OCTOBER 2023 M.Sc. in Electrical and Electronics Engineering EBRIMA JALLOW

REPUBLIC OF TÜRKİYE

GAZİANTEP UNIVERSITY

GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES

HEART ATTACK PREDICTION USING MACHINE LEARNING

M.Sc.

IN

ELECTRICAL AND ELECTRONIC ENGINEERING

BY

EBRIMA JALLOW

OCTOBER 2023

HEART ATTACK PREDICTION USING MACHINE LEARNING

M.Sc.

in

Electrical and Electronic Engineering
Gaziantep University

Supervisor

Prof. Dr. Ergun ERÇELEBI

by

Ebrima JALLOW

October 2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ebrima JALLOW

ABSTRACT

HEART ATTACK PREDICTION USING MACHINE LEARNING

JALLOW, Ebrima
M.Sc. in Electrical and Electronic Engineering
Supervisor: Prof. Dr. Ergun ERCELEBI
October 2023

52 pages

According to the World Health Organization (WHO) 2020 report, more than 17.9 million people lose their lives due to heart attacks each year, making it one of the deadliest diseases worldwide; furthermore, heart attack is classified as a cardiovascular disease that affects the lives of many people globally. Machine learning, vital to the healthcare industry, enhances people's quality of life along with scientific advancements. In this thesis, studies on predicting heart attacks using machine learning algorithms were conducted. The data used in the studies were obtained from the UCI Machine Learning database. The prediction of heart attacks was carried out using five well-known machine learning models mentioned in the literature. Additionally, a new conceptual model integrating data preprocessing, hyperparameter tuning, and Logistic Regression was proposed for predicting heart attacks in this thesis. Tests conducted on the proposed model yielded satisfactory results, achieving an accuracy rate of 93%. Furthermore, the results produced by the proposed model were compared with the outcomes of other machine learning algorithms, demonstrating that the proposed model outperforms other algorithms.

Key Words: Heart Attacks**,** Artificial Intelligence, Machine Learning, Data Pre-processing, Hyperparameter Tuning

ÖZET

MAKINE OĞRENME KULLANARAK KALP KRIZ TAHMIN

JALLOW, Ebrima
Yüksek Lisans Tezi, Elektrik ve Elektronik Mühendişliği
Danışman: Prof.Dr Ergun ERCELEBI
Ekim 2023
52 sayfa

Dünya Sağlık Örgütü (DSÖ) 2020 raporuna göre, her yıl 17.9 milyondan fazla insan kalp krizinden hayatını kaybetmektedir, bu da onu dünyadaki en ölümcül hastalıklardan biri yapmaktadır; ayrıca kalp krizi, dünya genelinde birçok insanın hayatını etkileyen kardiyovasküler hastalık olarak nitelendirilmektedir. Makine öğrenimi, sağlık endüstrisi için hayati öneme sahip olup bilimsel gelişmelerle birlikte insanların yaşam kalitesini artırmaktadır. Bu tezde, makine öğrenimi algoritmaları kullanılarak kalp krizinin tahminine yönelik çalışmalar yapılmıştır. Çalışmalarda kullanılan veriler UCI Makine Öğrenme veri tabanından alınmıştır. Literatürde bilinen beş makine öğrenimi modeli kullanılarak kalp krizinin tahminine yönelik yapılmıştır. Ayrıca tezde, kalp krizinin tahminine yönelik veri ön işleme, hiperparametre ayarlama ve Logistic regresyonun bütünleşmesinden oluşan kavramsal yeni bir model önerilmiştir. Önerilen model üzerinde yapılan testler tatmin edici sonuçlar üretmiş ve %93'lük doğruluk sonucuna ulaşılmıştır. İlaveten önerilen modelin ürettiği sonuçlar diğer makine öğrenmesi algoritmalarının sonuçları ile karşılaştırılmış ve önerilen modelin diğer algoritmalardan daha iyi sonuçlar ürettiği gösterilmiştir.

Anahtar Kelimeler: Kalp Krizleri, Yapay Zeka, Makine Öğrenme, Veri Ön Işleme, Hiperparametre Ayarlaması.

"Dedicated to my family"

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Dr. Ergun ERCELEBI for his guidance and support throughout the study. I am thankful for his encouragement and motivation.

I would like to express my love and gratitude to my family for their support, always best wishes.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

L_B	Lower Boundary
P(A)	Probability of an event before evident is known
P(B)	Probability based on evidence
$P\left(\frac{A}{B}\right)$	For evidence based probability that hypothesis is true
Q_1	First Quartile
Q_2	Second Quartile
Q_3	Third Quartile
U_B	Upper Boundary

LIST OFABBREVIATIONS

AI	Artificial Intelligence
CAA	Number of Major Vessels
CHOL	Cholesterol
CM	Confusion Matrix
CP	Chest Pain Type
CV	Cross Validation
CVD	Cardiovascular Disease
EDA	Exploratory Data Analysis
EXNG	Exercised Induced Angina
FBS	Fasting Blood Sugar
FN	False Negative
FP	False Positive
IQR	Interquartile Range
KNN	K-Nearest Neighbor
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
RF	Random Forest
ROC	Receiver Operating Characteristic
SLP	Slope
SVR	Support Vector Machine
THALACHH	Maximum Heart Rate Achieve
THALL	Thallanium
TN	True Negative
TP	True Positive
TRTBPS	Resting Blood Sugar
UCI	University of California Irvine
WHO	World Health Organization

CHAPTER 1
INTRODUCTION

1.1 Motivation of Study

Heart attack which is mainly referred to as acute myocardial infarction (AMI) is one of the most dangerous of cardiovascular diseases (CVD) and takes place as a result of the heart muscle being damaged by cessation of blood flow to the muscle [1]. According to World Health Organization (WHO) in 2020, cardiovascular diseases are now the leading cause of mortality globally accounting for 17.9 million fatalities annually [2]. Some of the symptoms of heart attacks are nausea, irregular heartbeat, vomiting, fatigue, sweating, anxiety and chestpain which travels to the leftside of the neck or leftside of the arm [3]. Life styles such as smoking, diet and lack of exercise increases the chances of having heart attack.

Both men and women are equally vulnerable of having heart attack [4]. However, some studies disagrees with this information by stating that heart attacks increases with age and affects men more than women [3], [5]. A dataset obtained from The UCI Machine Learning Repository that is used in this study showed that women have higher chances of having heart attack than men. Traditional methods such as electrocardiogram, stress test, echocardiography, magnetic resonance imaging, coronary angiography or intracoronary ultrasound has been used to detect cardiovascular diseases and the importance of early detection is significant [6]. Technologies such as Information and Communication Technologies and Artificial Intelligence (AI) are rapidly rising and are capable of predicting more accurate and faster outcome than experts which will soon transform the heart health sector [6].

AI is the ability of a computer program to develop intelligence that resembles that of the human brain, it enables thought in computers which increases their intelligence, it was proposed by Alex Mathison Turing almost 60 years ago, he stated that if the answers or solutions given by computers cannot be distinguish from human then the machine is referred to as intelligent [7].

The general feeling among people is that AI is capable of transforming our livelihoods and may even go on to perform task that experts in domains such as health sector cannot do and thus replacing them. It is not easy for experts to predict heart attack base on their own experiences which may be limited compare to large datas. However, with AI when fed with the right data it can easily predict heart attack no matter how big the data is and can predict more accurately than experts. For this reasons it is important for us to widen our scope in this field. AI can help experts to be able to predict quicker and faster thus leading to early detection of heart attack and leading to cure. It will also save a lot of time for the doctors by making their work easier and more efficient. Early detection can lead to safe of life and reduction of cost.

Machine learning is a branch of AI which has gain high popularity among researchers in all works of life. Traditional machine learning algorithms such as random forest classifier and logistic regression are mainly used for numerical datasets. This research is focused in this direction. Another type of advance machine learning algorithms is refers to as deep learning which performs better on unstructured datasets such as images and audios.

The primary problem of machine learning models is that feature engineering is frequently needed for optimal implementation which is a time consuming effort [8]. Feature engineering is used to handle missing values, outliers, future extraction and future selection etc.

Data analysis plays a huge role in AI. It gives insights about data and hidden informations and patterns that is very difficult to see with the naked eye. This can involve using statistics by looking at the relationship between features and labels in a dataset no matter how big or small the data is.

The main purpose of this study is to use a heart disease dataset obtained from UCI Machine Learning Repository and use it to extract information and also use machine learning methods to train this dataset. The best model is chosen and can be used to predict heart attacks based on this dataset within seconds. This in turn could be used by doctors to make faster and early diagnosis there by saving a lot of time and cost.

1.2 Literature Review

Several studies have been done is this area where by different authors obtained different probabilities of making the right predictions based on their datasets and different models have been proposed. The papers used in this studies ranges from 2014 to date and the limitations of some of the papers will also be discuss in this part.

Nandal et al. [1]; used four ML models such as Logistic Regression, Naïve Bayes, SVM and XGBoost classifiers to make heart attack predictions and at the end of the study, the authors proposed XGBoost classifier as the best model to predict heart attack with an accuracy of 92%. They collected their data from UCI repository which is publicly available. The data set contain 294 tuples with 14 attributes. The authors used data pre-processing to clean the dataset by handle missing values which was replaced by their mean values, duplicates values were also removed, outlier detection and handling with their minimum and maximum ranges. The limitation of this study is that it did not mention feature selection and handling which is important in selecting the best features to increase accuracy, also since the dataset contained categorical values no hot encoding was performed which converts categorical values to either zero or one and finally hyperparameter tuning was not conducted to try to increase the accuracy of the models used.

The Framingham dataset and UCI heart repository dataset were used by Gupta et al. [7]; to make a heart attack prediction model. They used both datasets and trained four models which are Gradient Boosting, Decision Tree, Logistic Regression and Random Forest algorithm. They split their datasets into training and testing sets, whereby 90% of the data was used for training and 10% of the data was used for testing. The authors first trained their model without optimization and feature transformations and vice-versa. Gradient Boosting algorithm achieved the highest accuracy of 85.5% in the framingham dataset that included both optimization and feature engineering. It also achieves the highest accuracy in the UCI repository dataset including both optimization and feature engineering with an accuracy of 85.6% with Random Forest Classifier. The shortcomings in this study include the fact that it did not mention handling outliers in the datasets, if any, and it also did not address data normalization for scaling the dataset. Additionally, the study did not discuss the process of hot

encoding the categorical data into dummy variables.

In another study Wang et al. [9]; studied about an imbalanced data-preprocessing algorithm for heart attack prediction in stroke patients. They used the Medical Information Mart for Intensive Care III dataset which is known as MIMIC III. This dataset is an imbalanced data and under sampling clustering oversampling (UCO) algorithm was used as it produces nearly a very balance dataset. Under sampling was performed due to the lower number of people with heart attacks compared to those without heart attacks. A sample size of 120 was selected, hence referred to as UCO(120), which demonstrated superior accuracy in handling unbalanced datasets compared to oversampling and clustering algorithms. The sample numbers were divided into a training set comprising 80%, a validation set comprising 10%, and a testing set comprising 10%. Five models were trained such as Random Forest, Adaboost , KNN, Decision Tree and SVM Classifiers and results showed that Random Forest Classifiers achieved the best accuracy with 70.29% and a precision score of 70.05%. The drawbacks of this paper was that hyperparameter tuning was not perform to increase the accuracy of the models and did not talk about outlier handling if there are any. Also they did not mention encoding categorical values if there is any.

A range of feature combinations and well know classification techniques were utilized to introduce a prediction model, Mohan et al. [10]; suggested using a hybrid random forest with a linear model (HRFLM). In order to improve prediction accuracy, this study uses machine learning approaches to identify important attributes. The authors proposed model achieved a maximum accuracy of 88.7% using the UCI dataset.

Takçi [11]; used feature selection methods to improve heart attack prediction. At the end of the study the paper concluded that SVM with linear kernel is the best machine learning model. It also mentioned that reliefF method was the best feature selection algorithm. The combination of both of them gave the best accuracy result of 84.81%.

In order to boost prediction ability of machine learning in heart attack, Dai et al. [12]; used the kernel-lasso feature expansion method which improved the prediction ability of SVM and DNN. R programming was use and it concluded that SVM with kernel-lasso feature expansion got the best accuracy of (84(%95C1:0.78-0.89)).

Ranga and Rohila [13]; trained five classification algorithms and two clustering methods, which were then simulated in WEKA using the UCI dataset. The dataset was split into a 60% training set and a 40% testing set. The classification algorithms employed were SVM, KNN, ANN, Decision Tree, and Random Forest Classifiers, while K-means and Expectation-maximizing clustering were utilized for clustering. Random Forest Classification exhibited the highest accuracy at 85.95%. However, the limitation of this study lies in the author's omission of hyperparameter tuning to enhance accuracy. Additionally, they did not perform essential feature engineering, a critical aspect in machine learning.

Hasan and Saleh [14]; uses the NHANES dataset which contained 37,080 records of diverse people and contained 50 features of each individual and a target value to make a heart attack prediction model that was based on hyperparameter tuning. The dataset was an unbalanced data which was handled and then divided into training set of 80% and a testing set of 20%. The four classification algorithms used were Naive Bayes, Decision Tree, XGBoost and CatBoost classification algorithms. The authors proposed Hperopt Optimization technique as the best technique to use in hyperparameter tuning and concluded that CatBoost classification algorithm as the best model with an accuracy of 97.1%. The constraints of this paper is since their dataset contained numerical values the data was not scaled and categorical values should have been hot encorded to convert categorical values to numerical values.

Similarly a feature selection approach (FCMIM-SVM) that is appropriate for support vector machine was introduced in another paper [15]. Their model was able to reach an accuracy of 92.37% after performing hyperparameter tuning.

Machine learning is just one of the ways in which studies have been made in order to make prediction of diseases. Data mining techniques have been also used by some studies [16], [17] and [18]; to make prediction of heart attack, heart diseases and diabetic prediction respectively.

Big data analytics has been used by Alexander and Wang [19]; to make heart attack prediction while Maihami and Rahimi [20]; used fuzzy systems to make heart attack prediction. Studies in this area generally have limitations especially conducting proper

feature engineering, hyperparameter tuning to increase accuracy of the model and also dealing with unbalanced datasets.

1.3 Problem Statement and Contribution of Thesis

One of the ways in which heart attacks can be prevented involves medicine, changing of lifestyle, and early detection of the disease. Early detection is vital in preventing heart attacks. Doctors have a huge challenge in being able to predict heart attacks base on their knowledge which may differ from person to person and may cause biasness. The problem that needs to be address is how to predict heart attack using a heart disease dataset.

The main objective of this study is to be able to predict whether a patient stands a chance of having a heart attack or not with a high accuracy and precision using a heart disease dataset. Datasets may have a lot of information that one can analyze and discover trends and hidden relationships in a data. This study may help doctors to predict heart attacks quickly and efficiently which can prevent bias and reduce cost.

1.4 Organization of Thesis

This study is made up of five (5) different chapters. Chapter one (1) has already been discussed above which involves motivation of study as to why this study has been done, literature review which evaluate the work that has been done by other researchers in the same field, their techniques and outlining their limitations. Finally problem statements and contribution of this research is outlined.

In chapter two (2), machine learning is discussed which includes the types of machine learning, including the techniques to be used in this study.

In chapter three (3), methodology used in the thesis is discussed in details. Starting from data collection, data-preprocessing, training and testing of data, evaluating metrics, hyperparameter tuning and finally model selection.

In chapter four (4), the results of the thesis is given in details and further compared with results of other techniques that were performed by other researchers.

Finally, in Chapter Five (5), the thesis will be concluded, and future work will be discussed.

CHAPTER 2
MACHINE LEARNING

2.1 Machine Learning

AI was proposed almost 60 years ago by Alex Mathison Turing [7]. Machine Learning is a subset of AI which allows it to learn from data rather than programming it carefully. To improve, clarify and anticipate outcomes using a variety of algorithms, Machine learning is a complicated algorithm that continues to learn [21]. In a nutshell, Machine Learning uses algorithm to try to learn about patterns of a particular data and uses what it learns to make a future prediction using a similar data. It looks at the input and output of a data and tries to figure out a set of instruction in between the input and output. Machine learning is a very complex and vast field whereby its scope and application is increasing daily [22]. Machine learning is mainly based on two principles that is how to build computer systems that will automatically improve outcomes based on prior experiences and the fundamental statistics, computational, and information theoretical laws that controls all learning systems, including those of humans, computers and organizations [21]. Machine learning algorithms is divided into three categories namely Supervised machine learning, Unsupervised machine learning and Reinforcement learning [13]. Supervise machine learning are referred to as prediction task since the objective is to predict or categorize a certain outcome of interest such as present or absent of a particular disease Jiang et al. [23]; Supervised machine learning consists of label data and also has an output and it is used for making classifications or predictions. Unsupervised machine learning has been developed to find novel patterns and relationships in regularly sampled data without the use of label data [24]. An unsupervised machine is used to try to understand relationships that are found in a dataset. In Reinforcement learning model which consists of an agent and an environment engage in interaction across a sequence of events. Based on a policy and an observable environmental state vector, an agent may decide what to do during each turn. In response to a chosen action, the environment creates a reward as well as a new environmental state vector.

When choosing a new action based on policy, the agent is given feedback on the reward and the environment's observed condition [25]. This policy of reward or punishment will go on until the intended target is achieved. This research is about heart attack prediction which is a classification problem and consists of label data. Therefore, supervised machine learning approach has been used since it suits our problem. Supervised Machine learning consists of many algorithms. However, only those that are used in this study will be discuss below.

2.2 Logistic Regression

Logistic regression is an effective and well known technique in supervised machine learning algorithm [2]. It is a machine learning approach that is developed from the field of statistics [5]. Although "regression" appears in the name, this algorithm is not one that uses regression [26]. It indicates the likelihood of occurrence or non-occurrence of specific instance and is an extension of standard regression modelling when applied to a dataset [27]. This method is useful for categorizing values into two groups, such as logistic regression and linear regression in binary classification. Here, a non-linear function, different from linear regression, is used by the logistic function to build output prediction. The values of its properties vary from 0 to 1 [5]. Logistic Regression can also be used for multi-class classification [7]. The logistic function or sigmoid can be defined as:

$$logistic\,(z) = \frac{1}{1+exp^{-z}} \qquad 2.1$$

The diagram below, Figure 2.1, displays a sigmoid function ranging from -20 to +20 on the X-axis, which represents the independent features or inputs. These independent features produce an output on the Y-axis, referred to as the dependent features. The sigmoid function, also known as the logistic function, produces outputs ranging from 0 to 1, making it suitable for binary classification.

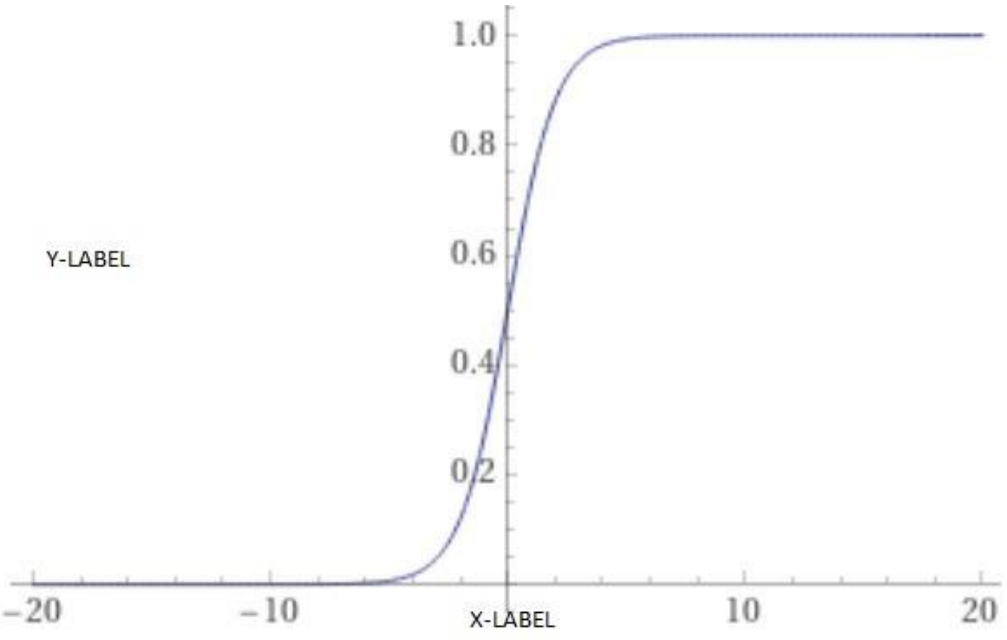

Figure 2.1. Logistic function of output y ranges from 0 to 1.

2.3 K-Nearest Neighbor

One of the earliest and most basic classification techniques is the KNN algorithm [2], [27]. The KNN rule was developed by Hodges et al. in 1951 as a non-parametric pattern categorization technique [5]. KNN algorithm is regarded as one of the lazy learning techniques [28]. The number of nearest neighbors considered in the KNN method is represented by the letter 'k' [13]. KNN is exceptional since it is straightforward and produces great results when it is applied to multiple problems [29]. The nearest k training samples in the feature space are used as the input vector for this method. The method finds the most prevalent class among the k closest neighbors to carry out the categorization [30]. There are several methods to choose the values for "k," but one can just run the classifier with various values several times to determine which value produces the best results [31]. The distance between two training and testing items is often determined using the Euclidean distance formula [28].

$$dxy = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad 2.2$$

Figure 2.2 shows how the nearest neighbors of a data point are used to categorize it before the votes of those neighbors are tallied. A query point is then assigned to the data class that has the most representation among its nearest neighbors [5].

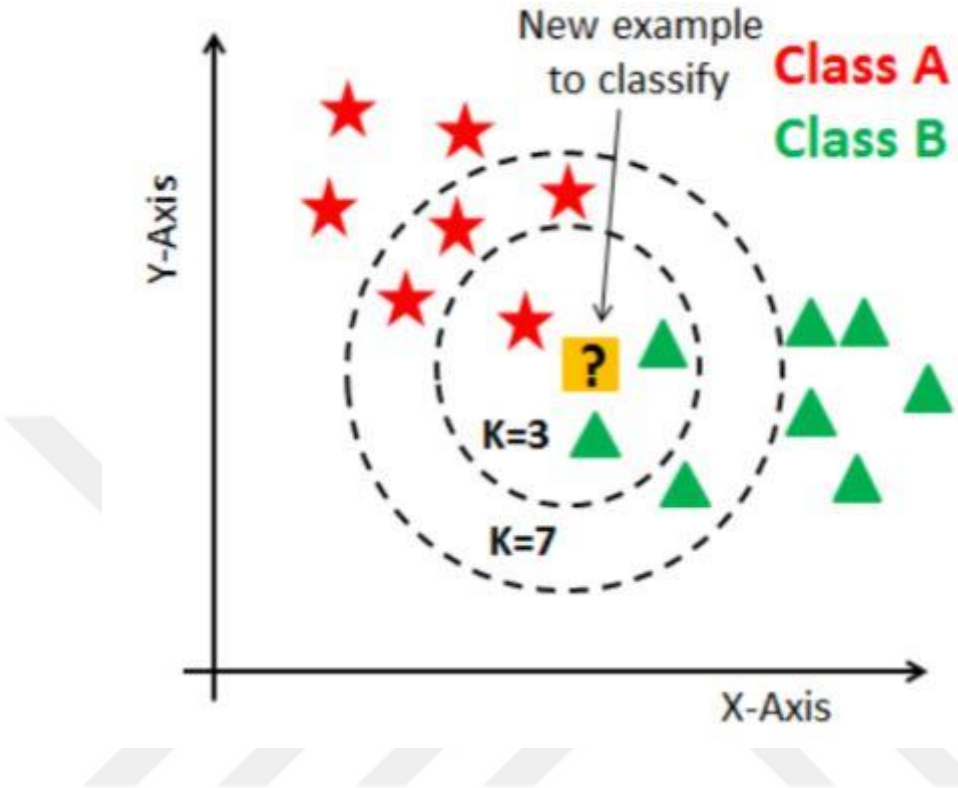

Figure 2.2. KNN before voting takes place [5].

2.4 Support Vector Machine

Support vector machine is a supervised machined learning technique that was developed by Vapnik in 1995 [5]. Support vector machines are designed for classification in which there are two classes (for example, cases or controls in a multidimensional space, or across many variables) [23]. Later, this was expanded to cover continuous outcomes and classification with more than two classes [32]. Support vector machines (SVMs) locate a hyperplane, which is a border that maximally divides classes (i.e., outcome categories), in order to distinguish between groups of individuals. To discover a separate decision surface, support vector machines use a data transformation that projects the input into a higher dimensional space. A kernel function is the term used to describe this procedure [23].

The algorithm's goal is to increase the margin of the hyperplanes, which reduces the issue with misclassification. To establish the decision boundary, the model selects

extreme points, known as support vectors [1]. Figure 2.3 shows how a simple SVM works. It tries to spot a hyperplane which increases the separation between the 'star' classes and the 'circle' classes.

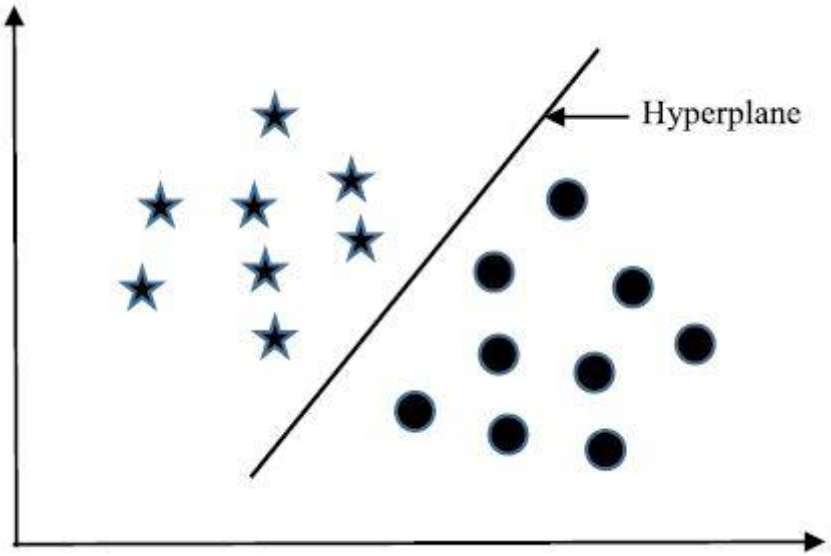

Figure 2.3. Hyperplane between two classes [27].

2.5 Random Forest

Random forest is a prominent machine learning technique that may be used to develop prediction models and was first introduced by Breiman in 2001 [33]. Random forest is a supervised machine learning technique that can handle both classification and regression tasks and it handles missing values, outlier values and uses dimensionality reduction techniques [34]. It involves combining several separated decision trees that are made in training time which functions as ensemble training [7]. These trees are produced using a variety of training set samples. Following that, a prediction is made using these decision trees majority voting and this algorithm is prone to overfitting [8]. Figure 2.4 is an example of random forest classification that contained three decision trees and each of the decision trees are trained using a random subset of the training data.

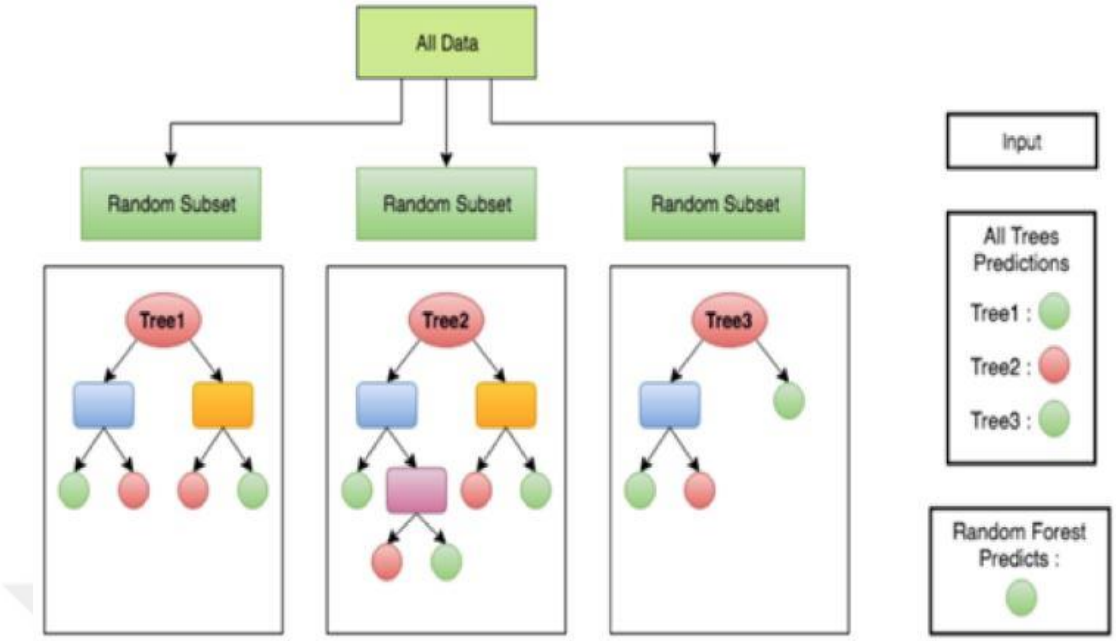

Figure 2.4. Random Forest Classification example [26].

2.6 Naive Bayes

Naive bayes is a bayes theorem based classification technique [5], [8], [27]. This theorem can estimate the likelihood of an event based on previous knowledge of the conditions related to that event [27]**.** From a classification standpoint, the fundamental objective is to identify the best mapping between a new piece of data and a group of categories inside a particular problem dormain [35]. The algorithm is refered to as naive due to the fact that it acts on the assumption that the characteristics are useless [14]. The fomular of Naive bayes is as follows:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)+P(A)}{P(B)} \tag{2.3}$$

Where by,

P(A): This is the probability of an event before the evidence is known.

P(B): This is the probality that is based on evidence.

$P\left(\frac{A}{B}\right)$: for evidence based probability the hypothesis is true.

CHAPTER 3
METHODOLOGY

3.1 Methodology

This section provides an intuitive discussions of the approach used to satisfy the objective of this thesis. As mentioned earlier, the purpose of this study is to predict the likelihood of a patient having a heart attack in the near future. This involves using a heart disease numerical dataset from the UCI Machine Learning Repository and training an ML model with it. This research starts from identifying a problem statement, collection of data and pre-processing, finally to model construction and predictive analysis. In this research five machine learning classification algorithms such as Logistic Regression, KNN, Support Vector Machine, Random Forest and Naïve Bayes are used. After identifying the problem statement, appropriate dataset was collected from the UCI Machine Learning Repository, enabling an efficient approach for solving the problem statement. The next phase is performing data pre-processing which involves cleaning the data, handling the missing values if any, outlier detection and removal, and finally removing the duplicated values if any. It is essential to note that during data pre-processing, exploratory data analysis was also done as it helps to identify hidden insights of the data that may be very difficult to see with the naked eye. For instance, in order to detect outliers it is important to conduct exploratory data analysis by using box plots which show if a data contains outliers or not. In this phase, categorical data was converted to numerical data while future scaling was also done using standard scaler. After data pre-processing, feature selection was performed, which involved looking at the correlation between independent values and dependent values. The dataset was split into training set of 70% and 30% of testing set with a random state of five. Hyperparameter tuning was performed, which aimed to increase the accuracy of the models. Finally, the model with the highest test accuracy, precision, recall, f1-score and ROC value was selected. Figure 3.1 depicts a high level representation of the proposed methodology of the research.

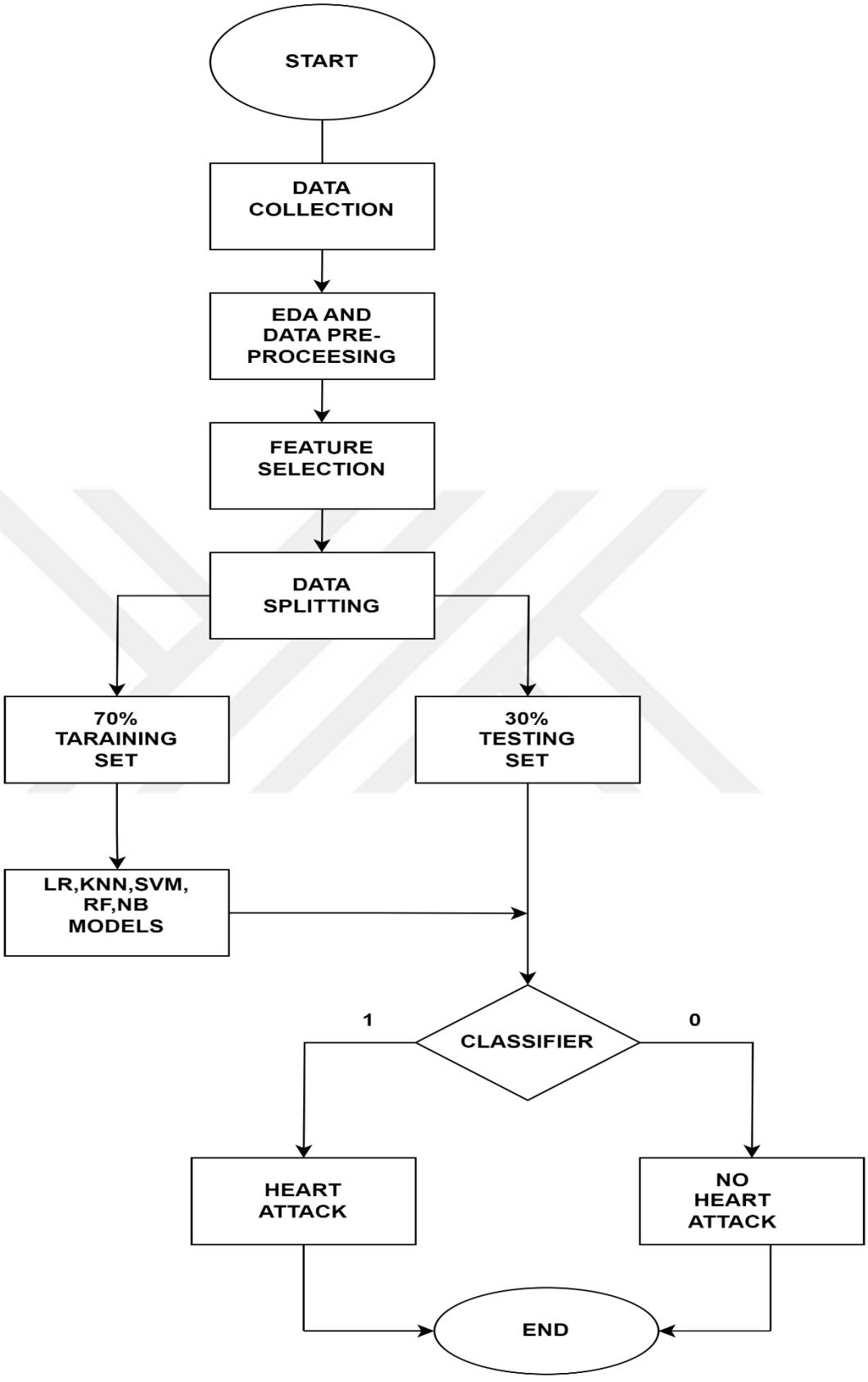

Figure 3.1. Flow chart of the proposed model.

3.2 Data Collection

In this study, an appropriate dataset was collected from the UCI Machine Learning Repository [36]; as a requirement to address our problem statement. The dataset is made up of 14 attributes with 303 patients record. This involves 207 (68.3%) males and 96 (31.7%) females. The number of patients deemed to have heart attack are 165 (54.5%) while those without were 138 (45.5%). Table 3.1 shows all the features in details.

Table 3.1. Description of UCI dataset in details.

No	Feature name	Type	Description
1	age	Numerical	Patient's age.
2	sex	Categorical	Gender of patients. Female = 0, Male = 1.
3	cp	Categorical	Chest pain type of patients. Typical angina = 1, Atypical angina = 2, Non-anginal pain = 3, Asymptomatic = 4.
4	trtbps	Numerical	Resting blood pressure (mm Hg).
5	chol	Numerical	Cholesterol (mg/dl).
6	fbs	Categorical	Fasting blood sugar (mg/dl). Fbs > 120 (mg/dl) =1, fbs < 120(mg/dl) =0.
7	restecg	Categorical	Resting electrocardiographic result. Normal = 0, having ST wave abnormality = 1, showing probable or definite left ventricular hypertrophy by Estes criteria = 2.
8	thalachh	Numerical	Maximum heart rate achieve.
9	exng	Categorical	Exercised induced angina. Yes = 1, No = 0.
10	oldpeak	Numerical	ST depression induced by exercised relative to rest.
11	slp	Categorical	Slope of the peak exercise ST segment. Unsloping = 1, Flat = 2, Downsloping = 3.
12	caa	Categorical	Number of major vessels (0-3).
13	thall	Categorical	Thallanium stress test. Normal = 1, Fixed defect = 2, Reversible defect = 3.
14	output	Categorical	Diagnosis of Heart attack. Heart attack = 1, No heart attack = 0.

The output feature of the dataset is known as the dependent feature as it is dependent on other features for its outcome to be known. The rest of the features are known as independent variable and all of them put together produce the results of the dependent value.

3.3 Exploratory Data Analysis

Exploratory data analysis (EDA) uses cutting-edge methods to examine data in order to reveal hidden structure, improve insight into a dataset, find anomalies, and create sparse models to test underlying hypothesis [37]. It can be used with the objective of extracting relevant and actionable information from the data [38]. EDA is divided into graphical and non-graphical, as well as univariate and multivariate categories. Univariate data only take into account one data column at a time, but multivariate methods analyze data using more than two factors [37]. Both univariate and multivariate analysis has been conducted so as to get hidden information in order to be able to make conclusions from the dataset. The EDA shows that there is no missing values in the dataset and all its variables are numerical, which are either integer or floating point values. Additionally, it revealed that, the dataset consists of 303 rows and 14 columns. The variables with high unique values are considered to be numerical variables while the ones with few unique values are termed as categorical values. Univariate and multivariate analysis has been used to analyze all the features of the dataset. The results of the exploratory data analysis will be discussed in the results section in detail.

3.4 Data Pre-Processing

Real world dataset are generally unclean due to the influence of natural disturbances and variation in the physical environment. This makes the dataset inconsistent for used for practical applications. This necessitates the use of proper data pre-processing methods to ensure a representative and high-quality dataset, which helps to increase the accuracy and effectiveness of the subsequent mining process [39]. It is important in machine learning to clean the data. Some of the data pre-processing techniques involve filling in missing values, outlier handling, categorical hot encoding, data normalization, feature selection and splitting of data. Since there is no missing values in the dataset used so there is no need to fill in missing values hence this step is skipped. The other steps will be explained below.

3.4.1 Outlier Detection and Handling

Outliers are observations that differ from the majority observations enough to warrant consideration that they are produced by a separate generating process [40]. They can be lower than 5%. Handling outliers helps to increase accuracies of machine learning

models. The boxplot was used in order to detect the outliers and after detection the interquartile range (IQR) method was used to remove the outliers. It is also the most commonly used method by researchers. For outlier detection the data is split into three quartiles where by Q_1 is referred to first quartile, Q_2 as second quartile and Q_3 as third quartile and can be formulated by IQR= $Q_3 - Q_1$ then the lower boundary L_B and upper boundary U_B is calculated using the following equations [41].

$$L_B = Q_1 - 1.5 * IQR \tag{3.1}$$

$$U_B = Q_3 + 1.5 * IQR \tag{3.2}$$

Any value that is lower than L_B and greater than U_B is considered to be an outlier. After detecting the outliers, those that are lower than L_B are equated as L_B and those that are greater than U_B are equated as U_B.

3.4.2 One Hot Encoding

One hot encoding is a technique that converts categorical values to numerical values by splitting the columns into multiple columns. It helps machine learning model to make a better and accurate prediction. The categorical variables of the dataset are converted to dummy variables using one hot encoding so as to increase the efficiency and accuracy of the prediction model.

3.4.3 Feature Selection

For classification or regression problems, supervised feature selection is typically used. Its objective is to choose a collection of features that will be able to separate samples from different classes or regression targets [42]. Feature relevance is known by observing the correlation between the independent features and the dependent features. The independent features that are deemed to have low correlation with the dependent features will be dropped while those with high correlation will be selected.

3.4.4 Data Splitting

The dataset as stated earlier consist of 303 rows and 14 columns. These were split into training and testing sets. The training set was 70% while the testing set was 30% and a random state of five was chosen.

3.4.5 Data Normalization

The reduction of features into a similar range, known as data normalization, is a crucial pre-processing step that prevents larger numeric feature values from dominating smaller numeric feature values. The primary objective is to reduce the bias of characteristics whose numerical contribution is larger in discriminating pattern classes [43]. The standard scaler method was used.

3.5 Modelling

This research validate and evaluate a range of state-of-the-art model, among which, five models were selected. These models includes, LR, KNN, SVM, RF and NB. They were selected based on their initial performance for carrying out a similar task.

3.6 Evaluation Metrics

In this research, the data is divided into training phase and testing phase. After training, the models are evaluated on the testing data which is not part of the training data so as to be able to evaluate the model's performance on data that it is not trained on and the metrics calculated.

3.6.1 Confusion Matrix

Confusion matrix measures the performance of an algorithm and it also helps to observe the factors that are affecting the performance of a machine learning algorithm. It is used for classification problems. The heart attack prediction dataset is a binary classification problem because it is either 0 or 1 where by 0 means absence of heart attack and 1 means presence of heart attack. For confusion matrix there are four entries that are relevant in determining the metrics for a binary classifier [44].

$$CM = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{3.3}$$

The confusion matrix in Figure 3.2 is a binary classifier. In the actual class the values are marked True (1) and False (0), while the predicted value are marked Positive (1) and Negative (0).

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.2. Confusion matrix for binary classification [45].

True positive (TP): This is when the outcome is predicted to be positive and the outcome becomes positive [14], [44], [45].

False positive (FP): This is when the outcome predicted is positive while the outcome becomes negative [14], [44], [45]. It is known as Type 1 Error [45].

False negative (FN): This is when the outcome predicted is negative while the outcome becomes positive [14], [44], [45]. This is known as Type 2 Error [45].

True negative (TN): This is when the outcome predicted is negative and the outcome becomes negative [14], [44], [45].

3.6.2 Accuracy

Accuracy is the ratio of the samples that were correctly classified to the total number of all the samples in the evaluation dataset. The accuracy is between 0 to 1 where 0 means predicting none of the samples correctly and 1 means predicting all the samples correctly.

$$Accuracy = \left(\frac{TN+TP}{TP+TN+FN+FP}\right) \tag{3.4}$$

3.6.3 Recall

Recall is also known as sensitivity or true positive rate. Recall is the ratio of number of accurate positive prediction to that of total number of positive class. The highest value of recall is 1 which means predicting the positive class correctly and lowest value is 0 which is predicting all the positive samples incorrectly.

$$Recall = \left(\frac{TP}{TP+FN}\right) \tag{3.5}$$

3.6.4 Precision

Precision is the ratio of the correctly predicted positive numbers to that of the total number of the positive prediction. The precision is between 0 to 1 were 0 means predicting all positive numbers wrongly and 1 means vice-versa.

$$Precision = \left(\frac{TP}{TP+FP}\right) \tag{3.6}$$

3.6.5 F1-Score

The harmonic mean of precision and recall is the f1-score. The f1-score is between 0 to 1. The value 1 represent maximum precision and recall value, while 0 value represent minimum precision and recall.

$$F1 - score = 2\left(\frac{Precision*Recall}{Precision+Recall}\right) \tag{3.7}$$

3.6.6 Roc Curve

Roc curve which means receiver operator characteristics (ROC) curves are used in various applications to illustrate how well a prediction matches the actual result. Threshold agnosticism is one of the major benefits of ROC analysis; performance of a predictor is calculated without a specified threshold and also provides a criteria to select an ideal threshold based on a given cost function or objective. A typical ROC analysis demonstrates how, at various thresholds, the sensitivity (true positive rate) and specificity (true negative rate or false positive rate) change [46].

3.7 Cross Validation

A resampling technique that uses several data subsets to evaluate and train a model across a number of different iteration is called cross validation. One of the most popular ways to test a prediction model's generalizability and avoid overfitting is through cross validation, which resamples data in a variety of ways [47]. It is primarily utilized for predicting outcomes and attempting to determine how well a predictive model will work in real-world situations. A methodological error is learning a prediction function's parameters and then testing them on the same set of data. In this research the k-fold cross validation was used and since the dataset used is small it is advisable to use to k-fold cross validation and it was applied in hyperparameter tuning. The k-fold is divided into 10 folds and in each of the 10 folds, 1 fold is used for

validation data and this continues until each of the 10 folds are used for validation purposes. Figure 3.3 shows a 10-fold cross validation score. The dataset is randomly divided into 10 different subset whereby each was containing approximately 10% of the dataset. After the model was trained on a training set it was finally applied to the validation dataset.

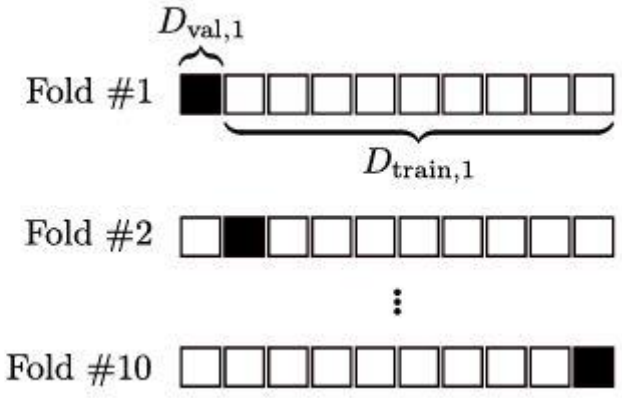

Figure 3.3. 10-fold cross validation score [47].

3.8 Hyperparameter Tuning

In order to increase the accuracy of machine learning models hyperparameter tuning is used. When training a machine learning model with scikit-learn default parameters are given and only those parameters are trained during the training process. Different parameters can be used to improve accuracy of the models. It is time consuming and may take hours, days or even weeks to get the right parameters as it is based on trials and errors. Some of the techniques of hyperparameter tuning are baby sitting (known as trial and error) , random search cross validation , grid search cross validation, gradient based optimization etc [48]. The techniques used in this research is grid search cross validation which exhaustively search a combination of all possible parameters and also trial and error known as baby sitting. It is time consuming and takes longer to train.

3.9 Model Selection

After training all the models and evaluating them, their results were looked at based on all the evaluation metrics as it was important to judge a model not only on accuracy but precision, recall, f1-score and ROC values. The second step carried out was to tune all the models using hyperparameter tuning to select the best model. This means models were trained without hyperparameter tuning first and evaluated, then trained again for the second time by trying to tune the models and the overall best model was selected.

CHAPTER 4
RESULTS

4.1 Results

In this chapter, the results of all the experimental findings and the performance of each model are presented and discussed. EDA findings will also be discussed. This is important as it will give further knowledge about the dataset that cannot be physically seen by humans. The results from the feature selection and outlier detection and handling stages are also presented in this section. The performance metrics of all the models before and after hyperperameter tuning are presented also in this section. Finally, the best model is selected and compared with results of existing studies as shown in Table 4.25.

4.2 Results of the Exploratory Data Analysis

Exploratory data analysis is vital in machine learning as it helps to understand the relationship between all existing features of a dataset that cannot be seen and helps to make decision based on the relationship of the features of these features. This can help in knowing if there is a missing value, presence of outliers, categorical values, irrelevant features and also if the values among features are large or small. Table 4.1 depicts the mean value, minimum values and maximum values of the dataset to two decimal places. The shape of the dataset is 303 rows and 14 columns**.** The exploratory data analysis is divided into univariate and multivariate data analysis**.**

4.2.1 Univariate Data Analysis

In univariate data analysis, features were examined individually. Features with a high number of unique values were categorized as numerical features, while those with fewer unique values were considered categorical. The dataset contain both numerical (age, trtbps, chol, thalachh and oldpeak) and categorical features (sex, cp, fbs, restecg, exng, slp, caa, thall). Figure 4.1 indicates a univariate distribution plots of the numerical features in the dataset. The age variable clearly shows that the majority of

the patients are in the 50 to 60 age range. It can also be said that few patients are of the age 30 below and 70 above.

Table 4.1. The mean, minimum and maximum values of the dataset.

Feature name	Mean value	Minimum value	Maximum value
age	54.4	29.0	77.0
sex	0.68	0.0	1.0
cp	0.97	0.0	3.0
trtbps	131.62	94.0	200.0
chol	246.26	126.0	564.0
fbs	0.15	0.0	1.0
restecg	0.53	0.0	2.0
thalachh	149.65	71.0	202.0
exng	0.33	0.0	1.0
oldpeak	1.04	0	6.2
slp	1.40	0	2.0
caa	0.73	0.0	4.0
thall	2.31	0.0	3.0
output	0.53	0	1.0

The trtbps (resting blood sugar of patients) variable indicates that most of the patients resting blood pressure is between 120 and 140. It could also be said that values that are more than 180 may be regarded as outliers. The chol (cholesterol) variable of most the patients are around 200 to 270. The thalachh (maximum heart rate achieve) variable of most patients are 140 to 160. Values less than 80 maybe regarded as outliers. Majority of the patients in the oldpeak variable is from 0 to 1 and values after 3 may be regarded as outliers. Figure 4.2 to 4.3 indicates a univariate distribution of categorical values using pie charts. The sex variable showing that 68.3% of patients consists of males and 31.7% is females. Therefore, the males are more than twice than the female patients. About half of the patients 47.2% in cp (chest pain type) are asymptomatic that is pain without any form of symptoms. 28.7% of the patients have atypical angina which includes shortness of breath. Finally, 16.5% of the patients have typical angina which is pain during physical activities and the rest of the patients 7.6% has non angina pain that does not results to heart attack or heart diseases. In the fbs (fasting blood sugar) 85.1% of the patients has an fbs of more than 120 mg/dl and the rest 14.9% has an fbs of less than 120mg/dl. Restecg (resting electrocardiographic) features shows that 50.2% of patients have abnormality while 48.5% are normal. Restecg type 2 is negligible since it is too small (1.3%) and it is not important. On the

basis of exng (exercised induced angina) 67.3% of patients does not experience pain caused by angina while the rest do. Pain caused by exercised induced angina (67.3%) is twice more than pain not caused by exercised induced angina (32.7%). The slp (slope) variable indicates that at least 46.9% ST wavelength of patients results is flat while 42.6% indicates that the ST wavelength is sloped upwards. The remaining 7% is downward slope. More than half of the patients 57.8% in caa (numbers of major vessels) variable is 0 which means that there are no big vessels that is colored by fluoroscopy. 21.5% of the patients results in caa is 1 which also cannot be observed by fluoroscopy technique. Most of the patients have occluded veins. Therefore, the fluoroscopy technique cannot be used to monitor big vessels. According to the thall (thallanium), 0 is not significant hence it is null. Most of the patients have a fixed defect of 2 which is 54.8%, others have a reversible defect of 3 which is 38.6% while the rest are normal which consists of 5.9% of patients. Finally the target variable shows that 54.5% of patients have heart attack and 45.5% of patients have no heart attack.

4.2.2 Multivariate Data Analysis

In the multivariate data analysis each feature in the dataset was analyzed with the output variable to see if there are any possible relationship among them. The numerical variables were compared with the output first using the KDE plot known as kernel density estimate. Figure 4.5 and 4.6 show the multivariate analysis of numerical data compared with output. The comparison of age and output variable indicated that from age 55 above chances of having heart attack reduces which is in contrast with popular believe. According to trtbps and output variable it is difficult to interpret chances of having heart attack due to resting blood sugar as the graphs are basically identical. In the cholesterol and target variable, a cholesterol value between 200 to 250 is not good for patients. The thalachh and output value relationship shows that the higher the thalachh value the greater the chances of having heart attack as there is an increase of up to the 150 value and then it started going down. This shows that there is correlation between thalachh and output variable. Finally, in the oldpeak and output value, from value 0 to about value 1.3 their is a huge increase in chances of having heart attack. The exploratory data analysis of categorical values was perfromed using the bar chart. This was done to check the relationship between output and categorical values.

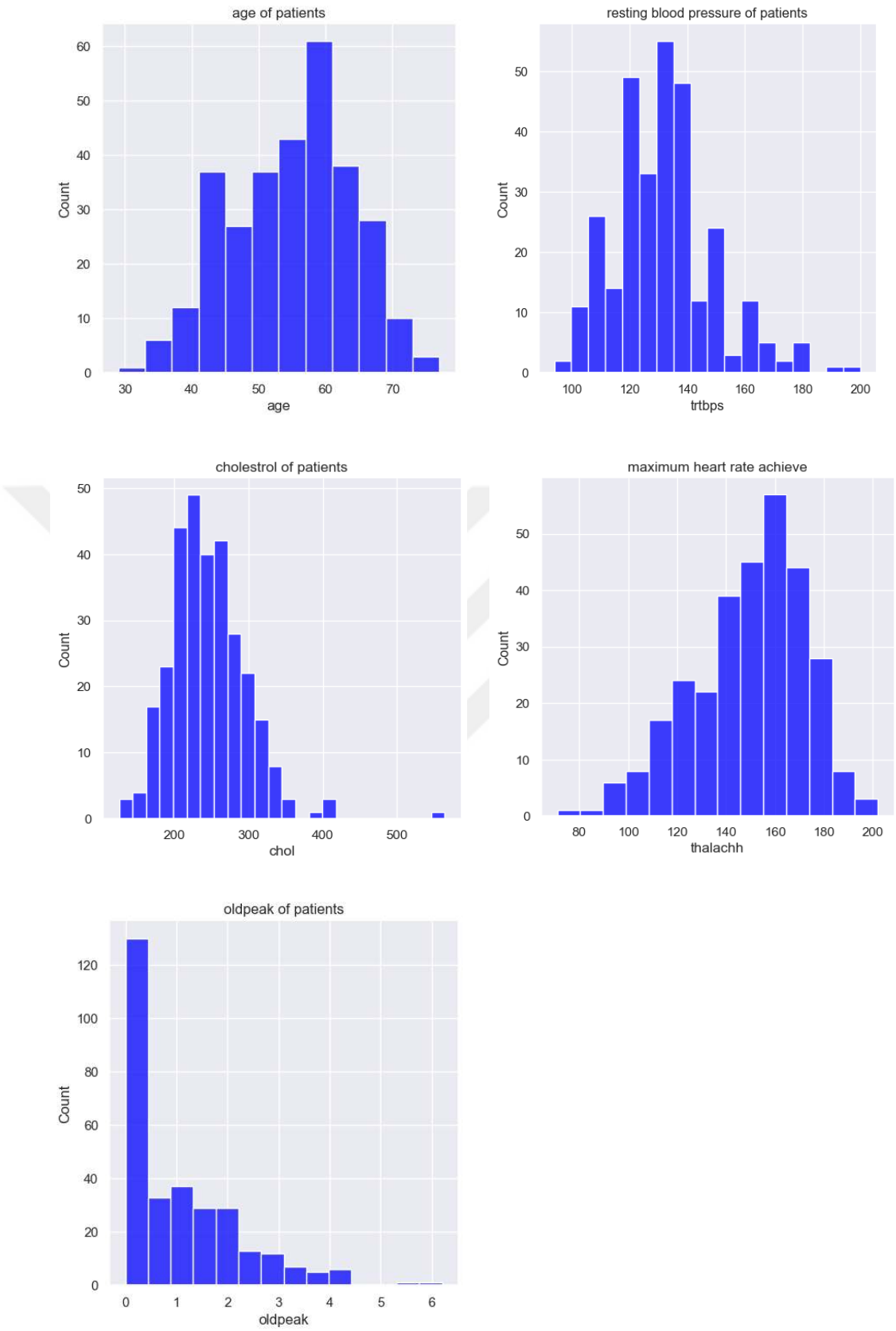

Figure 4.1. Univariate analysis of numerical values.

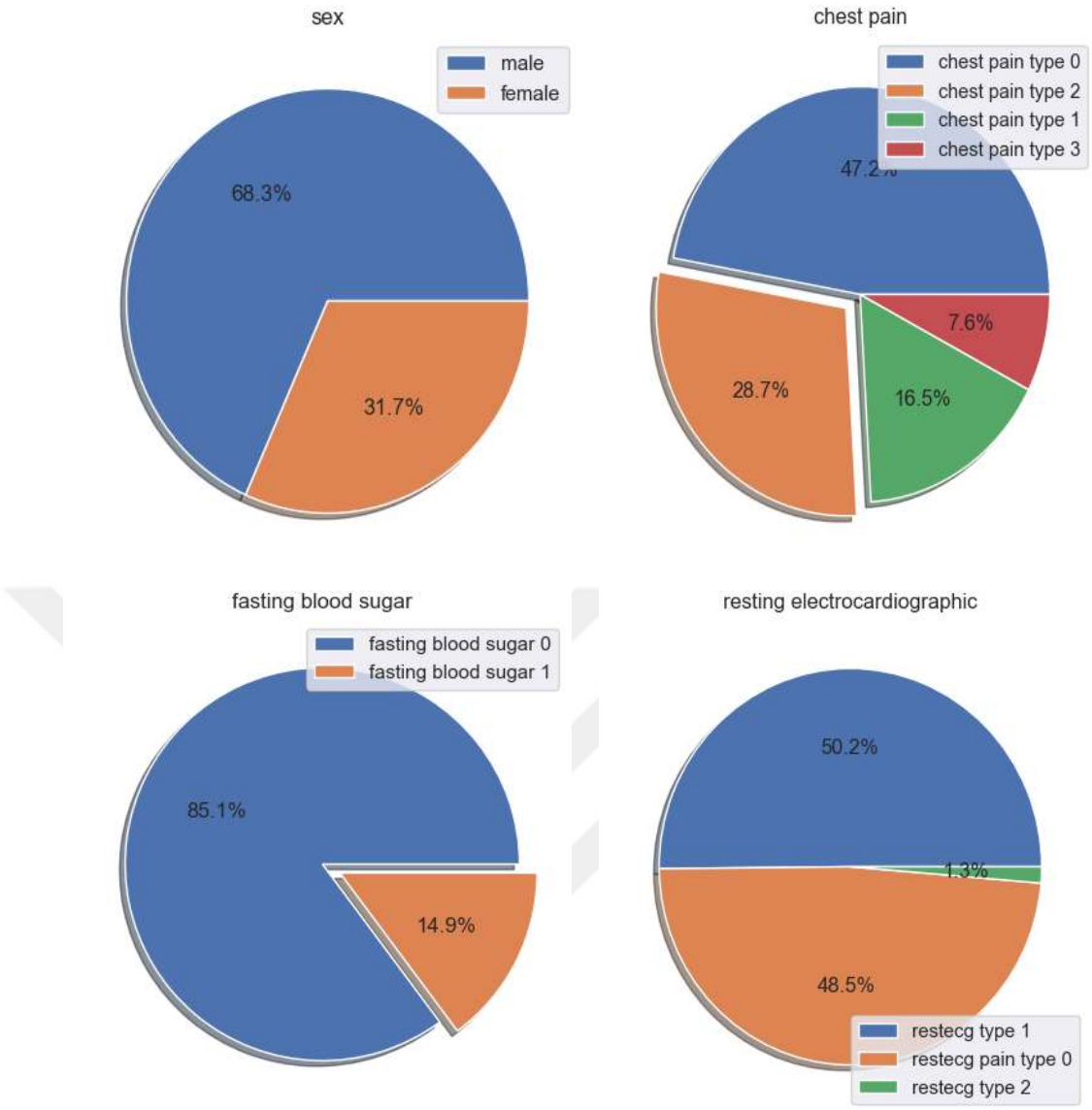

Figure 4.2. Univariate analysis of categorical values.

In the sex and output variables, results showed that women were twice more likely to have heart attack than not. Men have less chances of heart attack therefore, it can be concluded that women have higher chances of having heart attack than men. Patients with asymptomatic pains have less chances of having heart attacks while the other three types of pain have more chances of contracting the disease. Fasting blood sugar that is lower than 120 mg/dl have a higher chance of having heart attack while fasting blood sugar that is more than 120 mg/dl have equal chances of heart attack in fbs and output variable.

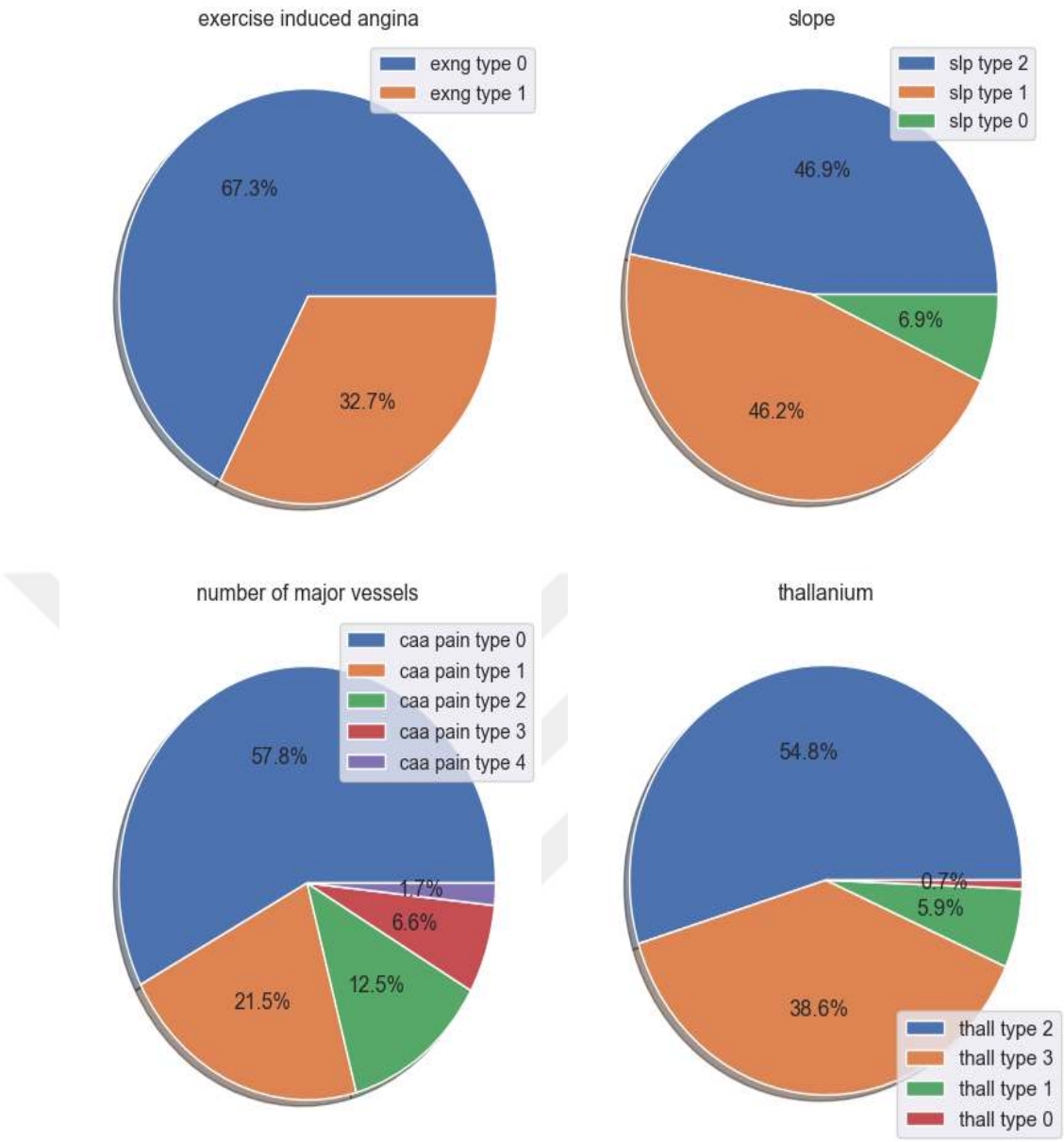

Figure 4.3. Univariate analysis of categorical value continuation.

Patients with restecg 1, indicating ST wave abnormality, are at a higher risk of experiencing heart attacks, whereas those with the other two values have a lower likelihood of suffering from a heart attack. The variables exng and the output variable indicate that chest pain caused by exercise does not lead to heart attacks; conversely, patients with non-exercise angina have a higher probability of experiencing heart attacks. It can be concluded that exercise due to pain does not cause heart attacks. Patients with downslope in the slp and output variable have twice more chances than the other values of having heart attacks.

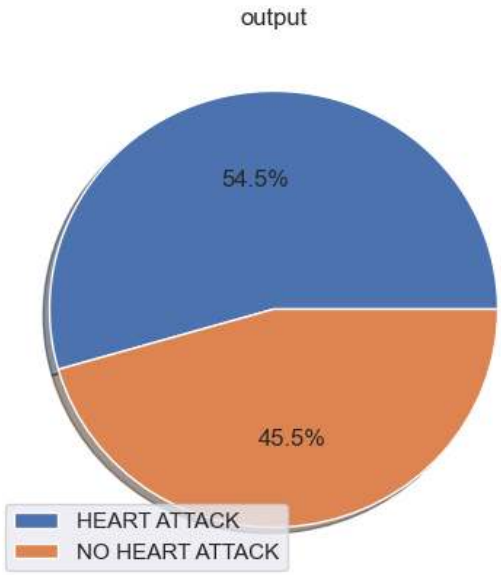

Figure 4.4. Univariate analysis of output value.

In ca and output value the patients with a value of 0 have greater chances of heart attack than the other 3 values which are twice less likely of having it. Patients with fixed defect that is those with observation value 2 are 3 times more likely of having the disease than the other patients. Both the univariate and multivariate data analysis have given a thorough analysis of both individual features and multiple features relationships. It is very import to do this as it gives a greater understand of the dataset which could not have been detected by the naked eye.

4.2.3 Outliers Result

Outliers are values that are actually very far from the normal distribution. The presence of outliers can hinder machine learning performance hence the need to view them and eliminate them. The boxplot was used to view whether outliers were present or not. Outliers may be true values or values that may have been entered wrongly. In this context, only the numerical values were considered when checking for presence of outliers due to the fact that it could be very difficult to get data on some of the categories hence categorical values were not considered in this study. The boxplot in Figure 4.9 shows that there are presence of outliers in trtbps, chol, oldpeak and thalachh variables. Since outliers are not needed and also to avoid deleting all the

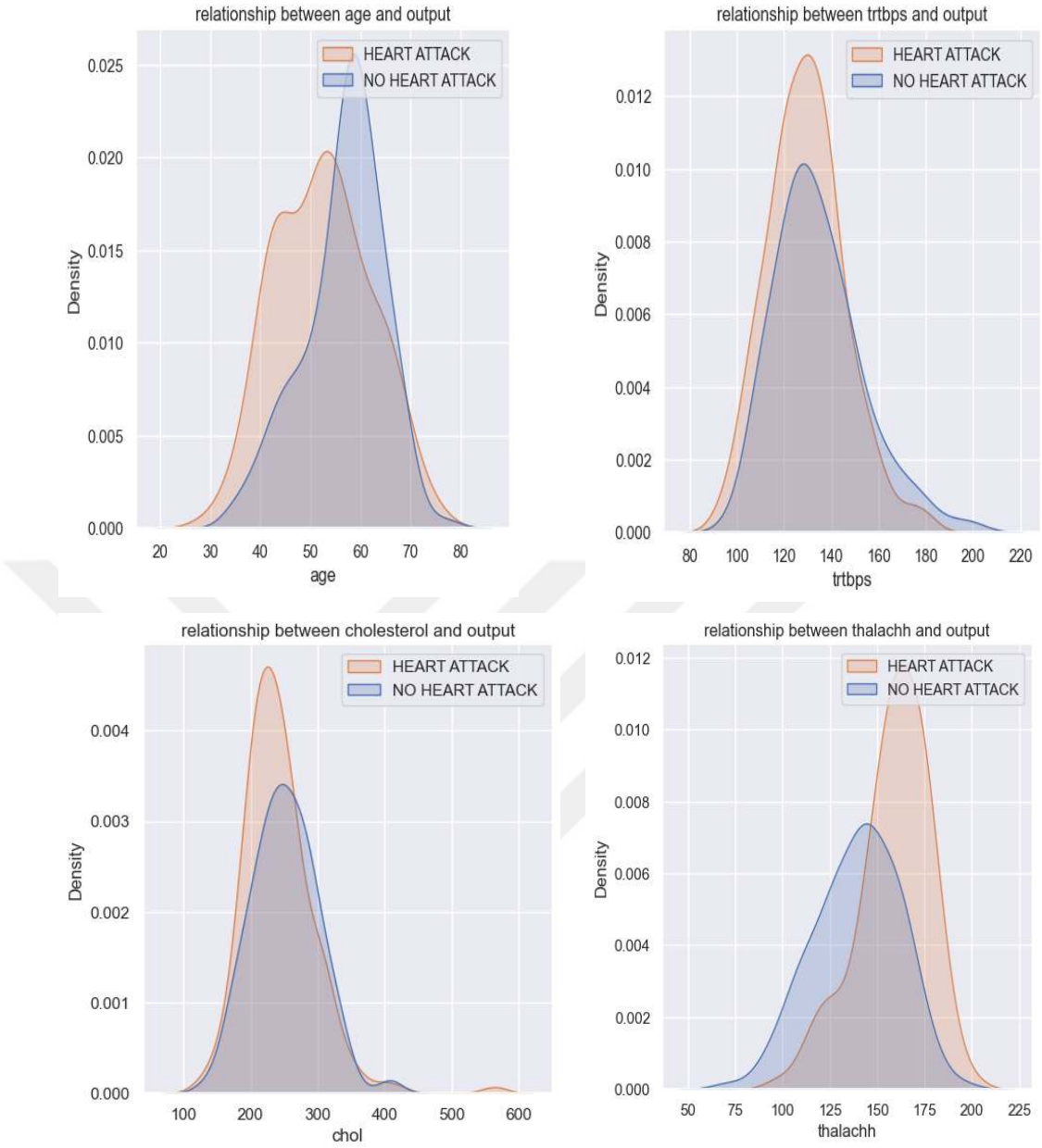

Figure 4.5. Multivariate analysis of numerical values.

features that contain outliers as those features may have significant importance, the IQR method is used and features that have outliers that are below the lower boundary L_B are made to be equal to the lower boundary and ones that are more than the upper boundary U_B were equated as upper boundary. This avoid the need of removing all the features that may be significant in the data. Figure 4.10 displays absence of outliers after IQR method has been done to remove the outliers.

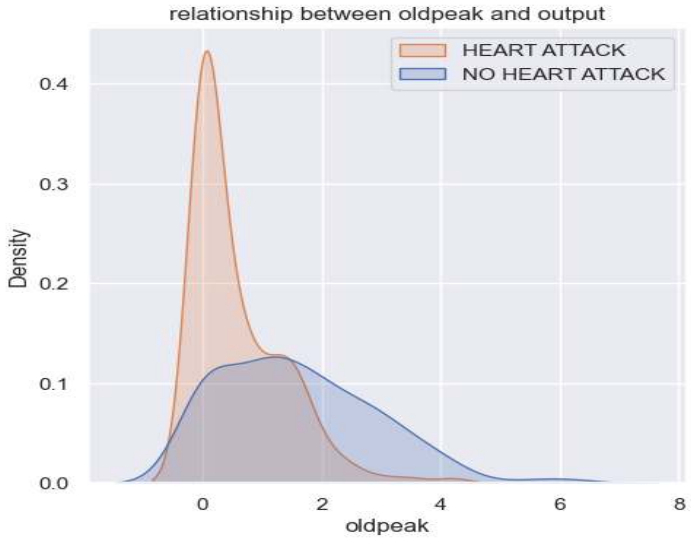

Figure 4.6. Multivariate analysis of numerical value continuation.

4.3 Results of Feature Selection

This is important as it gives the features that are less significant in the data. This helps in given better model performance. In this study, correlation between independent features and dependent features were looked at and those that were deemed to have higher correlation if any would be dropped. A threshold of 0.75 was set for the independent values and any feature correlation judge to be more than that would be dropped. However, no values were judge to be more than the threshold set and therefore, none of the independent values was dropped. Finally, for the correlation of independent and dependent values, three features were found to be lowly correlated with the output and were dropped. These were chol, fbs and restecg. Figure 4.11 shows the correlation matrix of all the features.

4.4 Results of Hyperparameter Tuning

Hyperparameter tuning is vital in increasing model performance. Five machine learning classification models were made once all the pre-processing approaches were completed. The machine learning classification models were Logistic Regression, KNN, SVM, Random Forest and Naive Bayes. Figure 3.1 shows the flow chart of the proposed study. The five machine learning algorithms were done and both grid search cross validation and trail and error method was use to tune the machine learning models to increase accuracy and performance. Table 4.2 depicts the parameters used to tune the machine learning models.

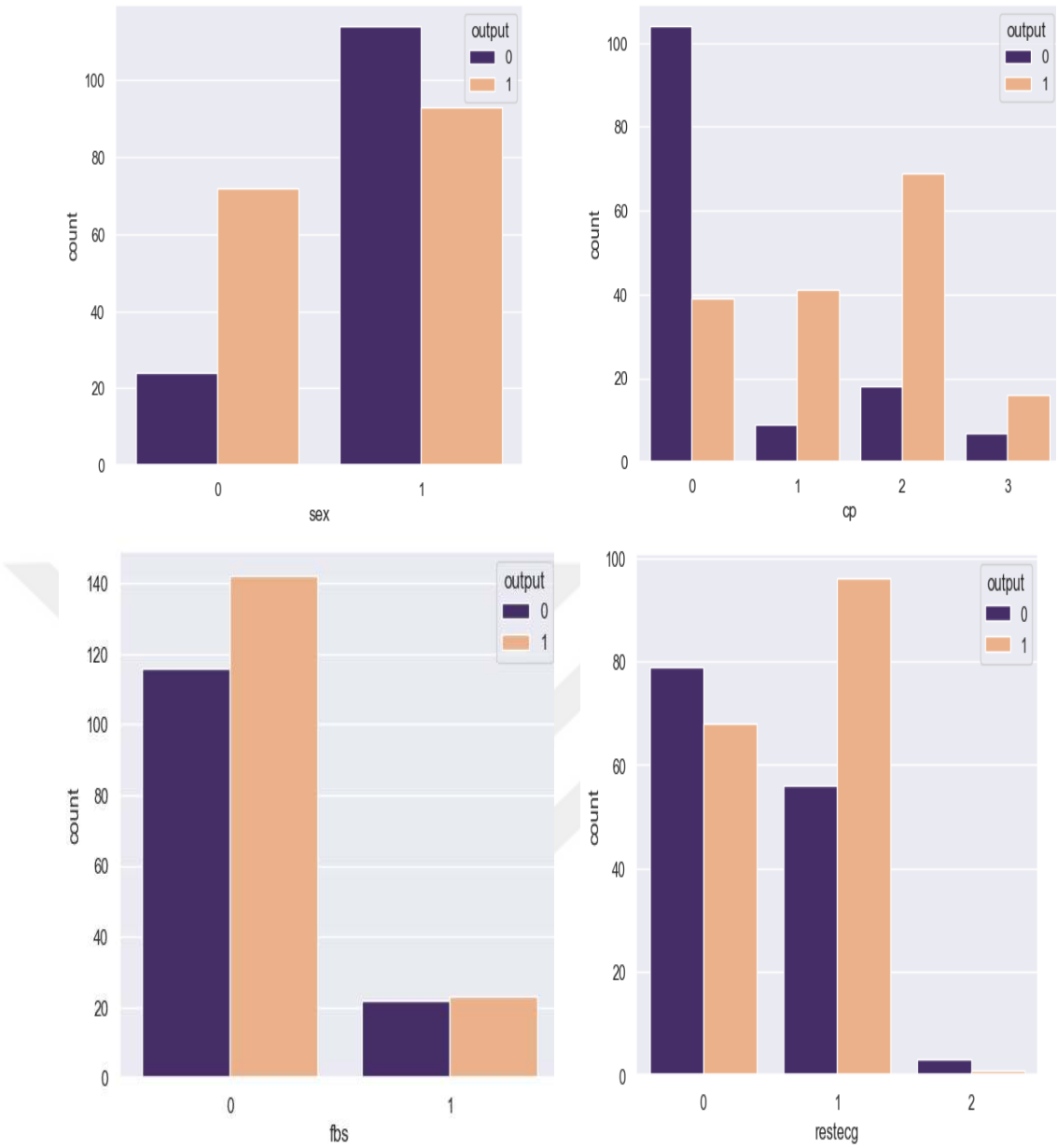

Figure 4.7. Multivariate analysis of categorical values.

4.5 Performance Metrics Analysis

One of the most frequent misconceptions in this field is the notion that every dataset, regardless of its form, can be assessed using the same evaluation criteria for machine learning models. The most common method for assessing machine learning models is accuracy. The performance of machine learning models should not be evaluated primarily by accuracy. It is important to consider accuracy, recall, precision, f1-score,

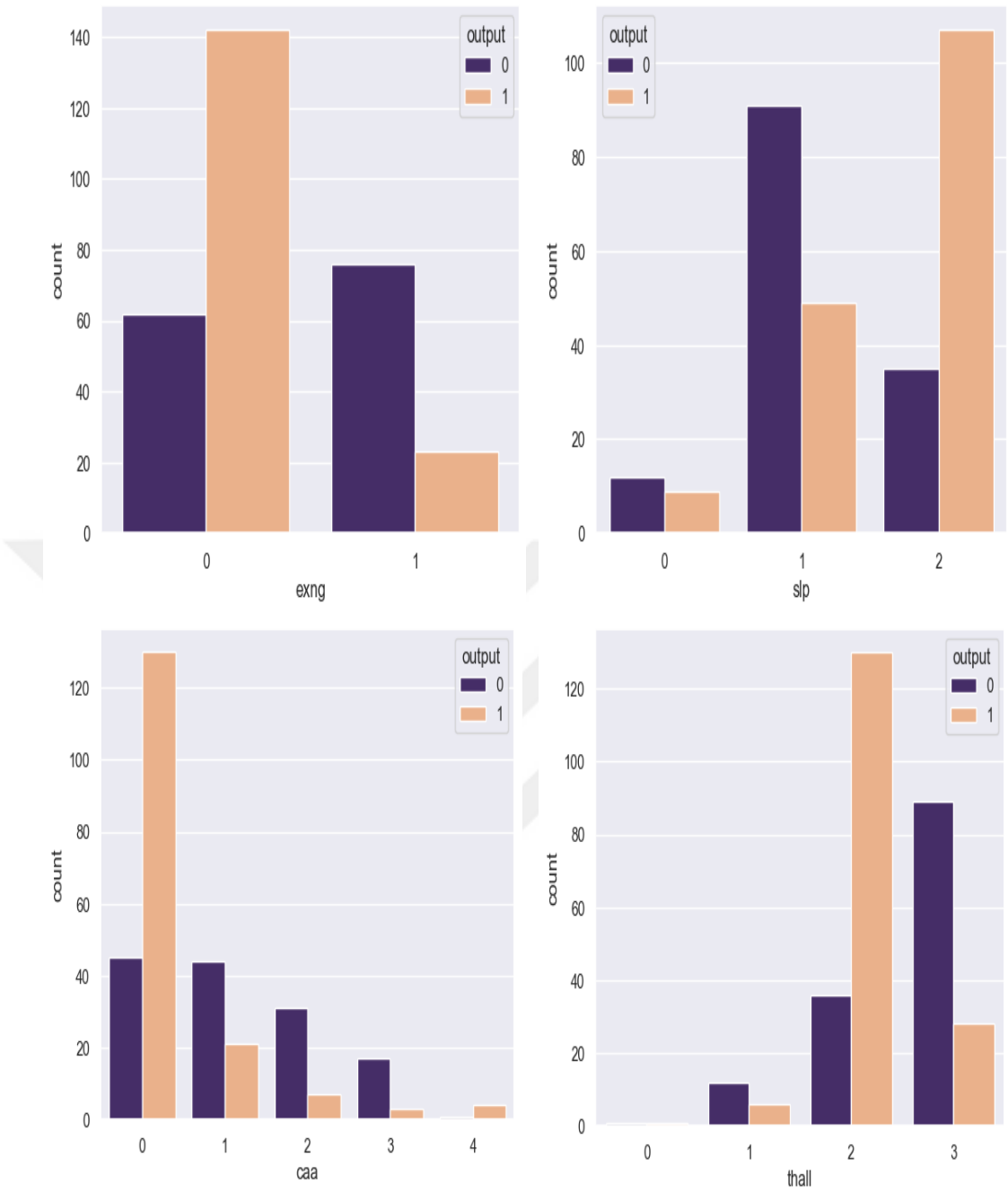

Figure 4.8. Multivariate analysis of categorical values continuation.

and ROC curve at the same time to evaluate the performance of machine learning models in classification task since this provides a fair picture of the model performance. The performance metrics is divided into two, the first is done before hyperparameter tuning while the second is done after hyperparmeter tuning. It is important to know if machine learning performance can increase after tuning.

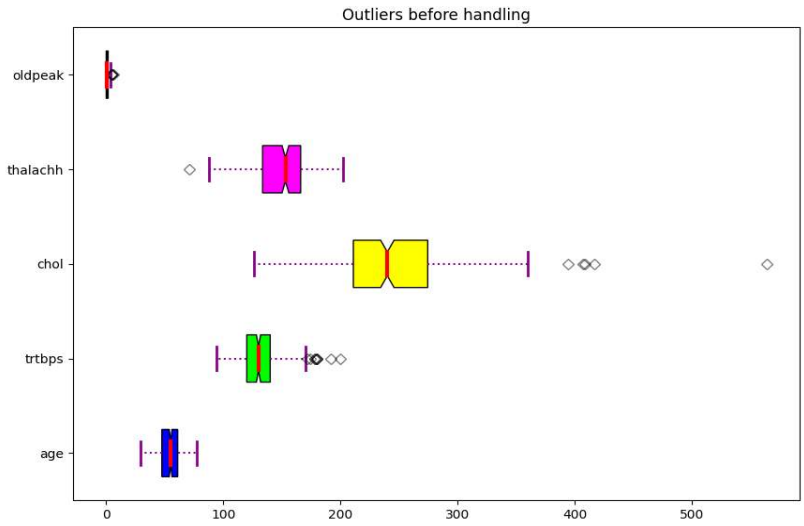

Figure 4.9. Outliers present in the dataset.

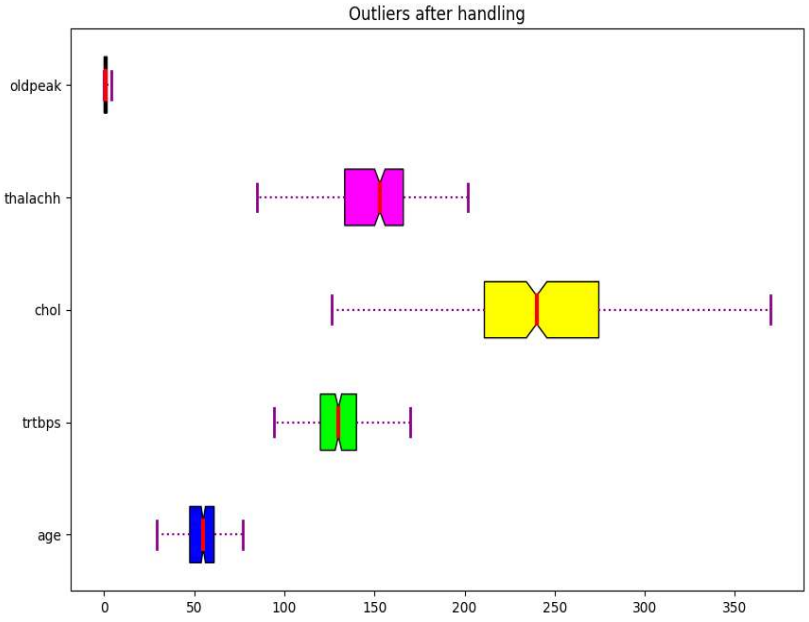

Figure 4.10. Absence of outliers after handling.

4.5.1 Models Performance before Tuning

Only the metrics that were done after machine learning classification techniques were made and did not involve tuning. The confusion matrix, accuracy, precision, recall, f1-score, and ROC curve were all used to evaluate the performance of the machine learning algorithms in this study. Table 4.3 to 4.7 lists the confusion matrix of all 5 machine learning classifiers trained. As shown from Table 4.3 to 4.7, the Logistic Regression produces the best result in the confusion matrix and out of 91 values 8 are missclassified. The second best result in the confusion matrix was produced by

Support Vector Machine and from 91 values 10 were wrongly identified.

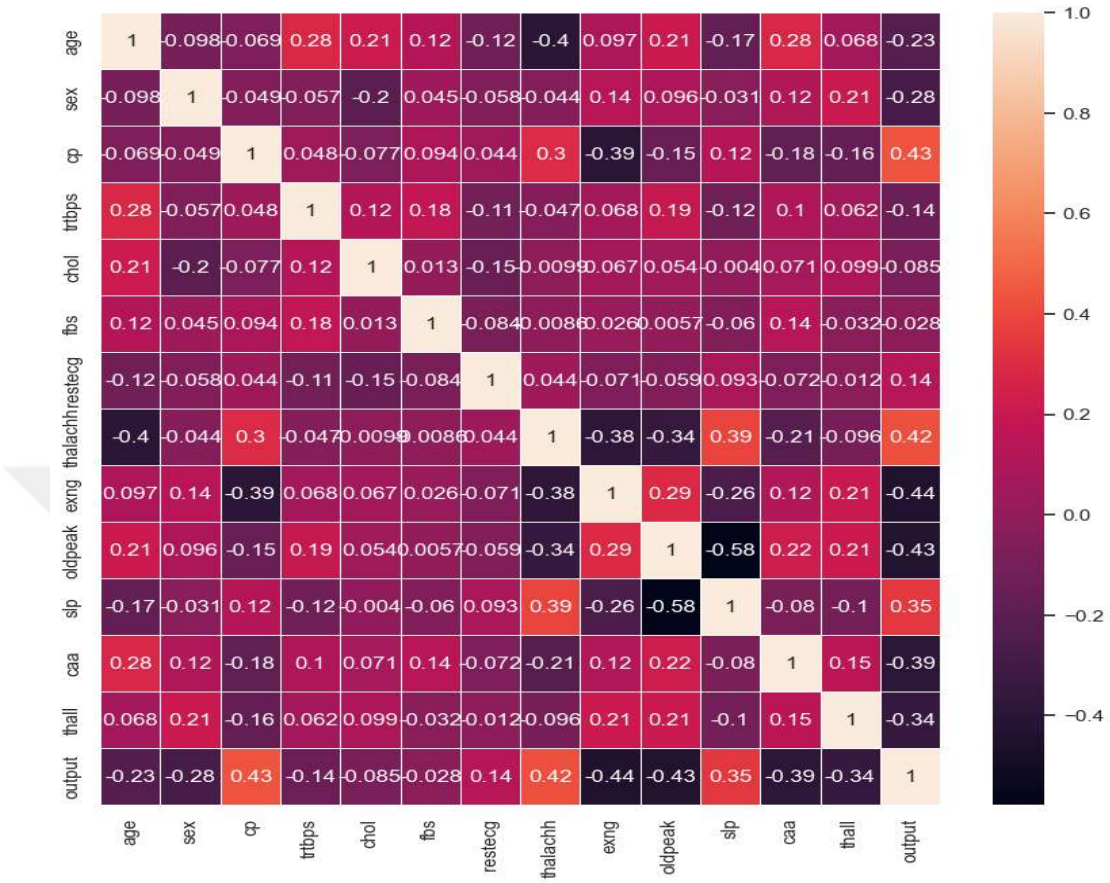

Figure 4.11. Correlation matrix of the features in the dataset.

Table 4.2. Parameters that were tuned.

No	Classification Models	Tuned Parameters
1	Logistic Regression	C = 7.7, penalty = l2, max_iter = 2000
2	K-Nearest Neighbors	n_neighbors = 12
3	Support Vector Machine	C = 1, gamma = scale, kernel = linear
4	Random Forest	n_estimators =110, max_features = sqrt, max_depth = None, criterion = gini, random_state = 5
5	Naïve Bayes	alpha = 2

Table 4.3. Confusion matrix of Logistic Regression.

Logistic Regression Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	40	3
	Negative (0)	5	43

Table 4.4. Confusion matrix of K-Nearest Neighbor.

K-Nearest Neighbour Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	36	7
	Negative (0)	6	42

Table 4.5. Confusion matrix of Support Vector Machine.

Support Vector Machine Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	39	4
	Negative (0)	6	42

Table 4.6. Confusion matrix of Random Forest.

Random Forest Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	38	5
	Negative (0)	6	42

Table 4.7. Confusion matrix of Naïve Bayes.

Naive Bayes Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	37	6
	Negative (0)	7	41

Random Forest produces the third best result from the confusion matrix and out of 91 values 11 were not identified correctly. Finally both K-Nearest neighbors and Naive Bayes had identical number of missclassification and out of 91 values 13 were missclassified. Table 4.8 to 4.12 lists both the positive and negative classes of precision recall, and f1-score. This is also important in deciding the best model for the data used.

Table 4.8. Logistic Regression precision, recall and F1-score.

	Precision	Recall	F1-score
0	89%	93%	91%
1	93%	90%	91%

Table 4.9. K-Nearest Neighbor Precision, Recall and F1-score.

	Precision	Recall	F1-score
0	86%	84%	85%
1	86%	88%	87%

Table 4.10. Support Vector Machine Precision, Recall and F1-score

	Precision	Recall	F1-score
0	87%	91%	89%
1	91%	88%	89%

Table 4.11. Random Forest Precision, Recall and F1-score

	Precision	Recall	F1-score
0	86%	88%	87%
1	89%	88%	88%

Table 4.12. Naïve Bayes Precision, Recall and F1-score.

	Precision	Recall	F1-score
0	84%	86%	85%
1	87%	85%	86%

The logistic regression also outperformed all the machine learning models in terms of both precision, recall and f1-score positive and negative class. Logistic Regression was followed by Support Vector Machine. The third best classifier was Random Forest. K-Nearest Neighbor was slightly better than Naive Bayes interms of averaging the positive and negative precision, recall, and f1-score. From this metrics it can be concluded that Naive Bayes is the worst performing classifier. Table 4.13 indicates the model performance on all of the metrics. This includes accuracy, ROC value, the average precision, recall, and f1-score. The average of the precision, recall and f1-score were taking by adding the positive and negative classes and divided by 2.

According to Table 4.13 the logistic regression has again outperforms all of the machine learning classifiers on every metric with an accuracy of 91% on all the metrics except mean recall which is 91.5% and hence it can be said that logistic regression is the best approach for this data without tuning the models. The worst model for this data based on these algorithms was Naive Bayes.

Table 4.13. Model performance of Machine Learning algorithms.

No	Classifiers	Accuracy	Mean precision	Mean recall	Mean f1-score	ROC curve
1	Logistic Regression	91%	91%	91.5%	91%	91%
2	K-Nearest Neighbour	86%	86%	86%	86%	86%
3	Support Vector Machine	89%	89%	89.5%	89%	89%
4	Random Forest	88%	87.5%	88%	87.5%	88%
5	Naive Bayes	86%	85.5%	85.5%	85.5%	86%

4.5.2 Models Performance after Tuning

After the machine learning models were trained and performance metrics analyzed, it was important to try to increase model performance and efficiency. Therefore, hyperparameter tuning was done to increase accuracy of the trained models. The confusion matrix, accuracy, precision, recall, f1-score and ROC curve were all taking into account. Table 4.14 to 4.18 shows the confusion matrix of the classifiers after hyperparameter tuning was applied.

Table 4.14. Confusion matrix of Logistic Regression.

Logistic Regression Classifier		Actual Class	
		Positive (1)	Absence (0)
Predicted Class	Positive (1)	41	2
	Negative (0)	4	44

Table 4.15. Confusion matrix of K-Nearest Neighbor.

K-Nearest Neighbor Classifier		Actual Class Positive (1)	Absence (0)
Predicted Class	Positive (1)	39	4
	Negative (0)	5	43

Table 4.16. Confusion matrix of Support Vector Machine.

Support Vector Machine Classifier		Actual Class Positive (1)	Absence (0)
Predicted Class	Positive (1)	39	4
	Negative (0)	3	45

Table 4.17. Confusion matrix of Random Forest.

Random Forest Classifier		Actual Class Positive (1)	Absence (0)
Predicted Class	Positive (1)	40	3
	Negative (0)	6	42

Table 4.18. Confusion matrix of Naïve Bayes.

Naive Bayes Classifier		Actual Class Positive (1)	Absence (0)
Predicted Class	Positive (1)	37	6
	Negative (0)	7	41

According to the confusion matrix tables, Logistic Regression has the best confusion marix after tuning. Also hyperparameter tuning has helped to increase the performance of Logistic Regression in terms of confusion matrix. All the other models performance have also been increased through hyperparameter tuning except the Naive Bayes. Therefore, it can be concluded that hyperparameter tuning can increase performance of machine learning models. After tuning second best performing model was Support Vector Machine, both K-Nearest Neighbors and Random Forest were equaled in terms of missclassified values. Naive Bayes again was the worst performing classifier. Table 4.19 to 4.23 outlines the positive and negative classes precision, recall and f1-score.

Table 4.19. Logistic Regression precision, recall and F1-score.

	Precision	Recall	F1-score
0	91%	95%	93%
1	96%	92%	94%

Table 4.20. K-Nearest Neighbor precision, recall and F1-score.

	Precision	Recall	F1-score
0	89%	91%	90%
1	91%	90%	91%

Table 4.21. Support Vector Machine precision, recall and F1-score.

	Precision	Recall	F1-score
0	93%	91%	92%
1	92%	94%	93%

Table 4.22. Random Forest precision, recall and F1-score.

	Precision	Recall	F1-score
0	87%	93%	90%
1	93%	88%	90%

Table 4.23. Naïve Bayes precision, recall and F1-score.

	Precision	Recall	F1-score
0	84%	86%	85%
1	87%	85%	86%

According to these performance metrics Logistic Regression yet again yielded the best results on all the metrics. Support Vector Machine was closely behind the Logisitc Regression model. K-nearest Neighbor and Random Forest were almost equal to each other except on f1-score metric were by K-Nearest Neighbor has just 0.5% more than Random Forest. Finally Naive Bayes was the least performing model for this data in this study. Table 4.24 exhibits the overall model performance on all the models after hyperparameter tuning. Therefore, it can be concluded that the best model for this data with or without hyperparameter tuning is the Logisitc Regression classifier. The proposed model in this study is Logistic regression with data pre-processing and

hyperparameter tuning (LR + DP + HPT).

4.6 Proposed Model Comparision with other Techniques

It is important to know if the accuracy of proposed model has reach an accuracy that is acceptable. For this reason the accuracy of the recommended algorithm was compared with 5 existing studies that were conducted previously. The suggested model outperformed all the studies it is compared with, as indicated in Table 4.25, which utilized the UCI Machine Learning Repository dataset.

Table 4.24. Model performance of Machine Learning algorithms.

No	Classifiers	Accuracy	Mean precision	Mean recall	Mean f1-score	ROC curve
1	Logistic Regression	93%	93.5%	93.5%	93.5%	94%
2	K-Nearest Neighbour	90%	90%	90.5%	90.5%	90%
3	Support Vector Machine	92%	92.5%	92.5%	92.5%	92%
4	Random Forest	90%	90%	90.5%	90%	90%
5	Naive Bayes	86%	85.5%	85.5%	85.5%	86%

Table 4.25. Proposed model compared with some previous studies.

Author	Proposed Technique	Accuracy
Mohan et al. [10]	A linear model with hybrid random forest (HRFLM)	84.81%
Nandal et al. [1]	XGBoost	92%
Takci [11]	SVM with linear kernel	84.81%
J. P. Li et al. [15]	FCMIM-SVM	92.37%
Samuel et al. [49]	FUZZY_AHP	91.0
Amin et al. [50]	Vote Method	84.7%
Proposed Technique	LR + DP + HPT	93%

CHAPTER 5
CONCLUSION

5.1 Conclusion

This study has proposed a Logistic Regression with data pre-processing and hyperparameter tuning (LR + DP + HPT) that can be used for detecting heart attacks using the UCI machine learning repository dataset. Five machine learning models have been trained. However, before training EDA was performed to understand the dataset. Understanding the data helps to make future decisions during data pre-processing. During this phase, it was realized that the dataset contained both numerical and categorical features, the presence of outliers, insignificant features and also there were both large and small numerical values. In data pre-processing all these problems were handled. The categorical features in the data were converted to numerical features which could help machine learning algorithms to perform better. Presence of outliers were detected using boxplot and calculated using IQR method. The study did not delete the features with outliers instead outliers that were judge to be more than lower bound were equated as lower bound and those that were more than the upper bound were also equated as the upper bound. This prevented in deleting features that may be important for this study. The insignificant features were removed after observing the correlation matrix. A threshold of 0.75 between independent variables were set and anything more than than 0.75 would be eliminated. However, none of the independent variables reached the set threshold, while the correlation of independent with output was also looked at. Features that were deemed to be lowly correlated with output value (fbs, chol, restecg) were dropped. These helps to avoid reducing performance of the models. Finally, feature scaling was done using standard scaler method to handle the difference between the big numeric features and small numeric features. This was conducted to ensure that the machine learning algorithms is not biased towards the larger numbers. Hyperparameter tuning was perform to increase the robustness and efficiency of the machine learning models. All the models performance increases in performance except Naive Bayes. According to Table 4.24, the logistic regression algorithm, when coupled

with data pre-processing and hyperparameter tuning, yielded the most favorable evaluation metrics. Therefore, it was chosen as the optimal algorithm for this dataset. When the proposed model was compared with some existing studies as indicated in Table 4.25, shows that the model has reached an acceptable accuracy and outperforms these studies with an accuracy of 93%. This study proofs that machine learning can be able to detect future heart attacks and is able to quickly predict these diseases faster than experts when fed with the right amount of data. This is both cost effective and saves a lot of time.

5.2 Future Work

It is crucial to acknowledge that the dataset used in this study falls under the category of small datasets. Hence, acquiring a larger dataset with more patients and additional features in the future can significantly enhance the research scope. Exploring other approaches like Internet of Things (IoT) and employing advanced optimization techniques could further improve accuracy. Lastly, implementing the recommended algorithm within a software system for heart attack prediction can provide valuable clinical guidance, aiding patients in taking preventive measures against heart attacks.

REFERENCES

[1] N. Nandal, L. Goel, and R. TANWAR, "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis," *F1000Res*, vol. 11, p. 1126, Sep. 2022, doi: 10.12688/f1000research.123776.1.

[2] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[3] L. Lu, M. Liu, R. R. Sun, Y. Zheng, and P. Zhang, "Myocardial Infarction: Symptoms and Treatments," *Cell Biochem Biophys*, vol. 72, no. 3, pp. 865–867, Jul. 2015, doi: 10.1007/s12013-015-0553-4.

[4] M. Jackson and B. J. Mcculloch, "Rural and Remote Health 14: 2560. (Online) 2014 Remote Health," 2014, doi: 10.3316/informit.351231526843522.

[5] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data," *Med Eng Phys*, vol. 105, Jul. 2022, doi: 10.1016/j.medengphy.2022.103825.

[6] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Cardiovascular Events Prediction using Artificial Intelligence Models and Heart Rate Variability," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 231–238. doi: 10.1016/j.procs.2022.07.030.

[7] S. K. Gupta, A. Shrivastava, S. P. Upadhyay, and P. K. Chaurasia*, "A Machine Learning Approach for Heart Attack Prediction," *Int J Eng Adv Technol*, vol. 10, no. 6, pp. 124–134, Aug. 2021, doi: 10.35940/ijeat.F3043.0810621.

[8] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," *Sci Program*, vol. 2021, 2021, doi: 10.1155/2021/6621622.

[9] M. Wang, X. Yao, and Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," *IEEE Access*, vol. 9, pp. 25394–25404, 2021, doi: 10.1109/ACCESS.2021.3057693.

[10] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[11] H. Takci, "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 1, pp. 1–10, 2018, doi: 10.3906/elk-1611-235.

[12] Z. Dai, L. Jiayi, T. Gong, and C. Wang, "Kernel-lasso feature expansion method: Boosting the prediction ability of machine learning in heart attack," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1955/1/012047.

[13] V. Ranga and D. Rohila, "Parametric analysis of heart attack prediction using machine learning techniques," *International Journal of Grid and Distributed Computing*, vol. 11, no. 4, pp. 37–48, 2018, doi: 10.14257/ijgdc.2018.11.4.04.

[14] O. S. Hasan and I. A. Saleh, "Developing Heart Attack Prediction Model Based on Hyperparameter Tuning and Machine Learning Approach," 2022.

[15] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[16] M. Ilayaraja and T. Meyyappan, "Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets," in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 586–592. doi: 10.1016/j.procs.2015.10.040.

[17] M. Raihan, S. Mondal, A. More, P. K. Boni, and M. O. F. Sagor, "Smartphone based heart attack risk prediction system with statistical analysis and data mining approaches," *Advances in Science, Technology and Engineering Systems*, vol. 2, no. 3, pp. 1815–1822, 2017, doi: 10.25046/aj0203221.

[18] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 449–457. doi: 10.1016/j.procs.2019.11.144.

[19] C. A. Alexander and L. Wang, "Big Data Analytics in Heart Attack Prediction," *Journal of Nursing & Care*, vol. 06, no. 02, 2017, doi: 10.4172/2167-1168.1000393.

[20] V. Maihami and E. Rahimi, "Designing an expert system for prediction of heart attack using fuzzy systems Factors affecting drug addiction and lack of referral to addiction centers in the city of Yasuj View project Improving the Ranking Results in Automatic Image Annotation Using Relevance Tags View project," 2016.[Online].Available: https://www.researchgate.net/publication/316961160

[21] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[22] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.

[23] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," 2020. [Online]. Available: www.elsevier.com/locate/bt

[24] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.

[25] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth, “Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning,” Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.08917

[26] B. Wu, X. Lv, A. Alghamdi, H. Abosaq, and M. Alrizq, “Advancement of management information system for discovering fraud in master card based intelligent supervised machine learning and deep learning during SARS-CoV2,” *Inf Process Manag*, vol. 60, no. 2, Mar. 2023, doi: 10.1016/j.ipm.2022.103231.

[27] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[28] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, “Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification,” in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, Nov. 2017, pp. 294–298. doi: 10.1109/ICITISEE.2017.8285514.

[29] J. Vitola, F. Pozo, D. A. Tibaduiza, and M. Anaya, “A sensor data fusion system based on k-nearest neighbor pattern classification for structural health monitoring applications,” *Sensors (Switzerland)*, vol. 17, no. 2, Feb. 2017, doi: 10.3390/s17020417.

[30] J. M. Johnson and A. Yadav, “Fault Detection and Classification Technique for HVDC Transmission Lines Using KNN,” 2018, pp. 245–253. doi: 10.1007/978-981-10-3920-1_25.

[31] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[32] N. Guenther and M. Schonlau, “Support vector machines,” 2016.

[33] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134. Elsevier Ltd, pp. 93–101, Nov. 15, 2019. doi: 10.1016/j.eswa.2019.05.028.

[34] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, IEEE, Oct. 2017, pp. 72–78. doi: 10.1109/ICISIM.2017.8122151.

[35] F. J. Yang, "An implementation of naive bayes classifier," in *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

[36] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," UCI Machine Learning Repository. Accessed: Apr. 29, 2023. [Online]. Available: https://doi.org/10.24432/C52P4X

[37] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 130–139. doi: 10.1016/j.procs.2020.06.017.

[38] Sarvam Mittal, "An Exploratory Data Analysis of COVID-19 in India," *International Journal of Engineering Research and*, vol. V9, no. 04, Apr. 2020, doi: 10.17577/IJERTV9IS040550.

[39] T. Iliou, C. N. Anagnostopoulos, M. Nerantzaki, and G. Anastassopoulos, "A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jan. 2015. doi: 10.1145/2797143.2797155.

[40] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognit*, vol. 74, pp. 406–421, Feb. 2018, doi: 10.1016/j.patcog.2017.09.037.

[41] M. S. Satu, S. T. Atik, M. A. Moni, and S. Tanjila Atik, “A Novel Hybrid Machine Learning Model To Predict Diabetes Mellitus,” 2020. [Online]. Available: https://www.researchgate.net/publication/335727823

[42] J. Li *et al.*, “Feature selection: A data perspective,” *ACM Computing Surveys*, vol. 50, no. 6. Association for Computing Machinery, Dec. 01, 2017. doi: 10.1145/3136625.

[43] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl Soft Comput*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[44] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-09954-8.

[45] Ž. Vujović, “Classification Model Evaluation Metrics,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

[46] J. Muschelli, “ROC and AUC with a Binary Predictor: a Potentially Misleading Metric,” *J Classif*, vol. 37, no. 3, pp. 696–708, Oct. 2020, doi: 10.1007/s00357-019-09345-1.

[47] D. Berrar, “Cross-validation,” in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[48] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: Theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020, doi: 10.1016/j.neucom.2020.07.061.

[49] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” *Expert Syst Appl*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.

[50] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.

CURRICULUM VITAE

Surname, Name: JALLOW, Ebrima

EDUCATION

Degree	Institution	Graduartion Year
B.Sc.	Milli Savunma Universitesi Deniz Harp Okulu	2020