**HEART DISEASE DETECTION USING MACHINE LEARNING ALGORITHMS**

**By**

**Sai Dinesh Bellamkonda**

# Abstract

This study makes a significant contribution to developing and assessing a wide range of technical methods in machine learning and suitable algorithms for heart disease diagnosis and its many symptoms. To provide an acceptable analysis of all the procedures that have been carried out, all the information on the project's specifications has been provided in this specific research with the necessary justification. In this research, the numerous symptoms found in the human body have been thoroughly explained with appropriate distinctions to understand the concept of heart disease detection. Various machine-learning technologies have also been used for this goal. The goal and objective have been explained in detail, along with other issues in the healthcare field relating to the development of heart disease, as well as how such research may bring about a revolution in the field of medicine by using an appropriate illustration. The method of diagnosing heart disease through identifying various symptoms in the human body is also briefly explained in this research. It also describes how deep learning techniques and tools will aid this process. This study also illustrates how distinct features can connect with the entire process and the various benefits of machine learning techniques for identifying heart disease.

**Table of Contents**

[Abstract 1](#_Toc120446219)

[CHAPTER 1: INTRODUCTION 1](#_Toc120446220)

[1.1 Introduction 1](#_Toc120446221)

[1.2 Background 1](#_Toc120446222)

[1.3 Problem Statement 2](#_Toc120446223)

[1.4 Aim and Objectives 3](#_Toc120446224)

[1.5 Feasibility 4](#_Toc120446225)

[1.5.1 Benefits of a Feasibility Study 4](#_Toc120446226)

[1.6 Chapters Overview 4](#_Toc120446227)

[CHAPTER 2: LITERATURE REVIEW 5](#_Toc120446228)

[2.1 Machine learning concept for heart disease detection 5](#_Toc120446229)

[2.2 Process of heart disease detection using machine learning 7](#_Toc120446230)

[2.3 Benefits of machine learning in detecting heart disease 10](#_Toc120446231)

[2.4 Issues in the detection process of heart disease using machine learning 11](#_Toc120446232)

[2.5 Literature Gap 12](#_Toc120446233)

[2.6 Research Design 13](#_Toc120446234)

[2.7 Research Strategy 13](#_Toc120446235)

[CHAPTER 3: METHODOLOGY 15](#_Toc120446236)

[3.1 Data collection 16](#_Toc120446237)

[3.2 Data Pre-Processing 16](#_Toc120446238)

[3.3 Machine Learning Algorithms: 16](#_Toc120446239)

[3.3.5 Hybrid Model 20](#_Toc120446240)

[3.4 Issues 20](#_Toc120446241)

[3.4.1 Ethical Issues 20](#_Toc120446242)

[3.4.2 Legal issues 20](#_Toc120446243)

[3.4.3 Professional issues 20](#_Toc120446244)

[3.4.4 Social Issues 20](#_Toc120446245)

[CHAPTER 4: RESULTS AND DISCUSSION 22](#_Toc120446246)

[4.1 Experimental setup 22](#_Toc120446247)

[4.2 Data Import 22](#_Toc120446248)

[4.3 Data Spliting and Scaling 25](#_Toc120446249)

[4.4 Models Performance 25](#_Toc120446250)

[4.5 Hybrid Model Performance 27](#_Toc120446251)

[CHAPTER 5: CONCLUSIONS 28](#_Toc120446252)

[5.1 Evaluation 28](#_Toc120446253)

[5.2 Summary of achievements 29](#_Toc120446254)

[5.3 Reflection 29](#_Toc120446255)

[5.4 Recomendations 30](#_Toc120446256)

[5.5 Future work 31](#_Toc120446257)

[References 32](#_Toc120446258)

**List of Figures**

[Figure 2.2.1: Detection of Heart Diseases 11](#_Toc118500931)

[Figure 2.3.1: Early Detection of Heart Diseases 14](#_Toc118500932)

[Figure 4.2.1: Data import process 22](#_Toc118500933)

[Figure 4.2.2: Initial data exploration 23](#_Toc118500934)

[Figure 4.2.3: Chest pain type distribution 23](#_Toc118500935)

[Figure 4.2.4: Fasting blood sugar distribution 24](#_Toc118500936)

[Figure 4.2.5: Exercise induces angina distribution 25](#_Toc118500937)

[Figure 4.2.6: Heart disease distribution 25](#_Toc118500938)

[Figure 4.2.7: Age column distribution 26](#_Toc118500939)

[Figure 4.2.8: Distribution of heart disease 27](#_Toc118500940)

[Figure 4.2.9: Scatter plot 28](#_Toc118500941)

[Figure 4.2.10: Correlation Map 29](#_Toc118500942)

[Figure 4.3.1: Dropping unnecessary values 30](#_Toc118500943)

[Figure 4.3.3: K nearest algorithm 31](#_Toc118500944)

[Figure 4.3.4: SVM 31](#_Toc118500945)

[Figure 4.3.5: Naive Bayes 32](#_Toc118500946)

[Figure 4.3.6: Decision Tree 32](#_Toc118500947)

[Figure 4.3.7: Random Forest 33](#_Toc118500948)

[Figure 4.3.8: Gradient Boosting 33](#_Toc118500949)

[Figure 4.3.9: Accuracy comparison 34](#_Toc118500950)

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction

The heart is a particular muscular organ that performs blood into the body and is central to the cardiovascular system. According to a survey by the World Health Organisation, 17.5 million total Global death mainly occur because of heart attacks as well as strokes. Therefore the proper detection of heart disease at a very early stage and the tools for the correct prediction of heart disease can save a massive amount of life and help the doctor design a very effective treatment plan that can reduce the mortality rate due to heart disease. Due to the specific development of the overall advanced Healthcare system, vast amounts of patient data are available on different websites. They are mainly big data that can be used to proto-design other predictive models for heart disease.Correctly mining for proper machine learning is a particular Discovery method that can be used for analyzing big data from various perspectiveperspectivessulating it in a very effective way into all the Useful information. The proper detection of heart disease will help all people, as well as doctors so that they can understand the severity of the disease at a very early stage and take proper precautions to mitigate the disease and help in the treatment process. Nowadays, big data regarding the customer's medical history is available on different websites as well as at the data storage of Any hospital, so all these data can be used for a data mining process which is a proper nontribal extraction of the previously unknown along with potential use for different information about the data. Recently, a vast amount of correct data about various disease diagnoses and patients has been directly generated by other Health Care industries. Data mining also provides many technologies that can discover hidden patterns and similarities in data and are beneficial to making the appropriate decision regarding the disease. In this research, the machine learning algorithm will propose the proper implementation of different heart diseases along with the prediction system, which can be validated over the heart disease prediction data set. This research will define and implement appropriate data set and then execute different models so that it can be beneficial to predict heart disease at a very early stage. The overall method is also required to analyze this research. A proper literature review will be performed, taking appropriate literature regarding the heart disease prediction system using machine learning to appropriate knowledge and Idea regarding the overall process.

## 1.2 Background

The importance, and advantage of the widespread application of heart disease detection, which is based on machine learning and prediction system, has been discussed in various research findings. In recent times, the amount of heart disease is growing day by day. Hence, it is essential to take proper advantage of technology along with the advanced system so that the doctor can detect heart disease at a very early stage and can take appropriate precautions as well as treatment plans which can reduce the mortality rate due to heart disease (Li *et al.* 2020). The main Specification of this research is that the study will analyze detailed data regarding the patient's health history of heart problems and then perform different implementation techniques and proper model development, which can help to detect the disease very effectively. These particular works aim to develop a helpful decision support system that can help detect the disease and make amakepriate decisions using the proper data mining technique with the best accuracy score. Different kinds of performance among Naive Bayes, Support vector machine random forest as well as simple Logistic regression is going to perform, which can help to detect the heart disease and the heart position. Different parameters of patients, such as age, ECG result, sex, Blood pressure, and blood sugar level, will be analyzed. They can help measure the possibility of getting directly affected by heart disease. This general detection uses a very efficient machine-learning technique to detect heart disease. It can be an effective tool for all patients, doctors, and medical students to diagnose heart disease. Performing the diagnosis of psychological condition as well as different symptoms and heart attack mainly require proper 24 hours monitoring of the oval patient's health data after transferring from the hospital to their home. All this patient data will be analyzed to provide an appropriate detection process that can analyze all these factors effectively and then perform a proper decision-making process for the doctors and patients (Alotaibi *et al.,* 2020). Another specification is that this particular research will help in different purposes in the entire Healthcare system where the data of other health disease parameters is going to be analyzed. Then the doctors can make appropriate decisions regarding the overall treatment plan. The entire process of detecting heart disease is going to be very easy through the advancement of machine learning and proper model implementation, which is very much necessary in recent times.

## 1.3 Problem Statement

In the current scenario, there is a different issue regarding adequately detecting heart disease. It becomes very hectic for doctors because many cases occur daily regarding heart disease. It is essential to take advantage of the different machine-learning processes for this purpose. The main issue is to perform the overall process manually because it takes proper knowledge and a medical certificate to check any patient. Still, through the machine learning process, it becomes straightforward, and it just requires the technical skill to analyze the data and develop a different model to make the proper detection (Bharti *et al.* 2021). A considerable number of people are suffering from heart disease, and the mortality rate is increasing day by day. Hence, it is essential to get the help of this detection process through machine learning and make decisions at a very early stage regarding the treatment plan of the patient, and it also reduces the stage from all the doctors. Another issue is the time of detection process because it has been observed that there is not enough availability everywhere to start the treatment as soon as possible. For this purpose, these detection processes can help to detect the severity level of the heart and make the proper decision by considering the result. Another issue that is associated with this purpose is the appropriate maintenance of massive data, which is called big data, which needs a suitable process such as data cleaning and other data mining techniques because without good data mining, it is not possible to use the data for different purpose of model development (Sarmah, 2020). All these issues are likely to mitigate by taking appropriate steps and action, and proper implementation of the machine learning model can help in this purpose to detect heart disease very early.

## 1.3 Research Question/Hypothesis:

***Research questions***

The research questions are discussed below:

* How may the effectiveness of heart detection of diseases be enhanced by a symptom-based strategy?
* How does reinforcement learning enhance the efficiency and specificity of heart disease detection?
* What will the research have over earlier methods in terms of improved performance?

***Research hypothesis***

The research hypothesis has discussed below:

* H0: The proposed work has predicated the chances of the heart disease and also classifies the risk of patient to implement the various data mining techniques.
* H1: The proposed work has not predicated the chances of the heart disease and also classifies the risk of patient to implement the various data mining techniques.
* H2: The research investigated the identification of cardiac disease using Python programming and a machine learning algorithm.

## 1.4 Aim

The project aims to detect human heart disease by analyzing the symptoms using machine learning algorithms such as K nearest algorithm, Random Forest, Gradient Boosting and Logistic Regression on the ‘Heart.csv’ dataset and additionally aims to design a hybrid model to improve the model accuracy.

## 1.4 Objectives

Here are my objectives for this work:

* To review the previous papers to gain knowledge of detecting disease, emphasizing heart disease detection and the relevant data and methods.
* To select a dataset containing the historical record of heart disease with several symptoms; Study the features and symptoms.
* To select machine learning models that will be applied to detect heart disease; also, proposed and designed the Hybrid Model, which will be used to the data for the same purpose.
* To apply the models to the data for the detection of heart disease; compare the performances and select the most effective model.
* To identify the research improvement by comparing the present best model with the existing approaches.

# 

# CHAPTER 2: LITERATURE REVIEW

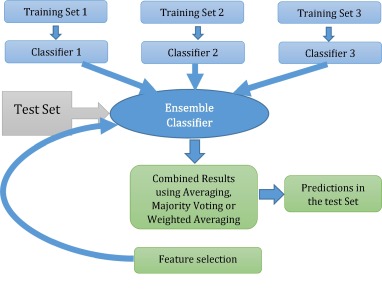
The literature review is the most significant element of the study since it thoroughly analyses all the prior research and incorporates new knowledge and concepts into the current study. The notion of heart disease detection will be put into practice in this section, and the correct information will be given by considering a variety of literature with the appropriate reasoning. In addition, the method of detecting cardiac disease through the deployment of machine learning will be examined using relevant references from established literature. Numerous machine learning techniques have recently been applied to the healthcare sector, and this research will investigate and explain the broader context of this implementation. There are several benefits of the machine learning implementation inside heart disease detection, which is also required to analyze in this particular section with a good idea. However, various issues of this implementation process and modeling also needed to study and explained by providing a proper critical analysis of the overall method.

## 2.1 Machine learning concept for heart disease detection

Machine learning can be essential to predict heart disease and many other diseases in the body. Suppose the expected oil is very effective in advance. In that case, all this information can effectively provide different important Insights to various doctors who can develop a proper diagnosis plan along with the treatment based on the detection result. Recently, heart failure disease has affected more people worldwide than all other autoimmune conditions (Segar *et al.* 2020). Proper cardiovascular disease affects the overall heart and obstructs the blood flow through different blood vessels. Heart attack and heart disease commonly occur when the specific heart fails to supply normal blood under very normal conditions, leading to high blood pressure Diabetes and other diseases.

There are different kinds of devices available in the market which can detect heart disease by considering various symptoms and the average rate of all these parameters. Still, the cost of this device is very high. It is not always possible to implement in a real-life scenario. However, such devices, as well as Technologies, are not just expensive. This is unfit for smaller clinics that can diagnose heart disease and is also very time-consuming. Several risk factors for every manual heart disease prediction can include inactivity in a proper physical form and very unhealthy eating habits, so it is essential to consider this factor for detecting the appropriate heart condition. The early identification of different heart diseases and the improved diagnosis, as well as high-risk individual using a proper prediction model, is very much recommended for the mortality rate reduction, and the overall decision-making is also going to improve for all the further treatment as well as prevention. Different libraries are present in the machine learning process, which can be very helpful for all these applications of the detection process (Abdar *et al.* 2019). However, in medical data, particularly data science, data privacy, and protection are necessary parameters that are impossible to ignore. Using the machine learning process, one can implement a proper Healthcare project where all the symptom, as well as other medical condition, is also going to analyze by taking appropriate precaution to make adequate detection regarding heart disease. It has been observed that the primary cause of any heart disease is an unhealthy diet, diabetes, being overweight, and the consumption of excessive alcohol as well as physical inactivity. The overall biological factor can be affected by all these causes, and the age of chest pain and pre-existing condition will also contribute significantly to this purpose. Several kinds of data set has been proposed to comprehensively train a particular machine learning model based on all these features and parameters that can be identified by an expert in the proper heart disease prediction and detection of heart disease (Baccouche *et al.* 2020). First, it is necessary to take the correct data set about the patient medical history and all the parameters that are the detecting factor of heart disease.

Furthermore, it is essential to analyze all this data, and with the proper machine learning model, it becomes possible to make an appropriate decision. In recent times, several data are available for the heart detection process, so it is essential to take those data and perform different functions inside the machine learning algorithm to make appropriate decisions and predict this data set. The overall concept of machine learning about heart disease detection is straightforward as its first takes the valid data set and then trains and tests it through taking appropriate machine learning algorithm so that it can analyze all these factors and features very effectively and provide an appropriate accuracy score which then can decide the proper treatment plan for the particular patient. In other words, it is possible to say that machine learning is a process that can simplify the overall data and Break it into two different segments so that it becomes possible to understand the outcome very easily (Terrada *et al.* 2019). The general concept is very much necessary to implement in the real-life scenario because without proper implementation, it is not at all possible to get the appropriate advantage of the overall process along with achieving high accuracy scored through the model implementation can help to get an accurate and proper measure.



#### Figure 1: Detection of Heart Diseases

(Source: Terrada *et al.* 2019)

## 2.2 Process of heart disease detection using machine learning

Different kinds of machine-learning approaches are associated with heart disease prediction. It is essential to analyze, implement, and execute all these approaches in real life to get a proper outcome. Various approaches of machine learning algorithms are random forest classifier, K nearest neighbors, decision tree classifier, and support vector machine.

***Random Forest Classifier***- The particular random forest algorithm can provide flexibility and robustness for proper classification Tasks with appropriate big data that all the other standard models can perform. The good simplicity, as well as versatility of the London for the specified, is now used very much in different kinds of other processes that can be predicting heart disease or fraud detection for different purposes (Joloudari *et al.* 2020). With the appropriate ensemble Learning theorem, the suitable random forest classifier mainly combines the result from various decision trees and different optimized training. This classifier aims to utilize various subsets and find the best combination to increase the predictive accuracy of the data set about the patient data as well as their medical history. The first step is to build mixing and optimize and match various decision trees. Further, it uses all these trees for proper prediction and accumulates all these results to get the final output prediction. This particular classification process is very much necessary in recent times to predict the appropriate heart disease through analyzing massive data about the patient history.

***K-Nearest Neighbors***- Through the name of this particular classification process, it becomes easy to understand that the neighbor's cassie fire Takes a good data point and also finds the k, another different data point near to it inside the proper vector space. In this Supervised particular fashion, the KNN develop a specific cluster of the overall data sample that has the very same target value, and the model needs to Have accurate data regarding the patient factors as well as all the factor that is associated with heart disease (Angraal *et al.* 2020). It enables any new data, as well as elements, to be classified effectively. It then assigns a proper distance metric to any of these classes. To perform heart disease detection, only two types are present that the KNN must build. In addition, it is appropriately robust and efficient for detecting heart disease and making the proper decision. Various metrics of distance will be used in this process, and the overall choice of this metric can impact the entire classifier feed for an extensive data set. However, this specific classification process is slower than all these contemporaries.

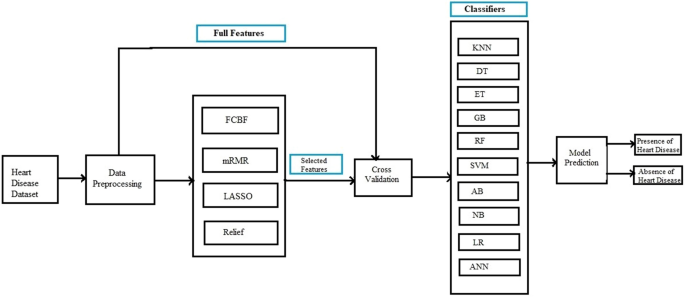
***Decision tree classifier***- The specific decision trees are every individual model which can make appropriate random forests after accurate ensembling. These specific decision tree classifiers can use the proper attributes of extensive patient information to create the right tree. The branches of the decision tree classifier end up in the suitable leaves, which are made up of different target values (Choi *et al.,* 2020). This classifier uses the proper visual components and the appropriate information gain index. This process tries to identify all the top features of different labels associated with each class. In addition, all these branches are developed to maximize the general information gained in each split and lead to the particular leaf node of that class. A decision tree classifier is very fast, and it is very much capable of performing the proper disease prediction through all the symptoms. If the data set has compelling features, it can complete the entire process most efficiently and effectively.

***Support Vector Machine***- The support vector machine algorithm is an excellent non-probabilistic classifier that can be very beneficial to generate proper hyperplanes that can divide the overall data points of any two classes in the specific vector space (Windmon et al. 2018). In particular, several features as well as it can directly create M1 in a dimensional Hymer plane, separating the data points of various classes from one another (Gonsalves et al. 2019). The support vector can be calculated in the proper margin between the vector of a particular type is the most and can optimize this particular margin metric to find the adequate and effective hyperplane for all the available categories. This specific support vector machine can perform the appropriate disease prediction as it can very effectively categorize all the data into various types, which is very helpful for this overall model implementation and getting the proper outcome for better analysis.

Researchers also utilized the advantage of neural networks, but the processing time remains the same as traditional ML algorithms. ***Artificial Neural Networks***- The specific artificial neural network is the most popular machine learning algorithm today because it has a wide application inside the deep learning process, which can help to perform different kinds of methods very effectively. This process can comprise a handful of linear notes that can function adequately comparable to the highest standard machine learning models. There is a specific structure of this neural network that can make up various linear notes. It can effectively perform the overall process by understanding the pattern between all these hidden layers (Dritsas and Trigka, 2022). There are different hidden layers between the output layer and the input layer to increase the overall complexity and the learning ability of the entire model. Adding more layers and nodes to the overall new comprehensive allow the whole model to learn more specifically the complex relationship between the input features and a categorical variable, which can help to get an effective accuracy score that is very much accurate. The capability makes the overall network very effective in capturing the in-between relationship of all these various biological and personal makers that can independently affect the overall probability of the entire presence of heart disease in any patient.

All these algorithms described above can perform the heart detection process by considering different symptoms of heart disease, and it has been observed that they can provide a proper outcome and accurate result (Haq *et al.* 2018). The doctor, as well as other people associated with this health industry, will benefit from this particular machine learning model, which analyzes all the symptoms and provides an appropriate result. The purpose of all these algorithms in this research is to evaluate different aspects of the data set and give the proper variable analysis to help determine the accuracy score. In the context of age distribution and its impact on heart disease detection, it is possible to implement the nearest neighbor classification process. on the other hand, the blood sugar level and its distribution are possible through the random forest process. This algorithm’s main context is finding the appropriate accuracy score to critically evaluate the data set’s attributes.

Researchers suggested that Neural networks should be utilized only when the traditional algorithms fail to achieve acceptable accuracies. Hence we may save the processing time of the training.



#### Figure 2: Early Detection of Heart Diseases

(Source: Haq *et al.* 2018)

## 2.3 Benefits of machine learning in detecting heart disease

There are several benefits of the machine learning model in detecting heart disease. In recent times, it is essential to see the problem at a very early stage so that the doctor can take appropriate precautions to mitigate the issue as soon as possible. The first benefit of the machine learning model is that it processes essential data, which is crucial in recent times as big data can help get the appropriate outcome. The accuracy score will also be very high if many attributes are present inside the data set (Chowdhury *et al.* 2019). Another benefit is that big data always provide an appropriate outcome and accurate accuracy because it is processed with different attributes of various features and parameters. The patient’s symptoms are also associated with this data set. Machine learning is not a complex process; proper skill and little knowledge can help determine the appropriate outcome through this model implementation and make the right decision regarding heart disease treatment. It has also been observed that the overall model implementation is not that time-consuming and very effective in different terms and conditions where manually, it is not always possible to implement those data to analyze the proper outcome and make future decisions. It has often been observed that big data always help to perform different processes because the model will train first. After The training process, it is required to complete the appropriate test of this data set to provide proper outcomes and feed the model. Machine learning can quickly review a large volume of data and discover specific patterns impossible for any human to detect (Abdel-Basset *et al.,* 2020). For example, suppose there is vast data about the patient’s older medical history. In that case, it can perform the property identification of different feature that is similar and then make a pattern that is associated with it and help to make the proper decision about it. No human is required to be related to this process of heart disease detection through machine learning because it can effectively perform the prediction and also improve the overall algorithm on its very own. Machine learning can perform various processes, and it can learn new things, which can train its model to provide the appropriate outcome.

A machine learning algorithm always gains experience and keeps improving its accuracy and efficiency to make better decisions for various purposes. It has also been observed that if there is important data and a vast amount of data available, the model keeps growing and can make proper accurate predictions significantly faster with a minimal time duration. A machine learning algorithm can also handle a multi-variety of data with multidimensional data available in this data set. Machine learning can also perform accurately in various uncertain conditions where the environment is not appropriate for implementing different algorithms. It can always predict heart disease correctly (Ali *et al.* 2020). The exposure, as well as flexibility associated with the machine learning algorithm, is also very high. All these features make it practical to perform in any environment with a total accuracy level.

## 2.4 Issues in the detection process of heart disease using machine learning

Machine learning algorithms benefit the whole Healthcare industry by predicting different ages and, most importantly,t heart disease because doctors make more appropriate decisions through the machine learning process. However, there are some flaws and a disadvantage associated with the machine learning process, and sometimes it also predicts the wrong exemption about heart disease and treatment plans. The first disadvantage is that machine learning requires a massive amount of data to train on, and all this data must be unbiased and of good quality (Chicco and Jurman, 2020). However, it has been observed that it is not always possible to create a proper data set that is unbiased and appropriate to the patient data. Machine learning also requires a considerable amount of time so that the data can train appropriately, perform, and pay its proper outcome. A significant amount of time is also associated with providing accurate accuracy scores. Many resources are also required for the machine learning algorithm; otherwise, it cannot perform effectively. All algorithms are not appropriate to provide tangible outcomes, so it is essential to choose the right product to interpret the model’s result. There is also a high chance of error. An appropriate algorithm implementation is necessary because a small wrong input inside the algorithm can predict inaccurate output results. All different steps are required to implement in the most effective manner forgetting appropriate outcomes in the real-life scenario and Making proper outcomes regarding heart disease (Cikes *et al.* 2019). Biased prediction is also possible, which is not at all recommended. Hence, carefully choosing the model implementation and the proper steps to follow is necessary for any unbiased heart disease prediction.

## 2.5 Literature Gap

In this research, different Literature has been analyzed to get an appropriate idea and the method that can help understand the overall process very quickly and effectively. However, through proper analysis, it has been observed that there are several years of gaps inside the literature, which can impact the overall quality of the research. Hence, reviewing this literature and adequately identifying the gap is necessary. The first thing observed is that this data set has not clearly defined the appropriate concept of machine learning in predicting heart disease by considering various symptoms. A very lack of information is present inside this description of the idea, which further impacts the overall quality of this literature and does not provide appropriate knowledge about the implementation techniques. The critical analysis has not been described in this literature, which can showcase the negative side of the literature and the disadvantage of machine learning algorithms which can also make different kinds of errors in the production process. The implementation technique and the approach of various algorithms in the heart disease prediction system have not been appropriately included in the literature so that anyone can get proper knowledge about the steps that need to execute in any machine learning software platform. The codes and an algorithm required to implement and run in the software platform have not been adequately described in this literature, and this gap can impact the quality of the research very much in a very negative way. The researcher has identified all these gaps in the literature and then implements appropriate knowledge and ideas so that this research does not contain this kind of gap and provides relevant information about the process of a machine learning algorithm to detect heart disease. This gap in the literature can be reduced by taking an appropriate approach and proper information about the machine learning implementation and the issue that anyone can face in this implementation process, which is also required to include in this research.

Based on the above discussion, this particular section of the literature review has very effectively caused the proper concept and execution process of a machine learning algorithm to detect heart disease in the most effective manner. All the machine learning algorithms that can be used for heart disease detection have been effectively described in the section with proper references and justification. The need for machine learning in the current scenario to perform various implementations that can evaluate the extensive data about the patient history and then provide the appropriate outcome is also explained with proper references. Miscellaneous issues associated with the machine learning process and all the flaws are also discussed with proper hotruein this careful research so that anyone can take appropriate precautions to reduce the severity of these issues.

## 2.6 Issues

## 2.6.1 Ethical Issues

Ethical issues might occur in this project. The test being run and its outcomes, such as incorrect detection results, might affect the study analysis and conclusions. These moral concerns are seen as crucial and may have a detrimental effect on the enterprise. Other than that, maintaining data secrecy is essential for the research, and if the data are leaked, it may raise ethical concerns (Kaur *et al*. 2021).

## 2.6.2 Legal issues

There might be fewer concerns regarding legal difficulties because this research will conduct a sufficient professional test that is regarded ethical. Although the researcher may produce problems that may breach the Act's terms if they do not adhere to the Data Privacy Act of 2018, Following the Act is essential, and all other requirements should be adhered to as well (Rath *et al.* 2021).

## 2.6.3 Professional issues

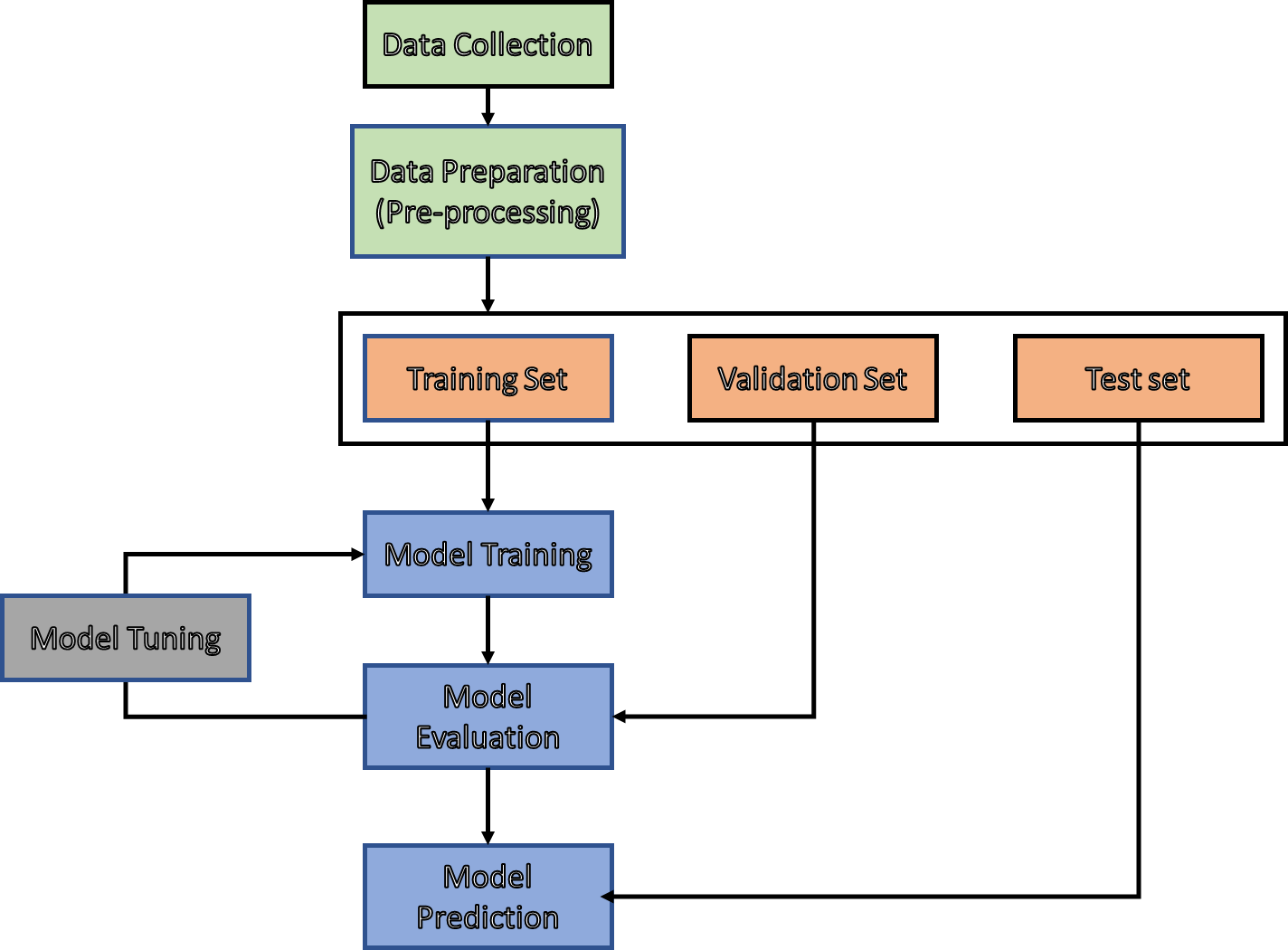
Due to the fact that the investigator must uphold their professional standards in all aspects of their study, there may be several professional concerns. Issues might arise if secondary research is included in the absence of primary data, or if knowledge from other studies is gathered at the levels of proficiency of the investigation (Kora *et al.* 2020). Aside from that, problems with approaches used wrongly will result in less professional work being displayed.

## 2.6.4 Social Issues

The project will be accomplished entirely on a technological basis, thus it is clear that there won't be any social concerns. There could be a situation where running the test via Wifi develops a connection to any nearby Wifi, which could be problematic (Nashif *et al.* 2019). If moral norms are upheld, however, this problem is nearly impossible. Aside from that, there is a slim probability that societal concerns will be explored for the project.

# CHAPTER 3: METHODOLOGY

The many types of approaches are distributed in this area to demonstrate the project by the project's order. The many research methodologies are distributed in this area to construct the data gathering and analysis methodologies. The cohesive component is examined here to offer the project a logical path. To administer the allotments, many kinds of analytical affiliations have been applied. The initiative clarifies that the agendas will be utilized for machine learning-based heart disease identification. The project indicates that machine learning will be employed in this segment to satisfy the criteria. The data analysis process will be demonstrated in this scenario using a variety of ways, and the data storing process will be developed in the next part. To address the issue, different types of agendas are distributed in this part. I have applied logical method to this study challenge. Here, I have created the scientific inquiry using a flowing approach. Here, existing theories will be examined to learn more about the project. To learn more about the heart disease monitoring system, I have used the various allocation types. With the use of ML methods, wearable technology has been improved to make it simpler to monitor heart disease. Among the most popular subjects that have received extensive attention since antiquity is the identification of heart disease (Abdar *et al.* 2019). The most recent improvements in technical approaches, however, have pointed to a number of useful applications. Additionally, one of the practical applications for numerous wearables in the realm of continual health monitoring is possible and involves using them all at once. As most of the authors suggested who are working in these areas, we followed the generic methodology in getting the higher accuracy as stated in the Figure 3.



#### Figure 3: Methodology followed for training

(Source: Self Developed)

## 3.1 Data collection

In this work, the Cleveland Heart Disease (ul et al. 2020) dataset is taken into consideration for testing purposes. There were 303 occurrences and 75 attributes when this data set was designed, but all published experiments only use a subset of 14 of these. Six samples were excluded from the data set in this work due to missing values after pre-processing. There are 297 samples from the remaining dataset of 13 characteristics and 1 output label. Two classes on the output label indicate if HD is present or not. As a result, the 297\*13 features matrix of extracted features is created.

## 3.2 Data Pre-Processing

The processing of the dataset beforehand, which is necessary for accurate depiction. The dataset went through some pre-processing steps, including the removal of attribute missing values, the application of Standard Scalar (SS), and the utilisation of Min-Max Scalar. if there are any needless points present, which have a further impact on the training process. Then the data is divided into three sets in the ratio of 8:1:1 for Training, Validation and Testing respectively.

## 3.3 Tools and libraries used

Different tools and libraries have been used in the entire process and it can be seen that the task has been done with the help of ***the "Jupyter Notebook"*** platform using the ***“Python”*** programming language. The libraries that have been used in the code as "NumPy", "pandas", "Matplotlib", and many more. The libraries have been used to perform a certain number of tasks such as data visualization, algebraic calculations and many more. Apart from these, several libraries have been imported as per the use of the machine learning algorithms and it can be seen that the libraries have been imported before the task for obvious reasons.

## 3.4 Machine Learning Algorithms:

This study built four prediction models using regression analysis. Each model is based on one of four machine learning techniques: Gradient Boosting, Logistic Regression, K-Nearest Neighbours Regression, and Random Forest. Regression analysis is a collection of statistical methods used to examine causality relationships. Potential relationships between variables can be modelled and used in forecasting with its help. Regression exemplifies how modifying explanatory variables can rectify the effects of modifying dependent variables. Here, heart disease is the independent variable, and the desired outcome is the dependent variable.

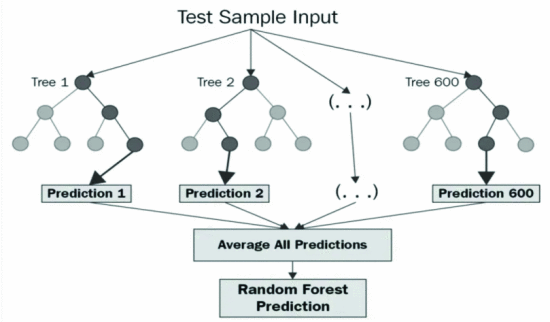
* + 1. **Random Forest Model**

The components of a Random Forest, a machine learning system, are individual decision trees. Each tree is constructed independently using a bootstrap sample and randomly selected features, leading to a prediction-improved, uncorrelated forest. The following is a diagrammatic description of the algorithm used to generate an N-tree Random Forest, When n is equal to 1..., N:

* It is advisable to generate a bootstrap sample Xn.
* Generate a bn tree for a given sample (Xn).
* You can use the specified criteria to determine which property is the best, then use that split to divide the tree and continue doing so until you've exhausted your sample.
* The tree is built until either a certain height is reached or each leaf has a maximum of nmin items.
* Determine the most effective method for classifying each subset using m randomly selected features based on a set of initial characteristics.

Weak correlations can be constructed by carefully choosing the features and hyperparameters to use. With the random subspace method, we can reduce tree correlation and avoid overfitting. The basic algorithm is taught on a variety of randomly selected subsets of the feature description. These models were built using the random subspace method, and their assembly procedure is as follows:

* First, we define N as the entire number of learning objects, and then we define D as the total number of features.
* The ensemble's size, L, requires selecting a subset of models.
* For each model l, you'll need to decide on a set of characteristics, or "features," by entering a value in the format dl (dl D).
* Model l must be trained by selecting feature dl from distribution D and constructing a training sample.
* It is crucial to combine the results of numerous L models by mixing the posterior probability.



#### Figure 4: Random Forest Architecture

(Source: *Sandy* *et al.* 2021)

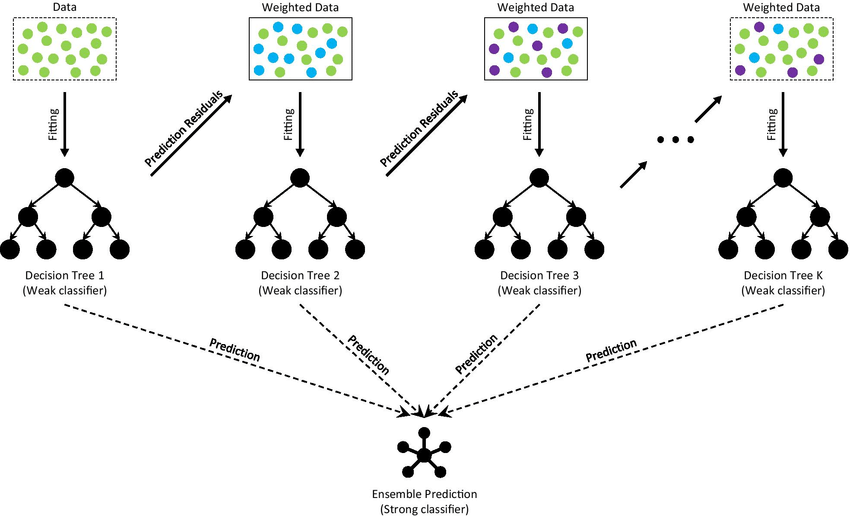
**3.4.2 K-Nearest Neighbours Model**

K-Nearest Neighbours, a machine learning technique, finds nearby objects having values for the target variables. In most cases, the average of the latest K samples of data is used as the predicted outcome of the regression issue. To create a model's connection between its "group of objects" and "dependent variable," a training sample is required. The distance function between the items in the collection must be explicitly defined in a way that is not shared with any other collection. The function takes a random item and sorts them in ascending order by the distance to all objects in a certain class.

In this case, the class boundaries will be overly complex, which is at odds with the simplicity of the technique. However, the paradox disappears once you realise that the objects in the training sample are also exceptional method parameters. Cross-validation is a method for evaluating an analytical model by subjecting it to tests on multiple data sets to ensure that all of the available data is being used fairly.

**3.4.3 Gradient Boosting Model**

Gradient Boosting is a machine learning strategy for classification and regression difficulties that builds a prediction model from an ensemble of weak predictive models. In this case, decision trees make up Gradient Boosting. The loss function is minimised using decision trees that are trained in an iterative fashion. Because of the characteristics of decision trees, Gradient Boosting can handle categorical information and nonlinearities. Enhancing the training of untrained models is what "boosting" is all about. In boosting, data from the original dataset is used to train each new tree. Given a set of feature pairs (x, y) with i=1,...,n, the dependence of the form y = f must be restored for each pair (x). This entails minimising the differentiable loss function L(y,f). It is required to find approximations that, on average, minimise the loss function given the facts at hand. Our search is constrained to a small set of parameterized functions.



#### Figure 5: Gradient Boost Architecture

(Source: *Deng* *et al.* 2021)

**3.4.4 Logistic Regression Model**

Logistic regression is a linear statistical method for modelling the association between a scalar response and one or more explanatory factors (also known as dependent and independent variables). When only one explanatory variable is available, simple Logistic regression can be utilised; when many are, multiple Logistic regression can be employed. This term is more precise than "multivariate Logistic regression," which predicts several interrelated dependent variables rather than a single scalar one.

Using Logistic predictor functions, we may model relationships and estimate their unknown values using existing data. Logistic models are the most common type. In many applications, the conditional mean of the response is used because it is assumed to be an affine function of the values of the explanatory variables (or predictors). However, unlike multivariate analysis, which considers the combined probability distribution of all these variables, Logistic regression analyses only the conditional probability distribution of the response given the values of the predictors.

## 3.4.5 Hybrid Model

In this study we also used the Hybrid models by combining the multiple proposed ML algorithms (LR, RF, GNB, KNN). By this we can compare the performance of the individual and hybrid model. The initial model will learn the features which are provided then it calculates the error percentage accordingly and pass the outputs to the second model which we used. I have used the same ratio of dataset as used for the individual models. As compared to the individual model training, the hybrid model provided the best results as we observed.

# CHAPTER 4: RESULTS AND DISCUSSION

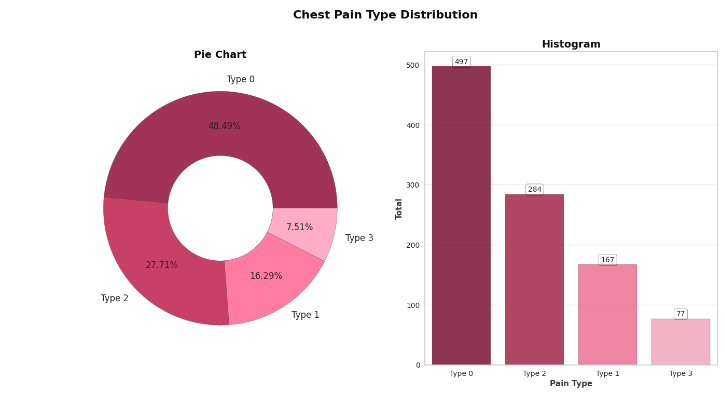
In this study, I used the heart disease dataset to compare the proposed models of Random Forest, Logistic Regression, Gradient Boosting, and K Nearest Neighbours. In order to calculate performance metrics like accuracy and precision, I extracted the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) from the code. I also compared the Hybrid model to the separate models.

## 4.1 Experimental setup

I have developed and executed the models using the advantage of jupyter in python language. Various required packages are installed like Matplotlib, Pandas, NumPy.

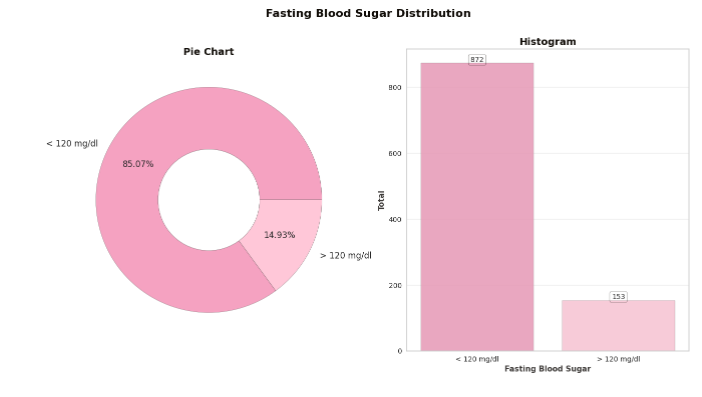
## 4.2 Data Import

The data set regarding different Heart function attributes has been imported inside the Jupyter platform. The importation process is the most important part of the whole implementation method because it helps to import the data within the software platform and further view it accurately. Describing the initial data exploration process where all the attributes have been clearly evaluated with the size and distribution of all the variables inside this data set. As well divided the columns into input features and targets. Further targets are provided as an output to the models. Some of the classes of individual features has been provided below with the help of pie charts.



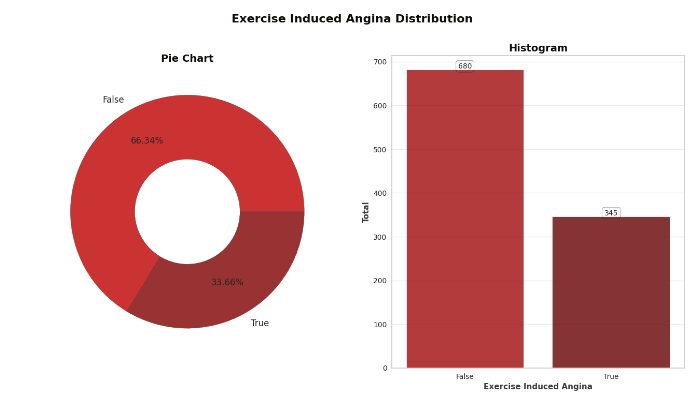
#### Figure 6: Percentage of the type of chest pain that generally occur inside the human body.

(Source: Self-Developed)



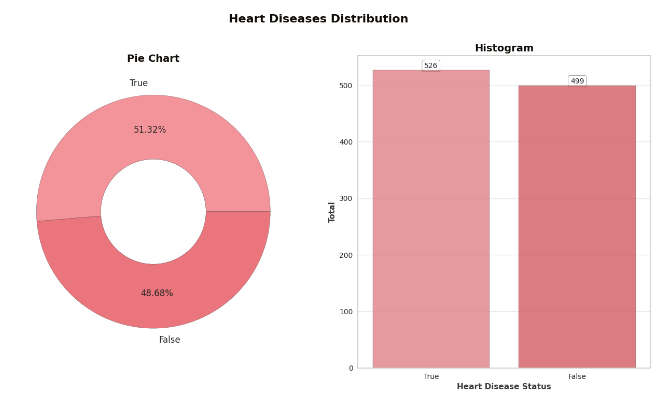
#### Figure 7: The blood sugar level connected with the level of heart disease and high blood sugar become the risk of a heart attack

(Source: Self-Developed)



#### Figure 8: Exercise distribution, which can help to reduce the risk of a heart attack in individuals

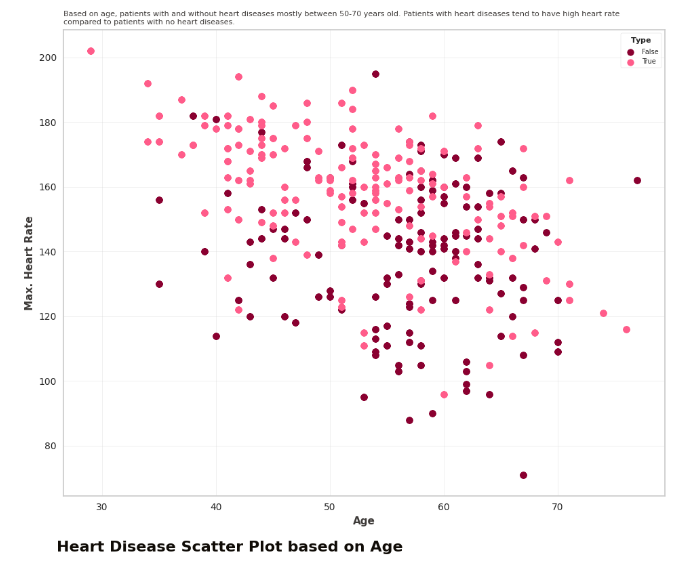
(Source: Self-Developed)



#### Figure 9: heart disease distribution, percentages showing the status of heart disease appropriately

(Source: Self-Developed)

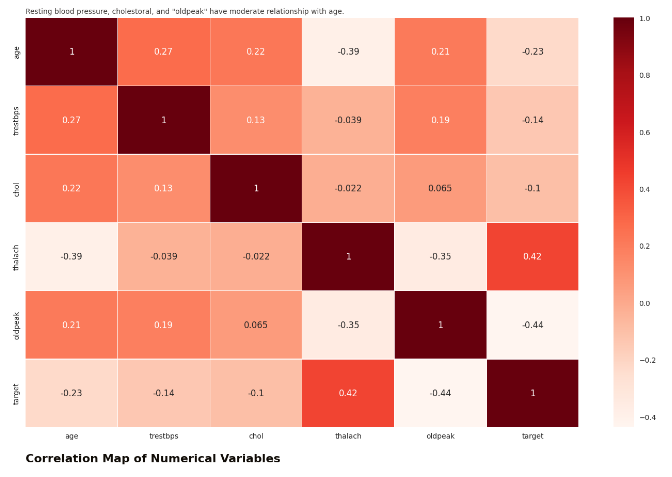
The scatter plot has been implemented regarding the heart disease distribution and the maximum heart rate has been evaluated through this graph. All this information regarding the maximum heart rate has been given in this particular picture.



#### Figure 10: Scatter plot

(Source: Self-Developed)

This is the correlation map which is clearly showing the relationship between all the attributes that are present inside the data safe and directly connected with the detection process of heart disease inside the human body. This correlation map also helps to identify the pattern as well as the relation between different variables which is going to evaluate for a better accuracy score.



#### Figure 11: Correlation Map

(Source: Self-Developed)

## 4.3 Data Spliting and Scaling

Data pre-processing has been clearly performed in this study, where all the unnecessary variable has been dropped in order to provide a better clean data set which is appropriate for the implementation process. After that, the data set has been split between train as well as a test which is also important for model implementation. It has been observed that without the splitting process it is not at all possible to implement different classification models inside the data set. Then we standardized the data before the training start, in order to measure similarities between observations and form clusters.

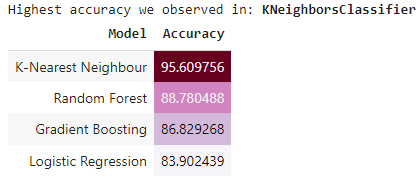
## 4.4 Models Performance

After pre-processing and standardizing the data, we fitted the heart disease dataset directly to the proposed models with the required hyperparameters. The performance and achieved metrics of individual models has been discussed in detail below:

Logistic regression that has been implemented inside the Google colab software platform. The accuracy score that has been implemented in this process of Logistic regression is 83.90 %. It has achieved very less as compared to the remaining models, it may be because of the less iteration of training or failing in finding the regression value.

Thereafter, the k-nearest algorithm has been performed using the appropriate Court so that it can provide very high accuracy. The accuracy score for the k-nearest algorithm is 95.6 1% which is remarkable and also defining that all the code has been executed accurately to achieve this high score. Reason behind the achieving the remarkable accuracy is it has an ability in finding out the nearest cluster values. As well our data points are easy in helping the model to find out the relations without any complexity.

Gradient Boosting and Random Forest has also been implemented in order to get the appropriate outcome of heart disease detection using all the machine learning algorithms. The accuracy score that has been obtained through this process is 86.82% and 88.78% respectively which are also high compared to all the other processes, but not higher than the KNN.



#### Figure 12: Model Performance

(Source: Self-Developed)

In the above-mentioned picture, a clear Idea and competition between all the models have been provided. there are different models that have been implemented in this particular research in order to understand the accuracy score along with the correct modal implementation for heart disease detection using different symptoms of the human body. So the overall analysis it has been observed that the K nearest neighbor algorithm has the highest accuracy score and on the other hand, the Naive Bayes has the lowest accuracy score so this proper competition provides the researcher a clear idea about the model that is accurate for this process (Krittanawong *et al.* 2020). It is also very effective to understand each and every aspect of this model implementation and how it can impact the overall process by effectively determining different attributes that are present inside the data set and providing clear inner relations between these attributes.

## 4.5 Hybrid Model Performance

We combine the proposed models to increase the accuracy, this type of technique is called as hybrid ML method. Initially, the KNN will be applied to the dataset and the output of KNN will be carried out to the next algorithms simultaneously. As we observed that more accuracy has been achieved in the hybrid model which is 98.54%. The models are learning from the error rate which is observed in Predecessor models.

# CHAPTER 5: CONCLUSIONS

In the long run, recognising the processing of raw healthcare data of cardiac information would allow for the saving of human lives through the early diagnosis of anomalies in heart diseases. In this study, we used machine learning approaches to analyse the raw data and provide novel understanding of cardiac disease. Cardiovascular disease prediction is both challenging and essential in the medical field. But the mortality rate can be drastically lowered if the condition is diagnosed early, and preventative measures are put into place as soon as possible. Developing this study further is highly desirable so that researchers may zero down on real-world datasets as opposed to only theoretical frameworks and simulations. Even though KNN achieved good accuracy in individual model approach of 95.61%. But, the features of the KNN, RF, GNB, and linear methods are integrated into the proposed hybrid model approach (LM) have chieved a high accuracy of 98.54%. Predictions of heart disease using the hybrid model proved to be very precise. In the future, researchers might explore this topic by combining different machine learning approaches to create more accurate prediction models. In addition, novel feature-selection methods can be developed to learn more about the most important factors and improve the predictability of cardiovascular diseases.

## 5.2 Summary of achievements

This research contributes to the development and evaluation of numerous technical techniques in the fields of machine learning and appropriate algorithms for diagnosing heart disease and its varied symptoms. All the information regarding the specification of this project has been given in this particular research with proper justification so that it can provide an appropriate analysis of all the processes that have been performed. In order to understand the whole concept of heart disease detection using different kinds of symptoms in the human body has been clearly described in this research with proper differences as well as various machine learning tools also used for this purpose. Different techniques of machine learning algorithms have been widely utilized for evaluating as well as developing the overall process of decision-making regarding heart disease and the benefits are also evaluated in this research. The aim along with the objective has been clearly discussed with the different current issue that is going on in the Healthcare domain regarding the heart disease process and how this research can bring revolution in the health sector also define through taking an appropriate example. The process of heart disease detection through identifying different symptoms in the human body also explain briefly in this research and how machine learning Tools and techniques are going to help in this process is also defined. Various advantages of machine learning Techniques in order to identify heart disease and how different attributes can clearly connect with the whole process are also explained in this project. There are also various issues involved with the whole execution process inside the machine learning software platform and all these issues clearly identified in this research so that the researcher can take appropriate M creation process in order to remove them. All the previous literature has been reviewed appropriately true taking a realistic approach so that all the gaps can be identified and can mitigate in this particular research in order to provide a high-quality process of heart disease detection.

## 5.3 Reflection

First, I conducted research using the assistance of earlier publications on this topic that are available on many reliable platforms and free of charge in order to understand the fundamental notion of this entire research job. Following this, I have chosen appropriate machine learning algorithms through study. In order to fully comprehend the subject and make connections to current, inescapable challenges, I have also mentioned several useful earlier research efforts. In the earlier parts of this comprehensive research effort, I have also highlighted numerous problems as well as numerous advantages of algorithms for machine learning in the field of health care. The full taxonomy of error metrics has been calculated by myself. The most crucial data pre-treatment step, which I used to apply to all of the independent variables in the particular dataset, is also stated. Additionally, I have chosen one of the best qualities for a technique of prediction that is considerably more accurate. These subjects were already covered in the earlier phase of my research project, along with screenshots of my work on the Google Colab platform.

## 5.4 Recomendations

The heart disease detection process is going to provide appropriate evolution in the Healthcare sector by identifying different symptoms in the human body which is directly connected with heart disease. In this research, a raw data set along with different resources has been taken in order to evaluate the appropriate process of this detection method (Cikes *et al.* 2019). It also utilized a huge range of different algorithm that is important for this process however there is some recommendation that is required to implement in all future research in order to improve the quality. The first recommendation is to take a more appropriate data set which can help to evaluate a better modelling process as well as a more appropriate evolution procedure. The accuracy score along with the result of the overall process is directly connected with the data set along with all the attributes that are present inside the data set so it is necessary to take an appropriate data set that consists of all the important attributes and variables. Another recommendation is to critically evaluate the complexity of the whole process which anyone can face in order to execute it inside the software platform (Uddin *et al.* 2019). All the classification algorithm that has been implemented in the section is required to explain more appropriately so that anyone without prior knowledge regarding the whole process can understand all these algorithms very well. Apart from these and another main recommendation is to clearly discuss each and every step which can be helpful to determine and identify different patterns which further improve the overall outcome of the research. Choosing the appropriate machine learning algorithm is also a recommendation that can help full to get the proper output result and to get the appropriate outcome from the data set regarding heart disease detection (Chowdhury *et al.* 2019). A proper technical approach is also needed to take which is very much important to provide the overall concept of these disease detection using different symptoms as well as the attributes of the human body such as age, blood pressure level, and many more.

## 5.5 Future work

There is a huge possibility and application associated with this research which is the proper heart disease detection through using various machine learning techniques, algorithms, and tools. All further work is required to evaluate the mode accessible approach which can provide more accurate results regarding this detection process so that the health sector can use the overall method very easily. The entire complexity level of the whole process is required to reduce in order to provide a proper realistic approach in the health domain with a reduced cost. Apart from this, all future work must include the proper details steps of these machine learning algorithm implementations regarding heart disease detection. Various methodologies as well as approach is also needed to evaluate so that it can provide a better outcome as well as result of this detection process. Accuracy is the main thing that can help positively impact the whole method and also provide a better analysis. All future work is also required to evaluate the complexity of this process so that anyone can understand the date of implementing the machine learning algorithm in the Google colab software to get proper correlation as well as classification model.

# References

Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P., 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *179*, p.104992.

Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P., 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *179*, p.104992.

Abdel-Basset, M., Gamal, A., Manogaran, G., Son, L.H. and Long, H.V., 2020. A novel group decision making model based on neutrosophic sets for heart disease diagnosis. *Multimedia Tools and Applications*, *79*(15), pp.9977-10002.

Ahmed, H., Younis, E.M., Hendawi, A. and Ali, A.A., 2020. Heart disease identification from patients’ social posts, machine learning solution on Spark. *Future Generation Computer Systems*, *111*, pp.714-722.

Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M. and Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, pp.208-222.

Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M. and Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, pp.208-222.

Alotaibi, F.S., 2019. Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, *10*(6).

Alotaibi, S., Mehmood, R., Katib, I., Rana, O. and Albeshri, A., 2020. Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and Machine Learning. *Applied Sciences*, *10*(4), p.1398.

Angraal, S., Mortazavi, B.J., Gupta, A., Khera, R., Ahmad, T., Desai, N.R., Jacoby, D.L., Masoudi, F.A., Spertus, J.A. and Krumholz, H.M., 2020. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC: Heart Failure*, *8*(1), pp.12-21.

Ayon, S.I., Islam, M.M. and Hossain, M.R., 2022. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, *68*(4), pp.2488-2507.

Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C. and Elmaghraby, A., 2020. Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information*, *11*(4), p.207.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, *2021*.

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), pp.1-16.

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), pp.1-16.

Choi, D.J., Park, J.J., Ali, T. and Lee, S., 2020. Artificial intelligence for the diagnosis of heart failure. *NPJ digital medicine*, *3*(1), pp.1-6.

Chowdhury, M.E., Alzoubi, K., Khandakar, A., Khallifa, R., Abouhasera, R., Koubaa, S., Ahmed, R. and Hasan, A., 2019. Wearable real-time heart attack detection and warning system to reduce road accidents. *Sensors*, *19*(12), p.2780.

Chowdhury, M.E., Khandakar, A., Alzoubi, K., Mansoor, S., M. Tahir, A., Reaz, M.B.I. and Al-Emadi, N., 2019. Real-time smart-digital stethoscope system for heart diseases monitoring. *Sensors*, *19*(12), p.2781.

Cikes, M., Sanchez‐Martinez, S., Claggett, B., Duchateau, N., Piella, G., Butakoff, C., Pouleur, A.C., Knappe, D., Biering‐Sørensen, T., Kutyifa, V. and Moss, A., 2019. Machine learning‐based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *European journal of heart failure*, *21*(1), pp.74-85.

Cikes, M., Sanchez‐Martinez, S., Claggett, B., Duchateau, N., Piella, G., Butakoff, C., Pouleur, A.C., Knappe, D., Biering‐Sørensen, T., Kutyifa, V. and Moss, A., 2019. Machine learning‐based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *European journal of heart failure*, *21*(1), pp.74-85.

Dritsas, E. and Trigka, M., 2022. Stroke risk prediction with machine learning techniques. *Sensors*, *22*(13), p.4670.

Gonsalves, A.H., Thabtah, F., Mohammad, R.M.A. and Singh, G., 2019, July. Prediction of coronary heart disease using machine learning: an experimental analysis. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies* (pp. 51-56).

Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.

Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.

Hill, N.R., Ayoubkhani, D., McEwan, P., Sugrue, D.M., Farooqui, U., Lister, S., Lumley, M., Bakhai, A., Cohen, A.T., O’Neill, M. and Clifton, D., 2019. Predicting atrial fibrillation in primary care using machine learning. *PloS one*, *14*(11), p.e0224582.

Joloudari, J.H., Hassannataj Joloudari, E., Saadatfar, H., Ghasemigol, M., Razavi, S.M., Mosavi, A., Nabipour, N., Shamshirband, S. and Nadai, L., 2020. Coronary artery disease diagnosis; ranking the significant features using a random trees model. *International journal of environmental research and public health*, *17*(3), p.731.

Joloudari, J.H., Hassannataj Joloudari, E., Saadatfar, H., Ghasemigol, M., Razavi, S.M., Mosavi, A., Nabipour, N., Shamshirband, S. and Nadai, L., 2020. Coronary artery disease diagnosis; ranking the significant features using a random trees model. *International journal of environmental research and public health*, *17*(3), p.731.

Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T. and Baber, U., 2020. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, *10*(1), pp.1-11.

Latha, C.B.C. and Jeeva, S.C., 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, p.100203.

Li, J.P., Haq, A.U., Din, S.U., Khan, J., Khan, A. and Saboor, A., 2020. Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, pp.107562-107582.

Martin-Isla, C., Campello, V.M., Izquierdo, C., Raisi-Estabragh, Z., Baeßler, B., Petersen, S.E. and Lekadir, K., 2020. Image-based cardiac diagnosis with machine learning: a review. *Frontiers in cardiovascular medicine*, p.1.

Mishra, S., Mallick, P.K., Tripathy, H.K., Bhoi, A.K. and González-Briones, A., 2020. Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Applied Sciences*, *10*(22), p.8137.

Quer, G., Arnaout, R., Henne, M. and Arnaout, R., 2021. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. *Journal of the American College of Cardiology*, *77*(3), pp.300-313.

Repaka, A.N., Ravikanti, S.D. and Franklin, R.G., 2019, April. Design and implementing heart disease prediction using naives Bayesian. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

Sarmah, S.S., 2020. An efficient IoT-based patient monitoring and heart disease prediction system using deep learning modified neural network. *Ieee access*, *8*, pp.135784-135797.

Segar, M.W., Patel, K.V., Ayers, C., Basit, M., Tang, W.W., Willett, D., Berry, J., Grodin, J.L. and Pandey, A., 2020. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning‐based unsupervised cluster analysis. *European journal of heart failure*, *22*(1), pp.148-158.

Siontis, K.C., Noseworthy, P.A., Attia, Z.I. and Friedman, P.A., 2021. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, *18*(7), pp.465-478.

Suri, J.S., Bhagawati, M., Paul, S., Protogeron, A., Sfikakis, P.P., Kitas, G.D., Khanna, N.N., Ruzsa, Z., Sharma, A.M., Saxena, S. and Faa, G., 2022. Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review. *Computers in biology and medicine*, p.105204.

Terrada, O., Cherradi, B., Raihani, A. and Bouattane, O., 2019, April. Classification and Prediction of atherosclerosis diseases using machine learning algorithms. In *2019 5th International Conference on Optimization and Applications (ICOA)* (pp. 1-5). IEEE.

Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), pp.1-16.

Windmon, A., Minakshi, M., Bharti, P., Chellappan, S., Johansson, M., Jenkins, B.A. and Athilingam, P.R., 2018. Tussiswatch: A smart-phone system to identify cough episodes as early symptoms of chronic obstructive pulmonary disease and congestive heart failure. *IEEE journal of biomedical and health informatics*, *23*(4), pp.1566-1573.

Ul Haq, Amin, JianPing Li, Muhammad Hammad Memon, Jalaluddin Khan, and Salah Ud Din. "A novel integrated diagnosis method for breast cancer detection." Journal of Intelligent & Fuzzy Systems 38, no. 2 (2020): 2383-2398.

Chumachenko, Dmytro, Ievgen Meniailov, Kseniia Bazilevych, Tetyana Chumachenko, and Sergey Yakovlev. "Investigation of Statistical Machine Learning Models for COVID-19 Epidemic Process Simulation: Random Forest, K-Nearest Neighbors, Gradient Boosting." Computation 10, no. 6 (2022): 86.

Deng, Haowen, Youyou Zhou, Lin Wang, and Cheng Zhang. "Ensemble learning for the early prediction of neonatal jaundice with genetic features." BMC medical informatics and decision making 21, no. 1 (2021): 1-11.

Amanoul, Sandy Victor, Adnan Mohsin Abdulazeez, Diyar Qader Zeebare, and Falah YH Ahmed. "Intrusion Detection Systems Based on Machine Learning Algorithms." In 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS), pp. 282-287. IEEE, 2021.

**Appendices**

**Libraries Imported**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

import os

import yellowbrick

import pickle

from matplotlib.collections import PathCollection

from statsmodels.graphics.gofplots import qqplot

from sklearn.preprocessing import MinMaxScaler

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

from sklearn.linear\_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.naive\_bayes import GaussianNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, ExtraTreesClassifier, VotingClassifier

from sklearn.metrics import classification\_report, accuracy\_score

from xgboost import XGBClassifier

from yellowbrick.classifier import PrecisionRecallCurve, ROCAUC, ConfusionMatrix

from yellowbrick.style import set\_palette

from yellowbrick.model\_selection import LearningCurve, FeatureImportances

from yellowbrick.contrib.wrapper import wrap

**Libraries Imported**

HD\_data = pd.read\_csv("/content/drive/MyDrive/Heart\_disease.csv")

print('\033[1m'+'.: Dataset Info :.'+'\033[0m')

print('\*' \* 30)

print('Total Rows:'+'\033[1m', HD\_data.shape[0])

print('\033[0m'+'Total Columns:'+'\033[1m', HD\_data.shape[1])

print('\033[0m'+'\*' \* 30)

print('\n')

print('\033[1m'+'.: Dataset Details :.'+'\033[0m')

print('\*' \* 30)

HD\_data.info(memory\_usage = False)

**.: Dataset Info :.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Total Rows: **1025**

Total Columns: **14**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**.: Dataset Details :.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1025 entries, 0 to 1024

Data columns (total 14 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 1025 non-null int64

1 sex 1025 non-null int64

2 cp 1025 non-null int64

3 trestbps 1025 non-null int64

4 chol 1025 non-null int64

5 fbs 1025 non-null int64

6 restecg 1025 non-null int64

7 thalach 1025 non-null int64

8 exang 1025 non-null int64

9 oldpeak 1025 non-null float64

10 slope 1025 non-null int64

11 ca 1025 non-null int64

12 thal 1025 non-null int64

13 target 1025 non-null int64

dtypes: float64(1), int64(13)

lst=['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal']

HD\_data[lst] = HD\_data[lst].astype(object)

**Initial Data Exploration**

# Gender

colors=['Green','Blue', 'Yellow', "Cyan", "Red"]

Head = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal', 'target' ]

Title = ['Sex (Gender) Distribution', 'Chest Pain Type Distribution', 'Fasting Blood Sugar Distribution','Resting Electrocardiographic Distribution','Exercise Induced Angina Distribution','Slope of the Peak Exercise Distribution','Number of Major Vessels Distribution','thal Distribution','Heart Diseases Distribution']

Xlabel = ['Gender','Pain Type','Fasting Blood Sugar','Resting Electrocardiographic','Exercise Induced Angina','Slope','Number of Major Vessels','Number of "thal','Heart Disease Status']

labels=[['Female', 'Male'],['Type 0', 'Type 2', 'Type 1', 'Type 3'],['< 120 mg/dl', '> 120 mg/dl'],['1', '0', '2'],['False', 'True'],['2', '1', '0'],['0', '1', '2', '3', '4'],['2', '3', '1', '0'],['True', 'False']]

for i in range (9):

  order = HD\_data[Head[i]].value\_counts().index

  # --- Size for Both Figures ---

  plt.figure(figsize=(16, 8))

  plt.suptitle(Title[i], fontweight='heavy',

              fontsize='16', fontfamily='sans-serif', color='#100C07')

  # --- Pie Chart ---

  plt.subplot(1, 2, 1)

  plt.title('Pie Chart', fontweight='bold', fontsize=14,

            fontfamily='sans-serif', color='#100C07')

  plt.pie(HD\_data[Head[i]].value\_counts(), labels=labels[i], colors=colors, pctdistance=0.7,

          autopct='%.2f%%',  textprops={'fontsize':12},wedgeprops=dict(alpha=0.8, edgecolor='#3E3B39'))

  centre=plt.Circle((0, 0), 0.45, fc='white', edgecolor= '#3E3B39')

  plt.gcf().gca().add\_artist(centre)

  # --- Histogram ---

  countplt = plt.subplot(1, 2, 2)

  plt.title('Histogram', fontweight='bold', fontsize=14,

            fontfamily='sans-serif', color='#100C07')

  ax = sns.countplot(x=Head[i], data=HD\_data, palette=colors, order=order,

                    edgecolor='#6D6A6A', alpha=0.85)

  for rect in ax.patches:

      ax.text (rect.get\_x()+rect.get\_width()/2,

              rect.get\_height()+4.25,rect.get\_height(),

              horizontalalignment='center', fontsize=10,

              bbox=dict(facecolor='none', edgecolor='#100C07',

                        linewidth=0.25, boxstyle='round'))

  plt.xlabel(Xlabel[i], fontweight='bold', fontsize=11, fontfamily='sans-serif',

            color='#3E3B39')

  plt.ylabel('Total', fontweight='bold', fontsize=11, fontfamily='sans-serif',

            color='#3E3B39')

  #plt.xticks([0, 1],labels[i])

  plt.grid(axis='y', alpha=0.4)

  countplt

# ****Numerical Variable****

**Descriptive Statistics**

HD\_data.select\_dtypes(exclude='object').describe()

|  | **age** | **trestbps** | **chol** | **thalach** | **oldpeak** | **target** |
| --- | --- | --- | --- | --- | --- | --- |
| **count** | 1025.000000 | 1025.000000 | 1025.00000 | 1025.000000 | 1025.000000 | 1025.000000 |
| **mean** | 54.434146 | 131.611707 | 246.00000 | 149.114146 | 1.071512 | 0.513171 |
| **std** | 9.072290 | 17.516718 | 51.59251 | 23.005724 | 1.175053 | 0.500070 |
| **min** | 29.000000 | 94.000000 | 126.00000 | 71.000000 | 0.000000 | 0.000000 |
| **25%** | 48.000000 | 120.000000 | 211.00000 | 132.000000 | 0.000000 | 0.000000 |
| **50%** | 56.000000 | 130.000000 | 240.00000 | 152.000000 | 0.800000 | 1.000000 |
| **75%** | 61.000000 | 140.000000 | 275.00000 | 166.000000 | 1.800000 | 1.000000 |
| **max** | 77.000000 | 200.000000 | 564.00000 | 202.000000 | 6.200000 | 1.000000 |

Describe = ['age','trestbps','chol','thalach','oldpeak']

colors=['#FF0000','#8A0030', '#4C0028', "#F38BB2", "#400000"]

Ytitle = ['Age','Resting Blood Pressure','Serum Cholestoral','Maximum Heart Rate','oldpeak']

for i in range(len(Describe)):

  # --- Variable, Color & Plot Size ---

  var = Describe[i]

  color = colors[i]

  fig=plt.figure(figsize=(12, 12))

  # --- Skewness & Kurtosis ---

  print('\033[1m'+'.: Age Column Skewness & Kurtosis :.'+'\033[0m')

  print('\*' \* 40)

  print('Skewness:'+'\033[1m {:.3f}'.format(HD\_data[var].skew(axis = 0, skipna = True)))

  print('\033[0m'+'Kurtosis:'+'\033[1m {:.3f}'.format(HD\_data[var].kurt(axis = 0, skipna = True)))

  print('\n')

  # --- General Title ---

  fig.suptitle('Age Column Distribution', fontweight='bold', fontsize=16,

              fontfamily='sans-serif', color='#100C07')

  fig.subplots\_adjust(top=0.9)

  # --- Histogram ---

  ax\_1=fig.add\_subplot(2, 2, 2)

  plt.title('Histogram Plot', fontweight='bold', fontsize=14,

            fontfamily='sans-serif', color='#3E3B39')

  sns.histplot(data=HD\_data, x=var, kde=True, color=color)

  plt.xlabel('Total', fontweight='regular', fontsize=11,

            fontfamily='sans-serif', color='#3E3B39')

  plt.ylabel(Ytitle[i], fontweight='regular', fontsize=11, fontfamily='sans-serif',

            color='#3E3B39')

  # --- Q-Q Plot ---

  ax\_2=fig.add\_subplot(2, 2, 4)

  plt.title('Q-Q Plot', fontweight='bold', fontsize=14,

            fontfamily='sans-serif', color='#3E3B39')

  qqplot(HD\_data[var], fit=True, line='45', ax=ax\_2, markerfacecolor=color,

        markeredgecolor=color, alpha=0.6)

  plt.xlabel('Theoritical Quantiles', fontweight='regular', fontsize=11,

            fontfamily='sans-serif', color='#3E3B39')

  plt.ylabel('Sample Quantiles', fontweight='regular', fontsize=11,

            fontfamily='sans-serif', color='#3E3B39')

  # --- Box Plot ---

  ax\_3=fig.add\_subplot(1, 2, 1)

  plt.title('Box Plot', fontweight='bold', fontsize=14, fontfamily='sans-serif',

            color='#3E3B39')

  sns.boxplot(data=HD\_data, y=var, color=color, boxprops=dict(alpha=0.8), linewidth=1.5)

  plt.ylabel(Ytitle[i], fontweight='regular', fontsize=11, fontfamily='sans-serif',

            color='#3E3B39')

  plt.show()

# ****EDA****

**Heart Disease Distribution based on Gender**

labels = ['False', 'True']

label\_gender = np.array([0, 1])

label\_gender2 = ['Male', 'Female']

# --- Creating Bar Chart ---

ax = pd.crosstab(HD\_data.sex, HD\_data.target).plot(kind='bar', figsize=(8, 5),

                                         color=['#FFD7D7', '#F17881'],

                                         edgecolor='#6D6A6A', alpha=0.85)

# --- Bar Chart Settings ---

for rect in ax.patches:

    ax.text (rect.get\_x()+rect.get\_width()/2,

             rect.get\_height()+1.25,rect.get\_height(),

             horizontalalignment='center', fontsize=10)

plt.suptitle('Heart Disease Distribution based on Gender', fontweight='heavy',

             x=0.01, y=0.1, ha='left', fontsize='16', fontfamily='sans-serif',

             color='#100C07')

plt.title('Female tend to have heart diseases compared to Male. In male, the distribution is not imbalanced compared to female\nthat have almost the same distribution',

          fontsize='8', fontfamily='sans-serif', loc='left', color='#3E3B39')

plt.tight\_layout(rect=[0, 0.10, 1, 1.025])

plt.xlabel('Gender (Sex)', fontfamily='sans-serif', fontweight='bold',

           color='#3E3B39')

plt.ylabel('Total', fontfamily='sans-serif', fontweight='bold',

           color='#3E3B39')

plt.xticks(label\_gender, label\_gender2, rotation=0)

plt.grid(axis='y', alpha=0.4)

plt.grid(axis='x', alpha=0)

plt.legend(labels=labels, title='$\\bf{Target}$', fontsize='10',

           title\_fontsize='9', loc='upper left', frameon=True);

**Heart Disease Distribution based on Major Vessels Total**

# --- Labels Settings ---

labels = ['False', 'True']

# --- Creating Horizontal Bar Chart ---

ax = pd.crosstab(HD\_data.ca, HD\_data.target).plot(kind='barh', figsize=(8, 5), fontsize='10',

                                         color=[ '#FFD7D7', '#F17881'],

                                         edgecolor='#6D6A6A', alpha=0.85)

# --- Horizontal Bar Chart Settings ---

for rect in ax.patches:

    width, height = rect.get\_width(), rect.get\_height()

    x, y = rect.get\_xy()

    ax.text (x+width/4, y+height/4, '{:.0f}'.format(width), fontsize='8',

             horizontalalignment='left', verticalalignment='center')

plt.suptitle('Heart Disease Distribution based on Major Vessels Total',

             fontweight='heavy', x=0.01, y=0.01, ha='left', fontsize='16', horizontalalignment='left',

             fontfamily='sans-serif', color='#100C07')

plt.title('Patients with 0 and 4 major vessels tend to have heart diseases. However, patients who have a number of vessels 1 to 3\ntend not to have heart diseases.',

          fontsize='8', fontfamily='sans-serif', loc='left', color='#3E3B39')

plt.tight\_layout(rect=[0, 0.04, 1, 1.025])

plt.xlabel('Total', fontfamily='sans-serif', fontweight='bold', fontsize='8', color='#3E3B39')

plt.ylabel('Number of Major Vessels', fontfamily='sans-serif', fontweight='bold',

           color='#3E3B39')

plt.yticks(rotation=0)

plt.grid(axis='x', alpha=0.4)

plt.grid(axis='y', alpha=0)

plt.legend(labels=labels, title='$\\bf{Target}$', fontsize='10', frameon=True,

           title\_fontsize='10', loc='upper right');

**Heart Disease Scatter Plot based on Age**

# -- Scatter Plot Size & Titles Settings ---

plt.figure(figsize=(5, 4))

plt.suptitle('Heart Disease Scatter Plot based on Age', fontweight='heavy',

             x=0.04, y=0.01, ha='left', fontsize='16', fontfamily='sans-serif',

             color='#100C07')

plt.title('Based on age, patients with and without heart diseases mostly between 50-70 years old. Patients with heart diseases tend to have high heart rate\ncompared to patients with no heart diseases.',

          fontsize='8', fontfamily='sans-serif', loc='left', color='#3E3B39')

plt.tight\_layout(rect=[0, 0.04, 1, 1.01])

# --- Creating Scatter Plot ---

plt.scatter(x=HD\_data.age[HD\_data.target==0], y=HD\_data.thalach[(HD\_data.target==0)], c='#8A0030')

plt.scatter(x=HD\_data.age[HD\_data.target==1], y=HD\_data.thalach[(HD\_data.target==1)], c='#FF5C8A')

# --- Scatter Plot Legend & Labels Settings ---

plt.legend(['False', 'True'], title='$\\bf{Type}$', fontsize='7',

           title\_fontsize='8', loc='upper right', frameon=True)

plt.xlabel('Age', fontweight='bold', fontsize='11',

           fontfamily='sans-serif', color='#3E3B39')

plt.ylabel('Max. Heart Rate', fontweight='bold', fontsize='11',

           fontfamily='sans-serif', color='#3E3B39')

plt.ticklabel\_format(style='plain', axis='both')

plt.grid(axis='both', alpha=0.4, lw=0.5)

plt.show();

**Heatmap**

# --- Correlation Map (Heatmap) ---

plt.figure(figsize=(7, 4.5))

sns.heatmap(HD\_data.corr(), annot=True, cmap='Reds', linewidths=0.1)

plt.suptitle('Correlation Map of Numerical Variables', fontweight='heavy',

             x=0.03, y=0.03, ha='left', fontsize='16', fontfamily='sans-serif',

             color='#100C07')

plt.title('Resting blood pressure, cholestoral, and "oldpeak" have moderate relationship with age.',

          fontsize='10', fontfamily='sans-serif', loc='left', color='#3E3B39')

plt.tight\_layout(rect=[0, 0.04, 1, 1.01])

# ****Dataset Pre-processing****

**One-Hot Encoding**

# --- Creating Dummy Variables for cp, thal and slope ---

cp = pd.get\_dummies(HD\_data['cp'], prefix='cp')

thal = pd.get\_dummies(HD\_data['thal'], prefix='thal')

slope = pd.get\_dummies(HD\_data['slope'], prefix='slope')

# --- Merge Dummy Variables to Main Data Frame ---

frames = [HD\_data, cp, thal, slope]

HD\_data = pd.concat(frames, axis = 1)

HD\_data.head()

**Dropping Unnecessary Variables**

# --- Drop Unnecessary Variables ---

HD\_data = HD\_data.drop(columns = ['cp', 'thal', 'slope'])

HD\_data.head()

**Spiliting and Normalizing the data**

# --- Seperating Dependent Features ---

x = HD\_data.drop(['target'], axis=1)

y = HD\_data['target']

# --- Data Normalization using Min-Max Method ---

x = MinMaxScaler().fit\_transform(x)

# --- Splitting Dataset into 80:20 ---

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=4)

# ****Model Implementation****

**Logistic Regression**

# --- Applying Logistic Regression ---

Models = ['LogisticRegression', 'KNeighborsClassifier','RandomForestClassifier','GradientBoostingClassifier']

classifiers = [LogisticRegression(max\_iter=1000, random\_state=1, solver='liblinear', penalty='l1'), KNeighborsClassifier(n\_neighbors=3),

               RandomForestClassifier(n\_estimators=1000, random\_state=1, max\_leaf\_nodes=20, min\_samples\_split=15),

               GradientBoostingClassifier(random\_state=1, n\_estimators=100, max\_leaf\_nodes=3, loss='exponential', min\_samples\_leaf=20)]

print('Model Accuracies: ')

Accuracy = []

for i in range(len(classifiers)):

  classifier = classifiers[i]

  classifier.fit(x\_train, y\_train)

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

  Acc = accuracy\_score(y\_pred, y\_test)

  Accuracy.append(Acc)

  print('\n' + Models[i]+':'+'\033[1m {:.2f}%'.format(Acc\*100))

**Model Comparison**

# --- Create Accuracy Comparison Table ---

compare = pd.DataFrame({'Model': ['Logistic Regression', 'K-Nearest Neighbour',

                                  'Random Forest', 'Gradient Boosting'

                                  ],

                        'Accuracy': [Accuracy[0]\*100, Accuracy[1]\*100, Accuracy[2]\*100, Accuracy[3]\*100]})

print("Highest accuracy we observed in: "+ "\033[1m"+Models[pd.Series(Accuracy).idxmax()]+"\033[0m" )

# --- Create Accuracy Comparison Table ---

compare.sort\_values(by='Accuracy', ascending=False).style.background\_gradient(cmap='PuRd').hide\_index().set\_properties(\*\*{'font-family': 'Segoe UI'})

# ****Hybrid ML Model****

Hybrid = VotingClassifier(estimators=[('knn', KNeighborsClassifier(n\_neighbors=2)), ('rf', RandomForestClassifier(n\_estimators=2000, random\_state=1, max\_leaf\_nodes=20, min\_samples\_split=20))

, ('gnb', GradientBoostingClassifier(random\_state=1, n\_estimators=2000, max\_leaf\_nodes=3, loss='exponential', min\_samples\_leaf=20))], voting='soft')

Hybrid = Hybrid.fit(x\_train, y\_train)

y\_pred\_hybrid = Hybrid.predict(x\_test)

Acc\_hybrid = accuracy\_score(y\_pred\_hybrid, y\_test)

print('Hybrid\_Model accuracy:'+'\033[1m {:.2f}%'.format(Acc\_hybrid\*100))