.

**Comparison :**

* Comparison between ejection fraction and heart problems are based on the medical data set which has been stated in two different graphs such as histogram and bar plot to show the variance.
* In the ejection fraction the heart-based illness ratio has been recorded at the level less than 150 out of 300 when the level was less than 35.
* In addition, according to the barplot in Figure(13), low heart-related disease ratio indicates that, in comparison to those whose ejection fractions are impacted, only 50% of the population is susceptible to heart-related illnesses.

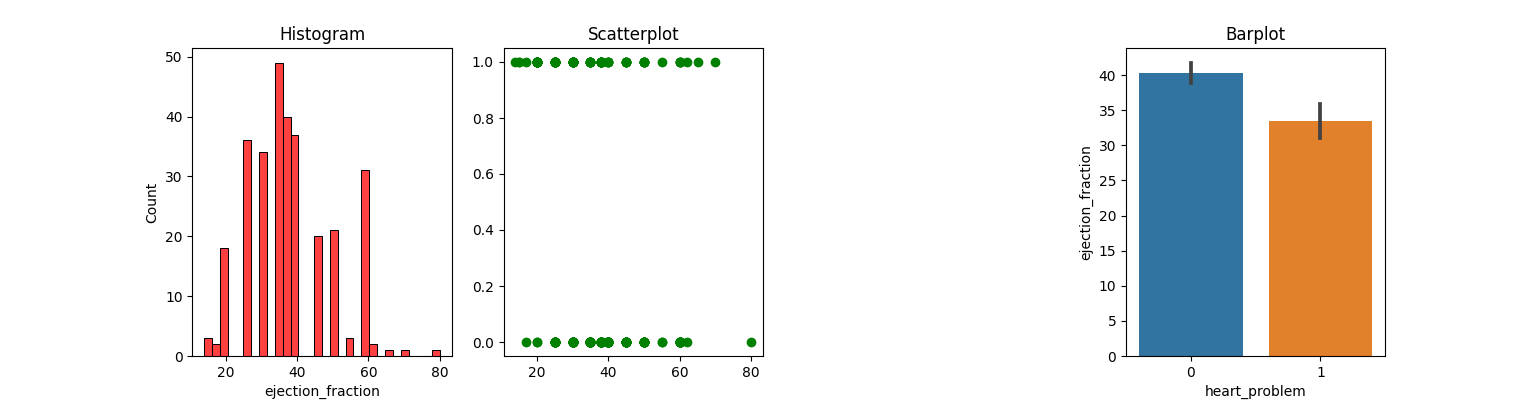


Figure (13): Represents people having heart problem when the ejection fraction is at these level in the dataset.

* Next comparison is between serum creatinine and heart problems. According to the histogram in Figure(14), if the data for creatinine were less than 1.5 level the heart problem occurrence is low and reached a peak of when the serum creatinine level is above 1.5 In this comparison topic I have compared and explained about the relation and difference between ejection fraction and heart problems and also with the creatinine and heart problems to show the difference in the algorithm and its accuracy.

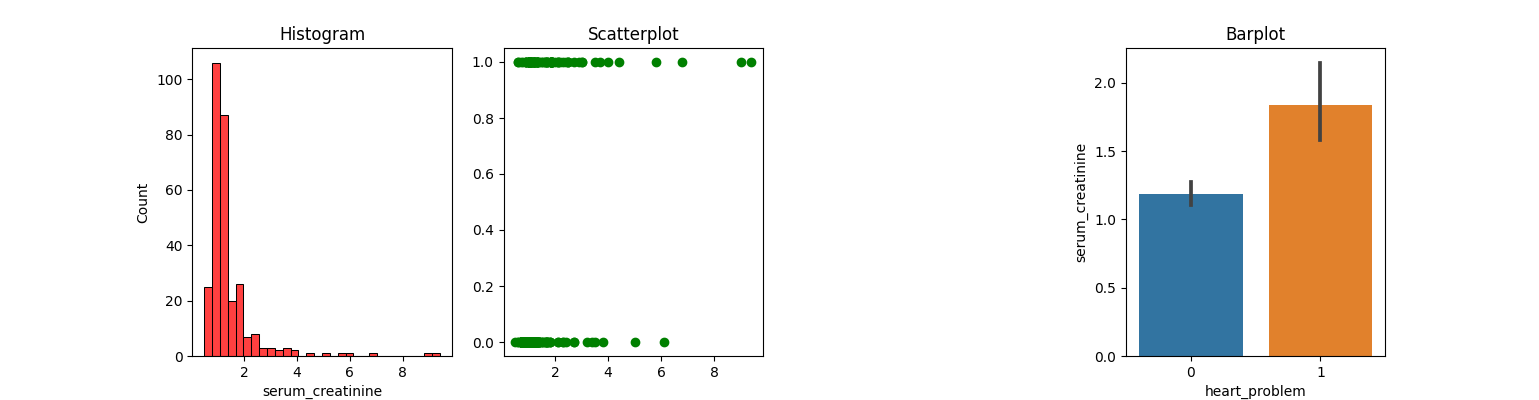


Figure (14): Represents people having heart problem when the serum creatinine is at these level in the dataset.

Age frequency graph in Figure(15) shows that risk of heart illness is at the peak at the age of 60, after the age of 45 and above the heart illness rate or the chance of getting the heart problem is more

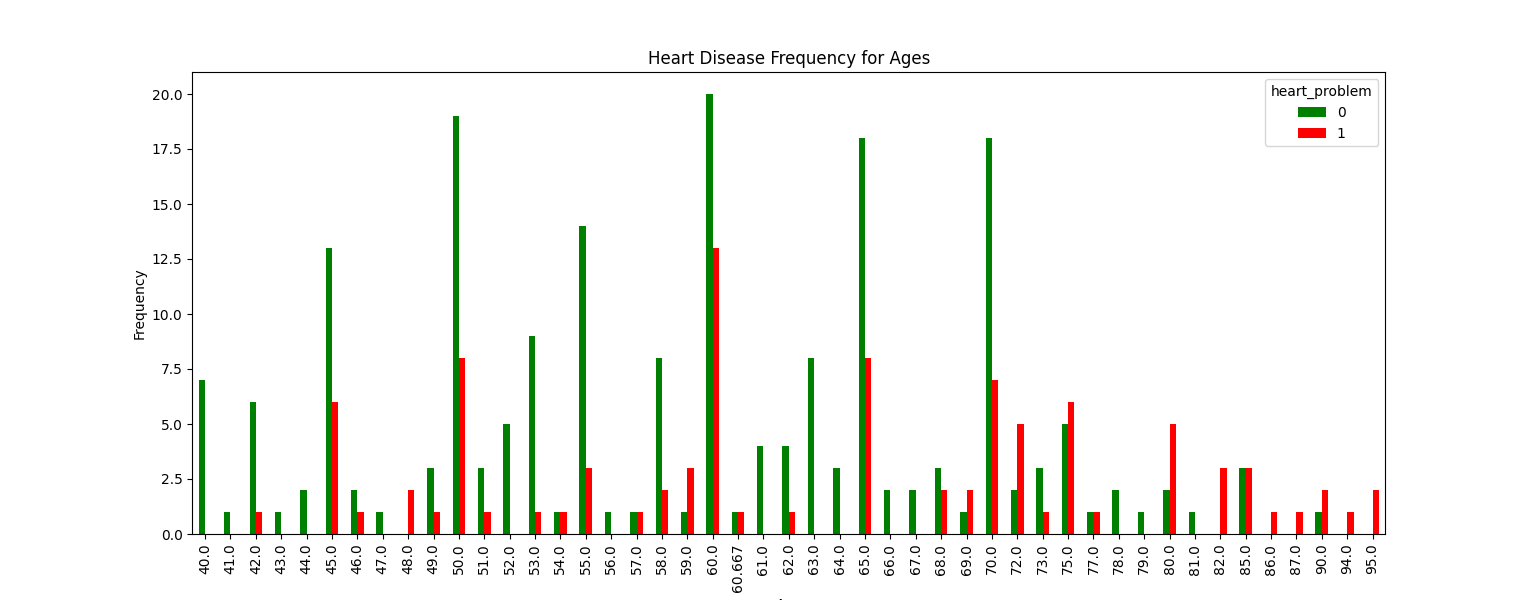


Figure (15): Represents whether the people having heart problem or not in each age in the dataset.

## Machine Learning Algorithm

Artificial intelligence (AI) in the form of machine learning trains computers to think like people do by learning from and improving upon prior experiences. It uses little to no human intervention and operates by examining data and recognising patterns.

Decision-making that is made more quickly is made possible by machine learning, which enables businesses to process and analyse data more quickly than ever before.

Depending on your use case, a variety of algorithms are available in machine learning, the majority of which fall into one of three categories such as supervised learning, unsupervised learning, and reinforcement learning. In these three types of algorithms, I am using supervised learning and it will be explained in detail.

## Supervised Learning

In supervised learning, the machine must learn how to translate the input to the desired output by being given both the input and the desired output. In order to achieve this, the machine is trained on a set of example inputs and outputs that are statistically representative.

The system can be better controlled with this method, and accuracy normally rises with the quantity of labelled samples or patterns offered. On the other hand, the instances or patterns that will be utilized in training must be labelled by competent individuals.

There are restrictions on scalability with this method, and it can be quite labour- and time-intensive.

Real-world applications of supervised learning have been quite successful. It is used in practically all fields, including text and web fields. Inductive learning or classification are other names for supervised learning in machine learning. This kind of learning is comparable to how people acquire new information by drawing on their past experiences to enhance their performance on tasks that they would encounter in the real world.

Machine learning, on the other hand, learns from data that have been collected in the past and in certain real-world applications represent prior experiences because machines do not have "experiences."

Several different supervised learning tasks exist. This learning paradigm, which has been the subject of machine learning research, is arguably the most popular one in everyday life. These supervised learning methods are described in this chapter.

In this supervised learning, there are two main types of supervised learning such as classification type and regression type. In that I am using classification type because it helps us to classify whether patients have heart problems are not by using this type of learning.

**Regression:**

Finding correlations between dependent and independent variables is the process of regression. It aids in the forecast of continuous variables like market trends, house values, and other things.

Finding the mapping function to transfer the input variable to the continuous output variable is the job of the regression algorithm (*Regression analysis in machine learning,* no date). In that there several types of regression algorithms such as,

* Simple Linear Regression
* Multiple Linear Regression
* Polynomial Regression
* Support Vector Regression
* Decision Tree Regression
* Random Forest Regression

**Classification :**

Finding a function to divide the dataset into classes based on several parameters is the process of classification. In classification, data is divided into various classes by a computer programme that has been trained on the training dataset.

Finding the mapping function to convert the input to the discrete output is the goal of the classification algorithm (*Classification algorithm in Machine Learning ,*no date). Machine Learning Classification Algorithm Types: The following types of classification algorithms are further subdivided as,

* K-Nearest Neighbor
* Support Vector Machines
* Kernel SVM
* Logistic Regression
* Decision Tree Classification
* Naive Bayes
* Classification with Random Forests

**Comparison between Classification and Regression:**

Supervised learning methods include regression and classification techniques. Both algorithms employ labelled datasets and are used for prediction in machine learning. However, how they are applied to various machine learning issues is where they diverge from one another.

The main distinction between classification and regression algorithms is that classification algorithms are used to predict or classify discrete values such as Male or Female, True or False, Spam or Not Spam, etc. while regression algorithms are used to predict/classify continuous values such as price, salary, age, etc.

In that I am using classification type and in classification type I am using K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Decision Tree Classification, Decision Tree Classification, Naive Bayes and Classification with Random Forest (*Regression vs classification in Machine Learning* no date).

**Methods commonly used in all the five algorithms**

**Feature Scaling:**

A technique for normalizing the variety of independent variables or features in data is called feature scaling. It is typically carried out during the data pre-processing step and is sometimes referred to as data normalization in the context of data processing.

**Confusion Matrix :**

* A table called a confusion matrix is used to describe how well a classification system performs. The output of a classification algorithm is shown and summarized in a confusion matrix.
* Confusion Matrix gives you values that show the differences between actual and predicted values. It has a table-like shape and measures how well our machine learning classification model performs.
* It gives you four values which are true positive , true negative, false positive, false negative. Using this value the accuracy of the algorithm is obtained.

## Machine Learning Algorithms Used

**Logistic Regression:**

One of the most well-known machine learning algorithms that falls within supervised learning techniques is logistic regression. It can be applied to Classification and Regression issues, but is primarily utilized for Classification issues.

With the aid of independent factors, categorical dependent variables are predicted using logistic regression. A Logistic Regression problem's outcome can only fall between 0 and 1.

When determining the probability between two classes is necessary, logistic regression can be utilized. such as true or false, 0 or 1, whether it will rain today, etc.

Maximum Likelihood estimate serves as the foundation for logistic regression. This estimation suggests that the observed data should be the most likely. In logistic regression, the weighted sum of the inputs is passed through an activation function that can transfer values between 0 and 1. Such an activation function is referred to as a sigmoid function, and the resulting curve is known as an S-curve or sigmoid curve (*Logistic regression in machine learning,* no date).

**Uses:**

* Using a given set of independent factors, logistic regression is used to predict the categorical dependent variable.
* Classification issues are solved using logistic regression.
* We forecast the values of categorical variables using logistic regression.
* In Logistic Regression, we discover the S-curve that allows us to categorize the sample data.
* For accuracy estimation, the maximum likelihood estimation method is applied.
* There is no requirement for a linear relationship between the dependent and independent variables in logistic regression.

**Assumptions for Logistic Regression:**

* The nature of the dependent variable must be categorical.
* There shouldn't be any multi-collinearity in the independent variable.

**Logistic Regression Equation:**

The linear regression equation yields the logistic regression equation. The following are the mathematical steps to obtain Logistic Regression equations:

* We are aware that the equation for a straight line is:

**y = b0 + b1x1+b2x2+b3x3+...+bnxn**

* Let's divide the preceding equation by (1-y) because y in Logistic Regression can only be between 0 and 1 in order to account for this:

**y/1-y; 0 for y = 0, and infinity for y = 1**

* However, we require a range between -[infinity] and +[infinity]. If we take the equation's logarithm, it becomes:

**log[y/1-y] = b0 + b1x1 + b2x2 + b3x3 + … + bnxn**

* The last equation for logistic regression is the one mentioned above.

**Type of Logistic Regression:**

Three different types of logistic regression can be distinguished based on the categories:

* **Binomial:** 0 or 1, Pass or Fail, etc., are the only two conceivable types of dependent variables in a binomial logistic regression.
* **Multinomial:** In multinomial logistic regression, the dependent variable may be one of three or more possible unordered kinds, such as "cat," "dogs," or "sheep."
* **Ordinal:** In ordinal logistic regression, the dependent variables can be categorized into one of three potential ordered classes, such as "low," "Medium," or "High."

In these types we are using the **binomial** type because the output values are either the patient will have heart problem or not.

**K-Nearest Neighbors Algorithm:**

The K-Nearest Neighbor is one of the simplest algorithms in supervised machine learning. The technique employs proximity to classify or anticipate how a single data point will be grouped. Since K-NN is a non-paramagnetic algorithm, it makes no assumptions about the underlying data. First knn Choose the neighbor with the Kth largest number then Calculate the Euclidean distance between K neighbors using the calculated Euclidean distance, select the K neighbors who are closest to you and assign the new data point to the category where the neighbor count is at its highest(*K-Nearest Neighbor(KNN) algorithm for Machine Learning* ,no date).

**Types of distance measures in KNN**

* **Euclidean distance**

The Euclidean distance measures the separation between two vectors with real values. When determining the distance between two rows of data with numerical values, such as floating point or integer values, you are most likely to use the Euclidean distance.

* **Manhattan distance**

The Manhattan distance, also known as the taxicab distance or city block distance, is the simplest way to determine the distance between two points. If you start at one location and travel to another, the distance between those two points is what is meant by the Manhattan distance. More mathematically speaking, we can say that it determines the absolute value between two points. Since we did not take a diagonal route, we determined the distance using the original path exactly.

* **Minkowski distance**

It is the generalized form of Euclidean and Manhattan distance because we take both the distance technique and the new technique for finding the distance between vectors.

* **Hamming distance**

Hamming distance is the method of measuring distance is more distinctive than the one we learned from others; hamming is typically employed in text processing or when using a Boolean vector.

In these four distance measuring types we are using the **Minkowski distance** which is a distance measured between two points in N-dimensional space. It is basically a generalization of Euclidean distance and the Manhattan distance(*K-Nearest Neighbor(KNN) algorithm for Machine Learning* ,no date).

**Equation of Minkowski distance**

**||x1 - x2 || = (i=1(SIGMA)^d |X1i-X2i|^p)^1/p**

**Advantages and Disadvantages:**

Similar to any machine learning method, k-NN has advantages and disadvantages. It might or might not be the best option, depending on the project and application.

**Advantages:**

* It is one of the first classifiers that a novice data scientist will learn due to its ease of use and accuracy.
* Easy to adjust: Since all training data is saved in memory, the algorithm can easily adapt when fresh training samples are introduced.
* Few hyperparameters: Compared to other machine learning algorithms, KNN just needs a k value and a distance metric.

**Disadvantages:**

* Lacks scalability: KNN is a lazy algorithm, which means that it uses more memory and data storage than other classifiers. Both in terms of time and money, this can be expensive. Business costs will increase with additional memory and storage, and processing more data may take longer. Different data structures, like the Ball-Tree, have been developed to alleviate the computational inefficiencies; nevertheless, depending on the business challenge.

The KNN method frequently experiences the dimensionality curse, which causes it to underperform when dealing with high-dimensional data inputs. The peaking phenomenon describes a situation in which, when the algorithm reaches the optimal number of features, adding more features causes a rise in classification mistakes, particularly when the sample size is less(*K-Nearest Neighbor(KNN) algorithm for Machine Learning* ,no date).

**Naive Bayes:**

Naive Bayes is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature.

Probability is typically represented by the letter P when it comes to probability calculations. Following are some scenarios where this would be likely: The likelihood of receiving two heads is 1/4.

Naive Bayes works well for multi-class prediction issues. It can outperform other models and needs a lot less training data if its assumption about the independence of features is correct. For categorical input variables as opposed to numerical variables, Naive Bayes is more appropriate (Vadapalli, 2022).

**Equation of naive bayes:**

**P(A/B) = (P(B/A) P(A))/P(B)**

**P(A/B)** is a posterior probability - probability of hypothesis A on observed event B**.**

**P(B/A)** is likelihood probability - probability of the evidence given that the probability of a hypothesis(Tutorialforbeginner.com, *Naive Bayes classifier algorithm in machine learning,* no date).

**Types of Naives Bayes Model:**

There are three different kinds of naive bayes models, and they are as follows:

* **Gaussian:** The Gaussian model presupposes that features are distributed normally. This indicates that the model assumes that predictor values are samples from the Gaussian distribution if they take continuous values rather than discrete ones.
* **Multinomial:** The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It indicates which category a specific document falls under, such as Sports, Politics, Education, etc., and is primarily used to solve document classification issues.
* **Bernoulli:** In contrast to the Multinomial classifier, the Bernoulli classifier uses independent Boolean variables as predictor variables. such as determining whether a word is used or not in a document. For tasks involving document classification, this model is renowned(Tutorialforbeginner.com, *Naive Bayes classifier algorithm in machine learning,* no date).

From the above three types we are using the **Gaussian model.**

**Advantages and Disadvantages:**

**Advantages:**

* One of the quick and simple machine learning algorithms to predict a class of datasets is naive Bayes.
* Both binary and multi-class classifications can be done using it.
* In comparison to other algorithms, it performs well in multi-class predictions.
* It is the most frequently used solution for text classification issues.

**Disadvantages:**

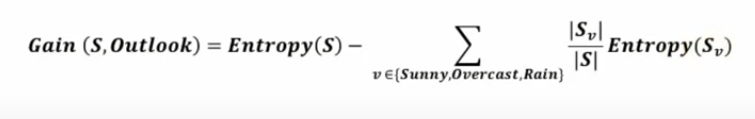
* Naive Bayes cannot discover the relationship between features because it presumes that all features are either independent or unrelated(Tutorialforbeginner.com, *Naive Bayes classifier algorithm in machine learning,* no date).

**Decision Tree:**

The non-parametric supervised learning algorithm used for classification and regression tasks is the decision tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes.

Decision Trees are a type of supervised machine learning in which the training data is continuously segmented based on a particular parameter, with you describing the input and the corresponding output. Decision nodes and leaves are the two components that can be used to explain the tree. It is a graphical representation of every option for making a choice based on specific circumstances. We attempt to create a condition on the features for each step or node of a classification decision tree in order to fully separate all of the labels or classes present in the dataset(*Decision tree algorithm in Machine Learning,* no date).

**Equation for decision tree:**



**Advantages and Disadvantages:**

**Advantages:**

Due to its many advantages, data scientists frequently start with a decision tree as their algorithm of choice. The decision tree benefits are as follows.

* **Interpretability:** The fact that decision trees are incredibly intuitive and simple to comprehend is one of their most important benefits. Additionally, data scientists and other experts can explain to the stakeholders the model's predictions by displaying the rules implemented by decision trees in a flow chart-like manner. In addition to giving stakeholders detailed information about what is occurring and why, this inspires confidence in them. As a result, it is simpler to understand, visualize, and interpret than other algorithms.
* **Less Data Preparation:** Data preparation is a significant challenge when creating a model involving other algorithms. Decision trees have an advantage in this situation because they follow the 'garbage in, garbage out' principle, which states that the quality of a model's predictions depends on the quality of the data it is fed when it is being trained. Because the decision tree does not require standard data preparation procedures like data normalization/standardization, missing value treatment, outlier capping, etc., it is a "go-to" algorithm for data scientists.
* **Non-Parametric:** Many assumptions must be met for algorithms like linear regression and naive Bayes, among others, to function properly. Since Decision Trees are a non-parametric algorithm—as was previously mentioned—there are no significant presumptions to be met or data distribution to take into account.
* **Versatility:** Another benefit of decision trees is their high degree of versatility, which allows them to serve many purposes beyond just making predictions. This includes its ability to explore data, use it as a benchmark model for speedy comprehension of data quality, etc. Additionally, a decision tree and its variations can solve segmentation problems in addition to classification and regression problems.
* **Non-Linearity:** Complex decision boundaries can be made using decision trees, making it simple to solve non-linear issues. While there are other algorithms that can solve the non-linear problem, the Decision Tree has the advantage of being interpretable(*Decision tree algorithm in Machine Learning,* no date)..

**Disadvantages:**

* These are the disadvantages, Overfitting, Data resampling and feature reduction. Optimization(*Decision tree algorithm in Machine Learning,* no date)..

**Random Forest:**

A supervised machine learning algorithm called a random forest is built from decision tree algorithms. This algorithm is used to forecast behaviour and results in a variety of industries, including banking and e-commerce.

A flexible, user-friendly machine learning algorithm called random forest typically yields excellent results even without hyper-parameter tuning. Due to its versatility and simplicity, it is also one of the most widely used algorithms.

Choose the number N of decision trees you want to build, and the random forest will choose R random data points from the training set and build decision trees associated with those data points(*Machine learning random forest algorithm,* no date).

**Difference Between Decision Tree and Random Forest:**

* While a random forest combines many decision trees, a decision tree combines some decisions. As a result, it is a slow but lengthy process.
* A decision tree, especially a linear one, is quick and works well with large data sets. Thorough training is required for the random forest model(*Machine learning random forest algorithm,* no date).

Figure (16) shows the flow chart for the process undergone to train each of the machine learning models.

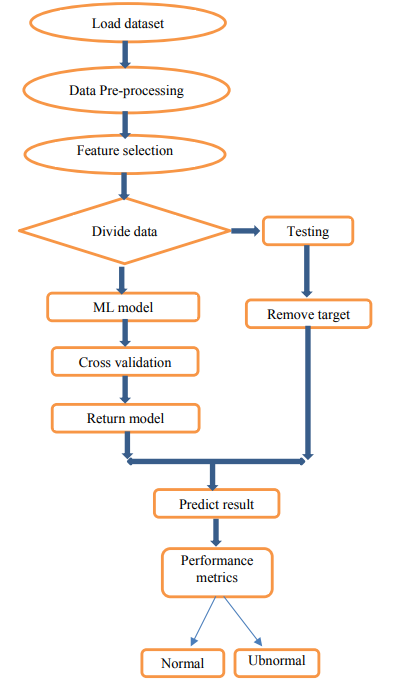


Figure (16) Flow chart representing process of training and implementing the machine learning models.

## Data pre-processing

In order to define the dependent and independent variables, by using the pandas we are reading the dataset from local in other words it is called data fitting. Then in order to define the dependent and independent variables, first analysing the dataset knowing what type of data is available and selecting the specific values from dataset which is **Ejection fraction** and **serum creatinine** as independent variables and **heart problem** as dependent variables. Then separating them as training and test set accordingly.

## Fitting the machine learning algorithms to prepare the models for training

Importing the algorithm from Sklearn library, assigning it to a variable, then calling the function and fitting those two variables. which are normalized using method feature scaling and separated as training set as x training and y training. By taking these values, the algorithms train them by solving the values in each of the machine learning algorithm equations one by one. For better understanding the visualization of training set is shown below which is trained by algorithm with 75% people records and it is randomly shuffled 30%.

**Fitting the logistic regression algorithm**:

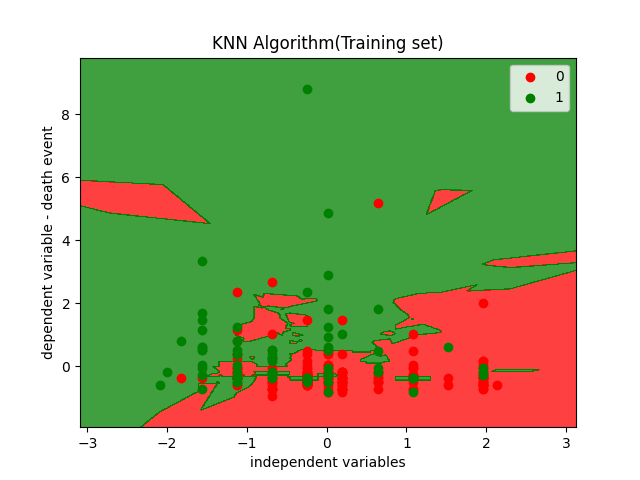
Figure (17) shows the training set for logistic regression. green dots are who doesn’t have heart problem and purple dots are who have the heart problem.



Figure (17): Training result for the selected variables.

**Fitting the KNN algorithm:**

Figure (18) shows the training set for KNN algorithm. The green dots and clusters shows the persons who have heart problem and red dots and clusters represents who does not have the heart problem



Figure(18): Training result for the selected variables

**Fitting the Naive bayes algorithm:**

Figure (19) shows the training set for Naïve bayes algorithm. The green dots shows the persons who have heart problem and yellow dots and represents who does not have the heart problem

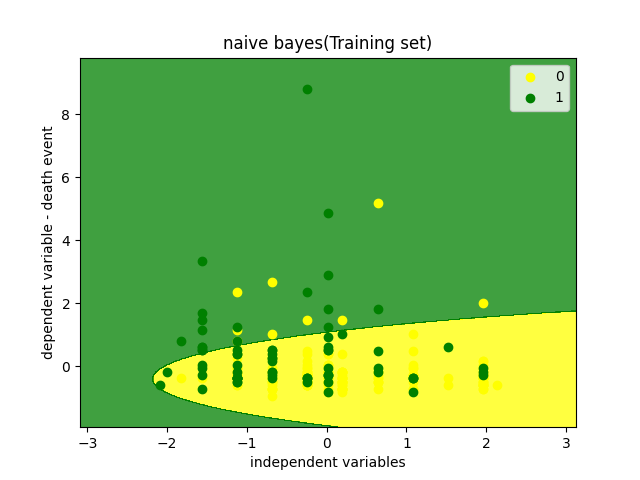


Figure (19): Training result for the selected variables

**Fitting the Decision tree algorithm**:

Figure (20) shows the training set for decision tree algorithm. The orange dots and segments show the persons who have heart problem and blue dots and segments represents who does not have the heart problem

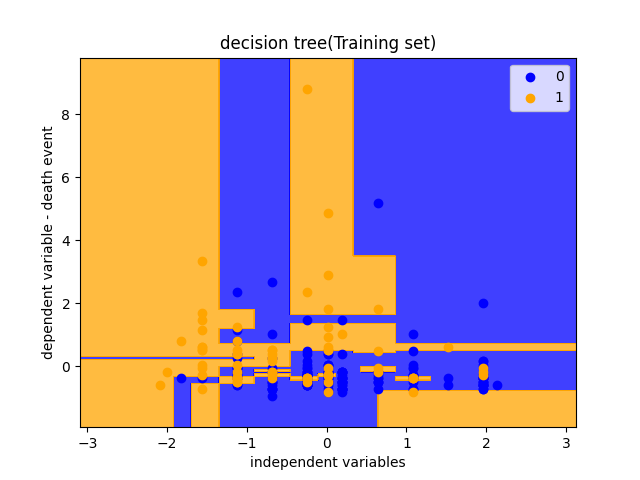
****

Figure (20): Training result for the selected variables

**Fitting the Random Forest algorithm**:

Figure (21) shows the training set for random forest algorithm. The orange dots and segments show the persons who have heart problem and blue dots and segments represents who does not have the heart problem

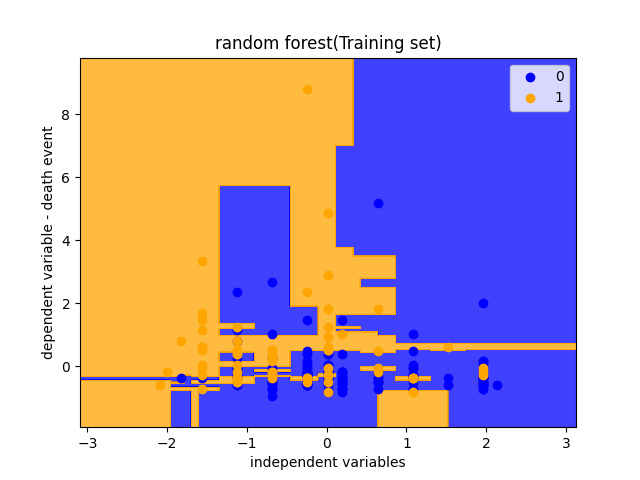


Figure (21): Training result for the selected variables

## Legal, Ethical, Social And Professional Implication

An ethical issue would be that data is collected from individuals in high-income countries may not perform well for individuals in low-income and middle-income settings.

A legal implication would be not providing the system and facilitate equitable use and access, irrespective of age, sex, gender, income, race, ethnicity, sexual orientation, ability or other characteristics protected under human rights codes.

A professional implication would be bias that creeps into the trained model. The developer must ensure that the data collected has a broad scope and makes sure that no bias does not creep into the trained model.

A social implication would be that a patient may develop a fear and negatively impact their mental and physical health when they are informed that they may possibly develop heart problems in the future.

## Commercial Considerations

This research may be used to develop a product that could be used in a medical setting such as clinics and hospitals where it can be used to detect heart problems before they appear in a patient. This can allow the medical professional to prescribe suitable medicines and advice to the patient so that they may prevent a heart problem.

# **Results**

In the work that is being presented, the research has been conducted using these five machine learning algorithms, Logistic Regression Model, K-Nearest Neighbors, Random Forest, Naives Bayes, and Decision Tree. It was created with the aid PyCharm for using python and the dataset in the form of csv file.

a) Projected results in the work that was just presented, we used five models to predict diseases. Each of the five machine learning models focuses on predicting heart disease. With open-source inputs from the Past Medical Dataset, heart diseases are predicted. I got some results variation and slow performance without using feature scaling. Then I used the feature scaling to normalize the independent values. After that I got the accuracy around 70% to 90% and performance speed is increased.

(b) Confusion matrix: The error matrix is another name for the confusion matrix. It is a matrix used to determine how well an algorithm performed. I distinguish between true positive, true negative, false positive, and false negative numbers with the aid of the confusion matrix.

**Equation of confusion matrix:**

Accuracy = (true positive + true negative) / (true positive + true negative + false positive + false negative)

## Predicting the test values and finding the accuracy

The 25% of values which are separated for testing as x test set is given and we are predicting the y test set value, so we are obtaining the y predicted value from the algorithm. By keeping this y predict set and the y test set, using the confusion matrix method we obtain the accuracy or performance of the classification model.

The visualization of test set of logistic regression is shown below in Figure (22).The green dots shows the persons who have heart problem and purple dots represents who does not have the heart problem, the line separates them shows either they have problem or not. The different colour in different part represents the results of getting error values like false positive and false negative.

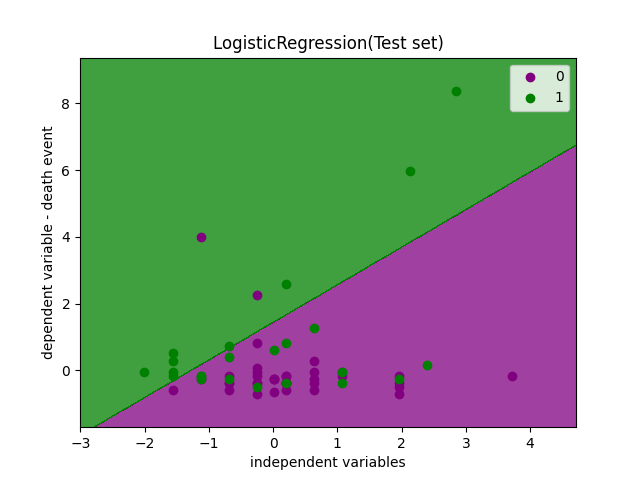
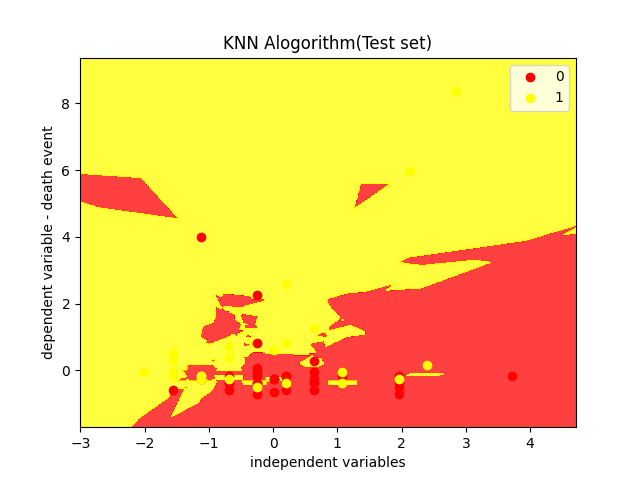


Figure (22): Test set result of dataset using logistic regression.

The visualization of test set of KNN algorithm is shown below in Figure (23).The yellow dots and clusters shows the persons who have heart problem and red dots and clusters represents who does not have the heart problem, the line separates them shows either they have problem or not. The different colour in different part represents the results of getting error values like false positive and false negative.



Figure(23): Test set result of dataset using KNN Algorithm

The visualization of test set of Naïve bayes algorithm is shown below in Figure (24).The yellow dots shows the persons who have heart problem and red dots and represents who does not have the heart problem, the curve separates them shows either they have problem or not. The different colour in different part represents the results of getting error values like false positive and false negative.

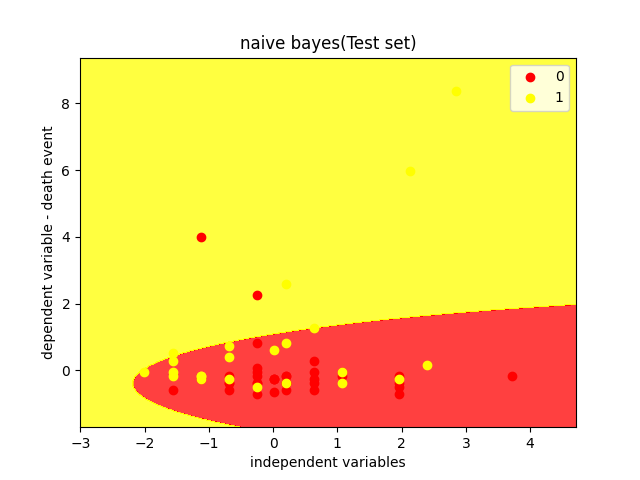


Figure (24): Test set result of dataset using naïve bayes

The visualization of test set of decision tree algorithm is shown below in Figure (25). The orange dots and segments show the persons who have heart problem and blue dots and segments represents who does not have the heart problem, the line separates them shows either they have problem or not. The different colour in different part represents the results of getting error values like false positive and false negative.

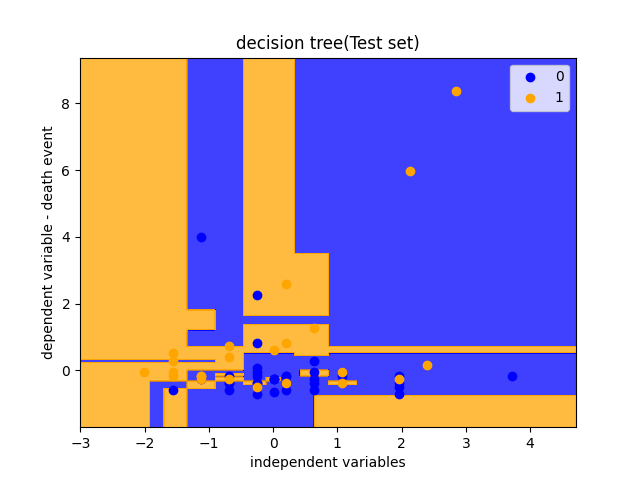


Figure (25): Test set result of dataset using decision tree.

The visualization of test set of random forest algorithm is shown below in Figure (26).The orange dots and segments show the persons who have heart problem and blue dots and segments represents who does not have the heart problem, the line separates them shows either they have problem or not. The different colour in different part represents the results of getting error values like false positive and false negative.

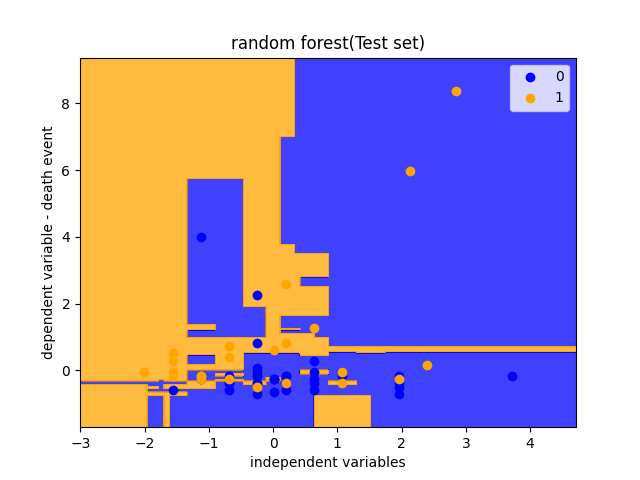


Figure (26): Test set result of dataset using random forest.

On training and testing the machine learning models, the accuracy of each of the models is as shown below:

**Logistic regression accuracy           = 84%**

**K - Nearest Neighbors accuracy     = 79%**

**Naive bayes accuracy                       = 73%**

**Decision tree accuracy                     = 78%**

**Random forest accuracy                 = 82%**

# **Discussion, Evaluation and Conclusion**

## Discussion

I chose this artificial intelligence field in specific because my specialism is artificial intelligence in that I happened to discover that I can use the Artificial intelligence to solve the various amounts of existing problems in that I chose this Heart based illness by using the machine learning algorithm which will be used to predict the heart disease.

I am using the open-source data set in this research method to predict heart-related illnesses. Additionally, I planned to use five machine learning algorithms rather than just two in order to find the best prediction of heart disease because I am using openly available medical data and pre-processing the data before using the machine learning algorithm. Because I am using five machine learning algorithms for better accuracy compared to the existing models, which only use one or two. For the purpose of identifying heart-related illnesses, the serum creatinine and ejection fraction are used.

According to the results without the use of feature scaling the algorithms go very slow and the percentage of accuracy differed however, when feature scaling was used, our accuracy increased and the algorithm’s working speed was strong. Using these five machine learning algorithms, we achieved 84% accuracy in logistic regression, 79% accuracy in KNN, 73% accuracy in Naives Bayes, 78% accuracy in Decision Tree, and 82% accuracy in Random Forest, for this dataset the logistic regression would be the optimal option.

In this I am going to explain why I chose these research question,

RQ 1: Can AI be useful in detecting heart problems using machine learning algorithms?

As per the stated result I found that machine learning algorithms are very much useful in predicting the heart-based illness with the good amount of accuracy than other methods and I used five machine learning algorithms in this research method.

RQ 2: What is the maximum accuracy machine learning algorithms can give?

I found that using multiples of machine learning algorithms, I happened to get the better accuracy to predict the heart-based illness. By using these five machine learning algorithms which help us to get better accuracy than the existing models.

## Evaluation and Conclusion

This Research methods is based on the machine learning algorithm which is from the artificial intelligence for the purpose of predicting heart-based illness. I found numerous number of similar existing projects but with less accuracy and that kind of existing methods has been used only one or two machine learning algorithm which gives the limitation for the existing model so I came up with the idea of using multiple machine learning algorithm such as using five machine learning algorithm to fill the gap between the accuracy of existing models and came close by getting multiple accuracy results but I found that our model also has few limitations but it predicts and produce better accuracy than the existing models.

In this Research, I have done the project based on the multiple algorithm process to get the accuracy in that I used the KNN, Logistic Regression, Naives Bayes, Random Forest and Decision tree to predict the heart based illness using the serum creatinine and ejection fraction to improve the model’s accuracy by using only ejection fraction and serum creatinine instead of using all the features in the dataset.

I am using the open-source data set in this research for the prediction to get the accuracy. One of the limitations of the dataset used is that it is a relatively small dataset. When I was starting the project, I faced an issue that these algorithms were new at the beginning and I was practicing all the machine learning algorithms by self-studying and building models with different datasets and features for all the machine learning algorithms. This helped me to build this project with best of my knowledge. These are the limitations and issues which I faced before and while developing this project and successfully completed the project with the help of the guidance from my mentor.

Future work can be done by using natural language processing for the self-learning process in this project and a better dataset with more patients can be used to train the machine learning models.

# **References**

“Analysis and prediction of heart disease using machine learning and data mining techniques” (2021) *Canadian Journal of Medicine* [Preprint]. Available at: https://doi.org/10.33844/cjm.2021.60500.

“A research of traffic prediction using Deep Learning Techniques” (2019) *International Journal of Innovative Technology and Exploring Engineering*, 8(9S2), pp. 725–728. Available at: https://doi.org/10.35940/ijitee.i1151.0789s219.

“A survey: Heart disease prediction using machine learning techniques” (2021) *Webology* [Preprint]. Available at: https://doi.org/10.29121/web/v18i4/148.

Babu, S. *et al.* (2017) “Heart disease diagnosis using data mining technique,” *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)* [Preprint]. Available at: https://doi.org/10.1109/iceca.2017.8203643.

Carchiolo, V. et al. (2019) “Medical Prescription Classification: A NLP-based approach,” Proceedings of the 2019 Federated Conference on Computer Science and Information Systems [Preprint]. Available at: https://doi.org/10.15439/2019f197.

Chapman, B. *et al.* (2019) “Clinical profiles in acute heart failure: An urgent need for a new approach,” *ESC Heart Failure*, 6(3), pp. 464–474. Available at: https://doi.org/10.1002/ehf2.12439.

Chicco, D. and Jurman, G. (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone - BMC Medical Informatics and decision making, BioMed Central. BioMed Central. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5 (Accessed: November 30, 2022).

Dhanabalan, G., Karutharaja, V. and Sakthimohan, M. (1970) *Realization of resource efficient block RAM based eight bit adder in FPGA: Semantic scholar*, *undefined*. Available at: https://www.semanticscholar.org/paper/Realization-of-Resource-Efficient-Block-RAM-Based-Dhanabalan-Karutharaja/a54a8612af7a5031d8e105f57a69a82563017765 (Accessed: November 30, 2022).

*Dr. M. Sakthimohan‬ -*

Dunn, W.B. *et al.* (2007) “Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate,” *Metabolomics*, 3(4), pp. 413–426. Available at: https://doi.org/10.1007/s11306-007-0063-5.

*Design and enactment of heart attack deduction using IOT measuring device* (no date). Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1427/1/012014 (Accessed: November 30, 2022).

Deep-learning based blood cells classification and initial edge device ... (no date). Available at: https://ieeexplore.ieee.org/document/9935610 (Accessed: November 30, 2022).

Doppala, B.P. et al. (2021) A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on Heart disease dataset - distributed and Parallel databases, SpringerLink. Springer US. Available at: https://link.springer.com/article/10.1007/s10619-021-07329-y (Accessed: November 30, 2022).

Tutorialforbeginner.com (no date) *Naive Bayes classifier algorithm in machine learning*, *www.tutorialforbeginner.com*. Available at: https://tutorialforbeginner.com/naive-bayes-classifier-in-machine-algorithm (Accessed: April 24, 2023).

*Enhanced prediction of heart disease with feature subset selection ...* (no date). Available at: https://www.researchgate.net/profile/Anbarasi-Masilamani/publication/50361284\_Enhanced\_Prediction\_of\_Heart\_Disease\_with\_Feature\_Subset\_Selection\_using\_Genetic\_Algorithm/links/54accada0cf2479c2ee853b1/Enhanced-Prediction-of-Heart-Disease-with-Feature-Subset-Selection-using-Genetic-Algorithm.pdf (Accessed: November 30, 2022).

*Decision tree algorithm in Machine Learning - Javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (Accessed: April 24, 2023).

*K-Nearest Neighbor(KNN) algorithm for Machine Learning - Javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (Accessed: April 22, 2023).

Farooq, K. et al. (2014) “A novel cardiovascular decision support framework for effective clinical risk assessment,” 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE) [Preprint]. Available at: https://doi.org/10.1109/cicare.2014.7007843.

*Heart diseases detection using naive Bayes algorithm - IJISET* (no date). Available at: https://ijiset.com/vol2/v2s9/IJISET\_V2\_I9\_54.pdf (Accessed: November 30, 2022).

Higuera, V. (2022) *Coronary artery disease: Causes, symptoms, and treatments*, *Healthline*. Healthline Media. Available at: https://www.healthline.com/health/coronary-artery-disease (Accessed: November 30, 2022).

Shen, C. *et al.* (2020) “An introduction to deep learning in medical physics: Advantages, potential, and challenges,” *Physics in Medicine & Biology*, 65(5). Available at: https://doi.org/10.1088/1361-6560/ab6f51.

Heart disease prediction using machine learning and deep learning ... (no date). Available at: https://ieeexplore.ieee.org/document/9844406 (Accessed: November 30, 2022).

Henriques, J. et al. (2015) “Prediction of heart failure decompensation events by trend analysis of Telemonitoring Data,” IEEE Journal of Biomedical and Health Informatics, 19(5), pp. 1757–1769. Available at: https://doi.org/10.1109/jbhi.2014.2358715.

Horgan, D. et al. (2019) “Artificial Intelligence: Power for civilisation – and for Better Healthcare,” Public Health Genomics, 22(5-6), pp. 145–161. Available at: https://doi.org/10.1159/000504785.

J., S.K. and S., G. (2019) “Prediction of heart disease using machine learning algorithms,” 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) [Preprint]. Available at: https://doi.org/10.1109/iciict1.2019.8741465.

J, S.K. and S, G. (1970) *Prediction of heart disease using machine learning algorithms.: Semantic scholar*, *undefined*. Available at: https://www.semanticscholar.org/paper/Prediction-of-Heart-Disease-Using-Machine-Learning-SanthanaKrishnan.-G./78291d0478ca3fc9d3b4abccb7b42e34b731ab58 (Accessed: November 30, 2022).

Johnson KW;Torres Soto J;Glicksberg BS;Shameer K;Miotto R;Ali M;Ashley E;Dudley JT; (no date) *Artificial Intelligence in cardiology*, *Journal of the American College of Cardiology*. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/29880128/ (Accessed: November 30, 2022).

Karthigadevi, K., Balamurali, S. and Venkatesulu, M. (2019) “Based on neighbor density estimation technique to improve the quality of service and to detect and prevent the sinkhole attack in Wireless Sensor Network,” *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)* [Preprint]. Available at: https://doi.org/10.1109/incos45849.2019.8951406.

Kaur, G., Sharma, A. and Sharma, A. (2019) “Heart disease prediction using KNN classification approach,” *International Journal of Computer Sciences and Engineering*, 7(5), pp. 416–420. Available at: https://doi.org/10.26438/ijcse/v7i5.416420.

Krishnani, D. et al. (2019) “Prediction of coronary heart disease using supervised machine learning algorithms,” TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON) [Preprint]. Available at: https://doi.org/10.1109/tencon.2019.8929434.

Liu, L. (2018) “Research on logistic regression algorithm of breast cancer diagnose data by machine learning,” *2018 International Conference on Robots & Intelligent System (ICRIS)* [Preprint]. Available at: https://doi.org/10.1109/icris.2018.00049.

MM;, D.S.M.R.V.L.R. (no date) Epidemiology of heart failure with preserved ejection fraction, Nature reviews. Cardiology. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/28492288/ (Accessed: November 30, 2022).

Nauta, J.F. *et al.* (2018) “Markers of left ventricular systolic dysfunction when left ventricular ejection fraction is normal,” *European Journal of Heart Failure*, 20(12), pp. 1636–1638. Available at: https://doi.org/10.1002/ejhf.1326.

Optimization-inspired compact deep compressive sensing | IEEE journals ... (no date). Available at: https://ieeexplore.ieee.org/document/9019857 (Accessed: November 30, 2022).

Patel, R.N. and Barot, M.P. (1970) *Development of methodology for precise diagnosis of ECG by artificial intelligent: Semantic scholar*, *undefined*. Available at: https://www.semanticscholar.org/paper/Development-Of-Methodology-for-Precise-Diagnosis-of-Patel-Barot/65810db24896a78ba987a67e842c9e804380722d (Accessed: November 30, 2022).

*Machine learning random forest algorithm - javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm (Accessed: April 24, 2023).

Prediction of Heart Disease using Machine Learning Algorithms (no date) Choudhary, G. and Singh, S.N. (2020) prediction of heart disease using machine learning algorithms. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, 25-26 April 2019, 197-202. - references - scientific research publishing. Available at: https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/referencespapers.aspx?referenceid=3170553 (Accessed: November 30, 2022).

Pandey, A.C. and Topol, E.J. (2019) “Dispense with supplements for improving heart outcomes,” *Annals of Internal Medicine*, 171(3), p. 216. Available at: https://doi.org/10.7326/m19-1498.

G, A. *et al.* (2022) “Logistic regression technique for prediction of cardiovascular disease,” *Global Transitions Proceedings*, 3(1), pp. 127–130. Available at: https://doi.org/10.1016/j.gltp.2022.04.008.

Raphael, C. *et al.* (2007) “Limitations of the New York Heart Association Functional Classification system and self-reported walking distances in chronic heart failure,” *Heart*, 93(4), pp. 476–482. Available at: https://doi.org/10.1136/hrt.2006.089656.

Reddy, R.L., Reddy, V. and Gupta, G. (1970) *Study of bio-plastics as Green & Sustainable Alternative to plastics: Semantic scholar*, *undefined*. Available at: https://www.semanticscholar.org/paper/Study-of-Bio-plastics-As-Green-%26-Sustainable-to-Reddy-Reddy/9e8cb3ddb5499e2f7226926feed35f87addbaad1 (Accessed: November 30, 2022).

Shahakar, M. (2021) “Heart disease prediction using advanced machine learning algorithms,” *International Journal for Research in Applied Science and Engineering Technology*, 9(VI), pp. 2160–2163. Available at: https://doi.org/10.22214/ijraset.2021.35495.

Sharma, H. (1970) *Prediction of heart disease using Machine Learning Algorithms: A survey*, *Semantic Scholar*. Available at: https://www.semanticscholar.org/paper/Prediction-of-Heart-Disease-using-Machine-Learning-Sharma/d0a5d4b8e8da3ee2a6bf8ac5d44196fb0365cf1c (Accessed: November 30, 2022).

Solanki, A. and Barot, M.P. (1970) *Study of heart disease diagnosis by comparing various classification algorithms*, *Semantic Scholar*. Available at: https://www.semanticscholar.org/paper/Study-of-Heart-Disease-Diagnosis-by-Comparing-Solanki-Barot/f7f3bf062aec1fcc049f526219f9a1c6a719c7f9 (Accessed: November 30, 2022).

*Sci-Hub | improvements in accurate detection of cardiac abnormalities ...* (no date). Available at: https://sci-hub.se/10.1109/ACCESS.2020.2965396 (Accessed: November 30, 2022).

Shah, D., Patel, S. and Bharti, S.K. (2020) “Heart disease prediction using Machine Learning Techniques,” *SN Computer Science*, 1(6). Available at: https://doi.org/10.1007/s42979-020-00365-y.

VL;, R. (no date) Epidemiology of Heart Failure: A contemporary perspective, Circulation research. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/33983838/ (Accessed: November 30, 2022).

Xu, S. et al. (2017) “Cardiovascular risk prediction method based on CFS subset evaluation and Random Forest Classification Framework,” 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)( [Preprint]. Available at: https://doi.org/10.1109/icbda.2017.8078813.

Xu, S. *et al.* (1970) *Cardiovascular risk prediction method based on CFS subset evaluation and Random Forest Classification Framework: Semantic scholar*, *undefined*. Available at: https://www.semanticscholar.org/paper/Cardiovascular-risk-prediction-method-based-on-CFS-Xu-Zhang/4f3a24460ebc7e2aac81142aa7f9f4b2b47788d6 (Accessed: November 30, 2022).

Taghizadeh, G. (2019) *Top 5 companies revolutionizing healthcare with Machine Learning*, *VentureRadar*. Available at: https://blog.ventureradar.com/2015/09/30/top-5-companies-revolutionizing-healthcare-with-machine-learning/ (Accessed: March 18, 2023).

Bhardwaj, R., Nambiar, A.R. and Dutta, D. (2017) “A study of machine learning in Healthcare,” *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* [Preprint]. Available at: https://doi.org/10.1109/compsac.2017.164.

*Artificial Intelligence in healthcare | accenture* (2018) *Industry Market Mapping Analysis*. Available at: https://www.accenture.com/\_acnmedia/pdf-49/accenture-health-artificial-intelligence.pdf?tracker\_id=1581887150757 (Accessed: March 18, 2023).

A. Batabyal, V. Singh, M. K. Gourisaria and H. Das, "Sleep Stress Level Classification through Machine Learning Algorithms," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 91-96.  
doi: 10.1109/OCIT56763.2022.00027  
  
P. S. Kumar and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, United Arab Emirates, 2017, pp. 508-513.  
  
A. Geramipour, S. Makki and A. Erfanian, "Neural network based forward prediction of bladder pressure using pudendal nerve electrical activity," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 2015, pp. 4745-4748.  
doi: 10.1109/EMBC.2015.7319454  
  
M. Shetty, Deekshitha, M. Bhat and M. Devadiga, "Detection of Alzheimer's Disease Using Machine Learning," 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), Karkala, India, 2022, pp. 117-120.  
doi: 10.1109/AIDE57180.2022.10060433

G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 217-221.  
doi: 10.1109/ICIPTM54933.2022.9754059

M, S. *et al.* (2022) “Forecast of heart sickness using machine learning,” *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* [Preprint]. Available at: <https://doi.org/10.1109/icacite53722.2022.9823777>.

A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319, doi: 10.1109/AEMCSE55572.2022.00071.

Khan, M. *et al.* (2020) “Severe analysis of cardiac disease detection using the wearable device by Artificial Intelligence,” *2020 IEEE International Conference for Innovation in Technology (INOCON)* [Preprint]. Available at: https://doi.org/10.1109/inocon50539.2020.9298388.

A. A. Ahdal, D. Prashar, M. Rakhra and A. Wadhawan, "Machine Learning-Based Heart Patient Scanning, Visualization, and Monitoring," *2021 International Conference on Computing Sciences (ICCS)*, Phagwara, India, 2021, pp. 212-215, doi: 10.1109/ICCS54944.2021.00049.

*Regression vs classification in Machine Learning - Javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/regression-vs-classification-in-machine-learning (Accessed: April 22, 2023).

Yousef, M. and Khaled Batiha, P. (2021) “Heart disease prediction model using naïve Bayes algorithm and Machine Learning Techniques,” *International Journal of Engineering & Technology*, 10(1), p. 46. Available at: https://doi.org/10.14419/ijet.v10i1.31310.

H. Singh, T. Gupta and J. Sidhu, "Prediction of Heart Disease using Machine Learning Techniques," 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 164-169, doi: 10.1109/ICIIP53038.2021.9702625.

Singh, R., Gehlot, A. and Buddhi, D. (2022) “Comparative analysis of different machine learning classifiers for the prediction of chronic diseases,” *Comparative Analysis of Different Machine Learning Classifiers for the Prediction of Chronic Diseases*, pp. 1–380. Available at: https://doi.org/10.13052/rp-9788770227667.

Chicco, D. and Jurman, G. (2020) “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Medical Informatics and Decision Making*, 20(1). Available at: https://doi.org/10.1186/s12911-020-1023-5.

*Logistic regression in machine learning - javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/logistic-regression-in-machine-learning (Accessed: April 22, 2023).

*Regression analysis in machine learning - javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/regression-analysis-in-machine-learning (Accessed: April 22, 2023).

*Classification algorithm in Machine Learning - Javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/classification-algorithm-in-machine-learning (Accessed: April 22, 2023).

Vadapalli, P. (2022) *Naive Bayes explained: Function, Advantages & disadvantages, applications in 2023*, *upGrad blog*. Available at: https://www.upgrad.com/blog/naive-bayes-explained/ (Accessed: April 22, 2023).

# **Appendix**

## Link to dataset

Below is the link to the dataset used in this research.

<https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>

## Code used in the research

# installing and importing lib

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

# loading data

df = pd.read\_csv('D:\heart\_failure\_clinical\_records\_dataset.csv')

df.head()

print(df.head())

df.info()

plt.figure(1,figsize=(10,10))

df['high\_blood\_pressure'].value\_counts().plot.pie(autopct="%1.1f%%",colors = ( "green", "blue"),labels = df['high\_blood\_pressure'], shadow = True)

plt.legend(title = "high\_blood\_pressure:")

plt.show()

plt.figure(2, figsize=(10,10))

df['sex'].value\_counts().plot.pie(autopct="%1.1f%%",colors = ('yellow', "orange"),labels = df['sex'].unique())

plt.legend(title = "sex:")

plt.show()

df.hist(figsize = (10, 10),color='green')

plt.show()

# function

def comparison\_plots(df, variable, target):

# The function takes a dataframe (df) and

# Define figure size.

plt.figure(figsize=(20, 4))

# histogram

plt.subplot(1, 3, 1)

sns.histplot(df[variable], bins=30, color='r')

plt.title('Histogram')

# scatterplot

plt.subplot(1, 3, 2)

plt.scatter(df[variable], df[target], color='g')

plt.title('Scatterplot')

# barplot

plt.subplot(1, 3, 3)

sns.barplot(x=target, y=variable, data=df)

plt.title('Barplot')

return plt.show()

comparison\_plots(df,'serum\_creatinine','heart\_problem')

comparison\_plots(df,'ejection\_fraction','heart\_problem')

pd.crosstab(df.age,df.heart\_problem).plot(kind="bar",figsize=(20,6),color= ['green','red'])

plt.title('Heart Disease Frequency for Ages')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

# extracting independent and dependent variable

x= df.iloc[:,[4,7]].values

# print(x)

y= df.iloc[:,12].values

# print(y)

# splitting the dataset into training and test set

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.25,random\_state=0)

# feature scaling

from sklearn.preprocessing import StandardScaler

st\_x= StandardScaler()

x\_train = st\_x.fit\_transform(x\_train)

x\_test =st\_x.transform(x\_test)

# print("scaled values")

# print(x\_test)

# logisticregersion

# fitting logistic regression to the training set

from sklearn.linear\_model import LogisticRegression

classifier = LogisticRegression()

classifier.fit(x\_train,y\_train)

# precdicting the the test result

y\_pred = classifier.predict(x\_test)

# creating confusion matrix

from sklearn.metrics import confusion\_matrix

cm= confusion\_matrix(y\_test,y\_pred)

print("logistic regression cm")

print(cm)

# v traing set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_train,y\_train

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('purple','green'))(i),label = j )

plt.title('LogisticRegression(Training set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# v test set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_test,y\_test

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha = 0.75,cmap=ListedColormap(('purple','green')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('purple','green'))(i),label = j )

plt.title('LogisticRegression(Test set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# knn

# fitting

from sklearn.neighbors import KNeighborsClassifier

classifier = KNeighborsClassifier(n\_neighbors=5,metric='minkowski', p=2)

classifier.fit(x\_train,y\_train)

# pred the test set result

y\_pred = classifier.predict(x\_test)

# confusion matrix

from sklearn.metrics import confusion\_matrix

cm= confusion\_matrix(y\_test,y\_pred)

print("knn cm")

print(cm)

# knn v training

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_train,y\_train

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha=0.75,cmap=ListedColormap(('red','green')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('red','green'))(i),label = j )

plt.title('KNN Algorithm(Training set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# knn test set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_test,y\_test

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha = 0.75,cmap=ListedColormap(('red','yellow')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('red','yellow'))(i),label = j )

plt.title('KNN Alogorithm(Test set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# naive bayes

from sklearn.naive\_bayes import GaussianNB

classifier = GaussianNB()

classifier.fit(x\_train,y\_train)

# pred

y\_pred = classifier.predict(x\_test)

# confusion matrix

from sklearn.metrics import confusion\_matrix

cm= confusion\_matrix(y\_test,y\_pred)

print("naive bayes cm")

print(cm)

# v training set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_train,y\_train

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha=0.75,cmap=ListedColormap(('yellow','green')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set == j,0],x\_set[y\_set == j,1],color= ListedColormap(('yellow','green'))(i),label = j )

plt.title('naive bayes(Training set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# v test set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_test,y\_test

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha = 0.75,cmap=ListedColormap(('red','yellow')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('red','yellow'))(i),label = j )

plt.title('naive bayes(Test set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# decision tree classifier

from sklearn.tree import DecisionTreeClassifier

classifier = DecisionTreeClassifier(criterion='entropy')

classifier.fit(x\_train,y\_train)

# pred

y\_pred = classifier.predict(x\_test)

# cm

from sklearn.metrics import confusion\_matrix

cm= confusion\_matrix(y\_test,y\_pred)

print("decision tree cm")

print(cm)

# v traing set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_train,y\_train

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha=0.75,cmap=ListedColormap(('blue','orange')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('blue','orange'))(i),label = j )

plt.title('decision tree(Training set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable -heart\_problem')

plt.legend()

plt.show()

# v test set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_test,y\_test

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha = 0.75,cmap=ListedColormap(('blue','orange')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('blue','orange'))(i),label = j )

plt.title('decision tree(Test set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# randomforest

from sklearn.ensemble import RandomForestClassifier

classifier = RandomForestClassifier(n\_estimators = 10, criterion="entropy")

classifier.fit(x\_train,y\_train)

# pred

y\_pred = classifier.predict(x\_test)

# cm

from sklearn.metrics import confusion\_matrix

cm= confusion\_matrix(y\_test,y\_pred)

print("random forest cm")

print(cm)

# v traing set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_train,y\_train

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha=0.75,cmap=ListedColormap(('blue','orange')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('blue','orange'))(i),label = j )

plt.title('random forest(Training set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()

# v test set

from matplotlib.colors import ListedColormap

x\_set, y\_set = x\_test,y\_test

x1,x2 = np.meshgrid(np.arange(start = x\_set[:,0].min()-1,stop=x\_set[:,0].max()+1, step =0.01),np.arange(start= x\_set[:,1].min()-1,stop=x\_set[:,1].max()+1,step=0.01))

plt.contourf(x1,x2,classifier.predict(np.array([x1.ravel(),x2.ravel()]).T).reshape(x1.shape),alpha = 0.75,cmap=ListedColormap(('blue','orange')))

plt.xlim(x1.min(),x1.max())

plt.ylim(x2.min(),x2.max())

for i,j in enumerate(np.unique(y\_set)):

plt.scatter(x\_set[y\_set ==j,0],x\_set[y\_set == j,1],color= ListedColormap(('blue','orange'))(i),label = j )

plt.title('random forest(Test set)')

plt.xlabel('independent variables')

plt.ylabel('dependent variable - heart\_problem')

plt.legend()

plt.show()