SOLENT UNIVERSITY, SOUTHAMPTON
FACULTY OF BUSINESS, LAW, AND DIGITAL TECHNOLOGIES

MSc Applied Artificial Intelligence and Data Science
Academic Year 2021-2022

Odesola Peter Adebayo

Heart Disease Classification Base on Machine Learning and Deep Learning

Supervisor: Dr Hamidreza Soltani

September 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Applied Artificial Intelligence and Data Science

ACKNOWLEDGMENT

All glory to God Almighty for his constant love towards me and for wisdom that can't be quantified bestowed for the completion of this thesis. My special appreciation goes to my supervisor **Dr. Hamidreza Soltani** for his relentless effort in making this thesis a reality. He was always available at any point he is needed, to guide me and give me the maximum supports needed. He motivated and mentored me to make sure the thesis was my own work. Without his keen interest, supports and knowledge, this project study could not have been successful.

I would also like to thank my supportive family who made this dream come through, supported me financially, morally, and in prayers. I wouldn't have survived this program without your love and supports, you have been so amazing. My special thanks to **Victoria Adebola Alade** who created this master's program platform, and for her massive interest and supports in my success.

My gratitude also goes to my darling wife **Hannah Olayinka Odesola** for always supporting me and her consistent prayers towards the success of this program. So much love to my son **Jayden Odesola** for being part of my success in this program.

Finally, I would like to thank all my friends for their moral supports in the United Kingdom and back in Nigeria. Thanks to my special friend **Dr Olalekan Olumide Oladejo** for his endless support, thanks for being a special and dedicated friend.

ABSTRACT

Cardiovascular disease has been the major cause of death in the past years. According to world health organisation statistical analysis in 2021, 17.9 million dies of heart disease annually (WHO 2021). It has been a major concern that most medical practitioners treat patients based on their knowledge and intuition, which is not good enough as so many lives have been lost based on error in diagnoses. The medical field have shown interest and vast development in using machine learning to solve some medical problems by using patient's medical records to create solution to some of the problems arising through the use of artificial intelligence. In the proposed study, an ensemble learning technique has been proposed in developing cardiovascular disease prediction. The researched study utilised a combined dataset from four database repositories, the Hungary, Switzerland, Cleveland, and Long Beach V. The study is a supervised classification problem that is based on predicting if a person has disease or not. Several algorithms were implemented for performance comparison which are Logistic Regression, Random Forest, Decision Tree, KNN, Multilayer Perceptron, Extreme Gradient Boost and Hybrid model (Hybrid of Random Forest, Decision Tree, and Extreme Gradient Boosting). The models used for ensemble techniques were chosen based on the best three models with good performances. Models were tuned to improve the performances using manual search and random search techniques for best parameters settings. Model evaluation was carried out by accuracy of models, precisions, F1-score, recall, confusion matrix and ROC curve. A cross validation was also used for estimation of performance of learning models. All models gave a good result, but the ensemble method was the best with 99% accuracy. A friendly Graphical user interface was developed for easy access to

users. The GUI was designed based on the open-source framework known as stream lit, with the ensemble model of Extreme Gradient Boosting, Random Forest, and Decision Tree to predict.

For more details on the dissertation artefacts, visit the link below:

https://github.com/peterodesola/Heart-disease-classification.git

Table of Contents

CHAPTER 1

1 INTRODUCTION

1.1 BACKGROUND

The cardiovascular system known to be a muscular organ which have the size of the fist of human hands, and it is composed of multiple layers of tissues. The heart supplies blood to the areas of human body through the arteries, the veins, and the blood capillaries of the cardiovascular system. This is essential in the body system which helps in the dissemination of oxygenated blood (purified blood) from the heart to the body system and collection of deoxygenated blood (unclean blood) from the body to the heart for purification. The rhythm and rate of the heartbeat is controlled by the cardiovascular electrical system. If the heart is unhealthy, affected by injury or diseases, the entire body system is affected. Heart disease keeps the heart unhealthy and affects the entire body system, such diseases are congenital heart anomalies, cardiovascular stroke, arrhythmias, coronary disease, pericardial disease, and cardiomyopathy.

In 2021, world health organisation reported a total number 17.9 million, 39 percent of global deaths caused by heart diseases yearly (WHO, 2021). Cardiovascular disease needs urgent attention, and solution because it is the major cause of death, and it is known to be the highest rated cause of death that is associated to the important tissues in the human body system. Majority of medical sicknesses have association with the heart, therefore it essential to develop a predicting system to diagnose disease of the heart and make a comparative research topic on heart disease a priority. The medical diagnostic has been so difficult due to numerous factors that are causing cardiovascular diseases **(Alqudah 2017, 215-222).** 20%-40% patients had been recorded with heart diseases due to

lack of proper diagnosing because of poor instrumental accuracy **(McClellan 2019.).** As a result, there is a necessity to further the study on more efficient algorithm to be implemented for heart disease prediction. Many studies had been previously done on cardiovascular disease prediction, but it is very critical to consider developing an effective and ideal algorithm for cardiovascular disease prediction to improve diagnosis and medical significant **(Pouriyeh et al 2017.)**

The features that can contribute to heart disease prediction are so many, but study was done by the researchers in the past to pay more attention on utilising the significant ones for heart disease predicting models **(Amin et al 2019, 36; 82-93).** At first, less attention was considered for feature importance, and how relevant the features are with the level of priorities to the predicting model **(Mohammed et al 2020.).** For accurate diagnosing, the weighted associative ruling mining (WARM) which was introduced to discover the relationship of various features and used in determining the associated rules with mining that had led to some outcomes through prediction **(Kannan 2018).** This has provided users with simple techniques to indicate the feature importance aiding heart diseases and considering accurate rules **(Altaf et al 2017).**

Unhealthy lifestyles or activities have been the major facilitator of heart diseases risk. Activities smoking, triglycerides level increase, obesity, unhealthy diet intake, alcohol consumption etc **(WHO 2020.).** The American Heart Association (AHS) announced some of the symptoms of cardiovascular diseases, such as insomnia, when legs are swelling and when heart beats are irregular **(AHS 2020).**

In recent years with the help of electronic medical records from hospital database, there have been great improvement in predicting heart disease to stop the disease from being fatal. Artificial intelligence has been impacting medical sectors positively, machine learning and deep learning can be used to predict, diagnose, or classify heart diseases. Individual records from hospital database can be converted and evaluated thoroughly to extract better predictions and implemented models can also be trained for knowledge pandemic prediction **(Shalev 2020).**

Heart disease is a deadly disease with highest death rate globally (WHO), and the society rely on medical domain to present reliable data in time. Medical information can be collected in the medical sector in the hidden pattern of dataset. For accurate results, pre-processed data are to be applied to machine learning or deep learning

algorithms, in which the medical practitioners can then make the right decision in treating the patients early before complicated. The implementation of data mining models for heart disease predictions are the reason for right prognosis of the disease **(Wu et al 2019, 7-11).**

Comprehensive research has been conducted on the risk factors importance to put into consideration based on medical records which can be considered for the creation of the predicting system using machine learning and model fitting. Implementation of an accurate predictive system can reduce death rates caused by the disease. It is important to make such services available for people at high risk of developing cardiovascular disease **(Maji 2009, 447-454).** It is of necessity to create awareness of heart disease risk factors, so everyone can be more sensitive to their lifestyles.

Due to the negative impact and high global death rate recorded as a of heart diseases globally (WHO, 2021), a proposed study on heart disease predicting system will be implemented. The dataset to be used for the algorithms of the predicting system are combined for the purpose of research, and the dataset date is from 1988, which was a combination of four repositories which are Cleveland, Switzerland, Hungary, and Long Beach V with 14 important features based on patients' medical records. Accuracy, precision, recall and F-1 score are the success metrics considered for effective performance of the models. Medical practitioners have been making decisions based on knowledge and experience which can lead to bias results, errors, and excessive medical cost. This study proposed creating an improved and accurate predictive heart disease system based on risk factors (features) from hospital records (dataset) using machine learning.

1.2 PROBLEM DEFINITION

Cardiovascular disease has been the major cause of life threats globally (WHO, 2021), many lives has been lost to heart diseases due to lack of early diagnosis. The major challenge faced by the medical health sector is inability to detect early stages of problems related to the heart. Detection of cardiovascular disease at early stage will reduce the rate of death statistics globally. For early cardiovascular disease detection, and death rate reduction caused by the heart problem, this study proposed implementation of improved, and accurate heart disease predicting system using machine learning knowledge.

1.3 PURPOSE AND RESEARCH QUESTIONS

In the proposed study, machine learning method would be implemented for heart disease prediction through the extracted medical knowledge history (dataset), and many machine algorithms will be developed to see algorithms with the best performance for the implementation of accurate predictive system. Many machines have been created for cardiovascular disease predictions, but recommendation for more research studies for improved and reliable model accuracy are being emphasized. Further, an approach on how to improve the performance of algorithms base on accuracy for implementation of a heart disease predictive system have been proposed (Mohan, 2021.), and base on this we address the following research questions:

- How reliable and accurate is a heart disease predictive system using the patient's risk factors or medical records for prediction?
- How can the accuracy of heart disease predicting system be improved with machine learning algorithms using patient's risk factor/historical records?
- Does feature reduction/ feature selection have any significant changes in the model performance?

1.4 AIM AND OBJECTIVES

This study is aimed at building an improved, accurate heart disease predictive system by using the hidden knowledge associated with various patient's historical records (dataset) collected.

The objectives are:

- To critically review previous works done on related topic.
- To carry out the exploratory data analysis of the heart disease dataset.
- To implement an algorithm that will accurately predict heart disease at early stage.
- To implement a friendly graphical user interface (GUI) for easy access to the users for diagnosis.

1.5 JUSTIFICATION

Cardiovascular diseases have been the major cause of death both in the young and adults due to lack of early diagnosis. Most decisions made by doctors or medical practitioners in hospitals are based on experience and insights when patients' records are filled with rich information that can be utilise in developing a predictive model. As a result of doctors using their initiatives to diagnose patients, there is bound to be bias, error, reduction in the quality of service provided and high cost of expenses. Implementation of a reliable and improved accurate predictive system will reduce the numbers of lives lost to heart disease. When heart disease is predicted rightly it saves life but when wrong predictions are made, lots of lives will be lost. To prevent or control the death rate globally, this research study proposed implementation of heart disease prediction using the voting ensemble classifier technique which was motivated by benchmark why trying to find ways of improving predicting machine with reliability. In this experimental research work, the machine learning algorithms and deep learning algorithm will be implemented but majorly machine learning algorithm will be mostly used because of the size of the dataset. The best three models will be ensembled using voting classifier to improve the performances of the models and to compensate each other's weakness, because model performances are more reliable and more accurate when different models are combined, compared to a performance of a single model **(Ho, Hull and Srihari 1994)**. Because this project is mainly to improve accuracy and reliability of heart

disease predicting system, ensemble learning is proposed due to its ability to overcome 3 machine learning problems which are statistics problem, computation issues, and representational challenge **(Dietterich 2002)**. Heart disease has been the highest cause of death globally and majority of the deceased was because of lack of diagnoses and incorrect diagnosis, and therefore this study has been proposed to eradicate or reduce high global death rate caused by heart disease. Model will be evaluated using precision, recall, accuracy, F1 score, ROC and cross validation to test the ability of the model to predict new data that has not been used for estimation, so that overfitting or bias selection can be flagged **(Cawley 2010)**.

1.6 PROPOSED ARTEFACT AND SOCIETAL IMPACT

A friendly graphical user interface (GUI) is proposed for this study. Stream lit is found in python language, it is a web application made for data science and machine learning with open-source framework. Stream lit can work perfectly with python libraries.

Development of a graphical user interface for cardiovascular disease prediction would grant the people in the society access to always examine themselves against heart disease symptoms. Development of such predicting system with a friendly easy to use web application will cause reduction in death rates, medical check-up cost and stress and will reduce biased results from medical practitioners.

The proposed software (web application) which would be developed using the stream lit graphical user interface library will have the following features:

- Feature display: This is where individual will enter their details base on the risk factors used in developing the predicting system.
- Title page: This will display the name or function of the system (heart disease predicting system)
- Prediction result display: This where the outcome of the heart disease tests will be displayed.

1.7 RESOURCES AND PROJECT IMPLEMENTATION

The resources needed for the development of heart disease predictive system proposed are basically software.

- Computer system (Windows 11, RAM: 8GB, Processor Inteli3: 2.59HGz, System type: 64bit)
- IDE: Pycham
- Programming language: Python 3.8
- Heart disease dataset: The dataset was collected through secondary source, it is a combination of five databases (Cleveland, Hungary, Switzerland, Stalog and Long Beach V) based on medical history, and it is in a csv format.
- Major python libraries needed
    - Matplotlib
    - Numpy: For mathematical performance based on arrays
    - Pandas: Use for loading files in various format
    - Seaborn: Visualisation function
- Machine learning algorithms for classification
    - Logistic Regression
    - Naïve Bayes Classifier
    - Random Forest Classifier
    - Extreme Gradient Boost Classifier
    - K Nearest Neighbour Classifier
    - Decision Trees Classifier
    - Stacking classifier
    - Voting Classifier
- Journals/Article: IEEE, Google scholar, Research gate, IJERT, Hindawi, National library of medicine and PARC.

CHAPTER 2

2. LITERATURE REVIEW

The healthcare sectors have huge data that can be turned to information through machine learning which can help to make informed decision and predictions **(Shalev 2020).**

The diagnosis of heart disease has led to numerous research study finding remedy to high death rates caused by heart disease using machine learning. Various data mining techniques have been applied for diagnosis and different results have been achieved based on the methods implemented.

(Reddy and Kanimzhi, 2022) had a study on dynamic KNN (D-KNN) with improved accuracy over supervised vector machine (SVM) for novel intelligent algorithm for cardiovascular disease prediction. Heart disease dataset was collected from UCL/Kaggle repository with more than 303 rows and 14 key features. The study was divided into two groups, while the data was split to two, the train and the test data. Twenty samples were allocated to training and 20 samples allocated to testing dataset, they were split and fit algorithms to predict accuracy values. Mean, standard deviation and standard error were the considered success metrics, while dynamic KNN (D-KNN) had the best performance with 80.29% mean accuracy and SVM had 69.05% mean accuracy. DKNN performed better than supervised vector machine for cardiovascular disease prediction.

(Suri et al, 2022) proposed a study on cardiovascular paradigm for powerful stratification danger with the multi-label, multi-class, and ensembled based machine learning paradigm application signifies that conventional approach to stratification of heart risk showed inferior result talk of performance compared to the fast-moving artificial intelligence method. Three paradigms which were multiclass, multi-label, ensembled techniques in office-based and stress test laboratory were proposed. 256 cardiovascular based studies were utilised by development of preferred reporting items to be viewed systematically and meta-analysis (PRISMA) model application. The attributes considered were architecture, pro and cons, application, validating scientifically, evaluation clinically, and bias danger which were comprehensively reviewed. Performance metrics considered were sensitivity, accuracy, specificity, AUC, and F1-score. The study showed that artificial intelligence based for heart-based assessment risk came out in great

success and seems promising when the three cardiovascular disease patterns in a non-cloud and cloud-based frameworks were used.

(Uddin et al, 2021) had a study using ensembled method which was based on multilayer dynamic system to predict cardiovascular disease using machine learning. Three classifier models were combined to form the ensemble which were Random Forest, Naïve Bayes and KNN. Also, the idea of increasing knowledge in every layer using ensembled method based multilayer dynamic system (MLDS). Correlation attributes evaluator (CAE), Gain Ratio Attribute Evaluator (GRAE), Information Gain Attribute Evaluator (IGAE), Lasso and Extra Trees Classifier (ETC) for feature selection were applied. Collection of datasets was via Kaggle, and the model came out with 88.84%, 89.44%, 91.56%, 92.72% and 94.16% accuracy respectively, while trained and test ratio were split on ratios of 50:50, 60:40, 70:30, 80:20 and 87.5:12.5. AUC 0.94 was achieved by the model which was in probability 94% to classify positive and negative classes correctly. The model had high accuracy with satisfactory performance. The use of neural network model was highly recommended.

(Bharathi, 2021) suggested the application of algorithms such as random forest, Naïve Bayes, decision trees and logistic regression for prediction of level of patient's risk and the chance of having heart disease. Data was collected through secondary source from UCL repository with 14 features. The machine algorithms performance was compared and analysed (Random Forest 90.16%, logistic regression 85.25%, decision tree 81.97% and Naïve Bayes 85.25%), Random Forest had the best accuracy with good performance when compared with the other algorithms.

(Kavitha et al, 2021) had study on prediction of cardiovascular disease using hybridisation model. Dataset was extracted from Cleveland with 14 features and 303 instances, they proposed the use of two algorithms and a hybrid model. Data pre-processing was carried out, dataset was split into test and train data before model fitting. Decision Tree and Random Forest algorithms were implemented, while Decision Tree had 79% accuracy, Random Forest 81%, and hybrid (Decision Tree + Random Forest) had the best performance based on accuracy with 88%.

(Bhoyar et al, 2021) in their study used multilayer perceptron to predict real time heart disease system. They collected dataset from UCL with 303 instances and 13 relevant features. The multilayer perceptron algorithm applied had better accuracy

than decision tree model. Multilayer perceptron had 85.71% while Decision tree had 73.45%. The second dataset was collected from Kaggle with 12 features and 70000 observations. In the second prediction multilayer perceptron had better accuracy with 87.30% while Decision Tree had 72.85% accuracy.

(Mohan et al, 2021) proposed a study on the use of supervised machine learning algorithms to diagnose cardiovascular disease and dataset was extracted through Kaggle. Four different algorithms were used, K-Nearest Neighbour, Naïve Bayes, Random Forest, and Logistic Regression. Logistic regression came out with the best predicting accuracy 90.2%, while other accuracies were Random Forest 86.9%, Naïve Bayes 86.9% and KNN 82.0%. The study proposed the use of other machine learning algorithms for future study in improving performances.

(Krithika and Rohini, 2021) proposed a study on application of bigdata analytics in cardiovascular prediction using ensembled method. Dataset consists of 14 attributes and 96655 observations and various model were applied for the disease prediction. Data was well pre-processed, while logistic regression, Naïve Bayes, Random Forest, K-Nearest Neighbour, Decision Tree, support vector machine, ANN and HPTRF (Hyper parameter tuned random forest classifier) were the algorithms implemented for the predictive system. HPTRF can out with the best performance, it had 96.54% trained data and 80.43% test accuracy. ROC curve showed HPTRF had the prediction value. The paper suggested the use of more factors for future works.

(A. Lakshmanarao, A. Srisaila and T. S. R. Kiran 2021) had study on forecasting heart disease by feature selection and ensembled method. This study was carried out with two datasets, one from Kaggle with 16 features and 4238 observations while the other dataset was collected from UCI with 303 instances and 14 features.

UCI dataset was pre-processed, applied classification algorithms with sampling techniques. Success metrics was based on accuracy which are, Logistic Regression 85%, KNN 64%, SVM 70%, Decision Tree 80%, Naïve Bayes 81%, Random Forest 77%, Adaboost 85%, Stacking classifier 84.5% and voting classifier with highest accuracy 90%. Kaggle dataset result in reference to accuracy are, Logistic Regression 60%, KNN 79%, SVM 64%, Decision Tree 91%, Naïve Bayes 60%, Random Forest 96%, Adaboost 64%, voting classifier 99% and stacking classifier with highest accuracy 99%. The study showed that ensembled techniques was efficient.

(Prakash et al, 2020) proposed study on visualisation and variant models, heart

disease predictions which was conducted with dataset collected from UCI repository. Dataset contains 300 records and 14 attributes. Data cleaning was carried out, data exploration, and attribute selection was carried out. Five models were selected for the study, logistic regression, support vector machine, Adaboost, Naïve Bayes and random forest. The success metrics used were precision, recall and F1-score, in which logistic regression and support vector machine came out with best performance based on precision, recall and f1-score. Enhancement of accuracy with some other algorithm was proposed.

Islam et al, 2020 proposed a study on the use of machine learning paradigms to forecast cardiovascular disease with the help of machine learning. Dataset collection was from UCL with 301 instances and 13 features. Logistic regression, Naïve Bayes, support vector machine and decision trees where the machine learning algorithms used. The accuracy of the algorithms was Logistic regression 86.25%, Naïve Bayes 73.77%, support vector machine 83.61% and decision trees 75.41%. Logistic regression was the best algorithm based on accuracy.

Looking into the literature reviews of the study done on heart disease prediction, it is very clear that more research work is needed in this area, in creation of an improved and reliable predictive system.

(Deepika and Sasikala, 2020) proposed a study on the use of decision tree and swarm particle optimisation to boost model prediction and classifying heart disease. Dataset was extracted from Cleveland with 7 features, in which the data was pre-processed, and the machine learning algorithm (Decision trees) came out with good accuracy of 83.61% with 0.02 seconds prediction time.

(Chakarverti et al, 2019) proposed study on the use of data mining for heart disease classification technique in which supervised vector machine and K-nearest neighbour algorithms were implemented. They used data mining technique in getting useful information from a raw dataset, both alike and unlike data types were classified using KNN classifier. The study showed that KNN used same number of parameters which gave higher accuracy compared to SVM likewise execution time was low. The study suggested enhancing hybrid design classifier for heart disease prediction.

(Dinesh 2018) suggested study on cardiovascular disease forecast using machine learning, the dataset was collected from UCL repository. The dataset had 14 features with 920 observations, and appropriate machine learning approach was intensely

carried out. Classifier algorithms used were support vector machine, logistic regression, and random forest. Logistic regression had the best performance with 91.6% accuracy compared to the other algorithms.

(Ambekar and Phalnikar, 2018) proposed a study on using Convolutional Neural Network in predicting heart disease risk. Dataset was collected from UCL repository with 12 features and CNN-UDRP algorithm was used, and the performance was great based on recall (60%) and precision (62%). It had accuracy of 65%. They suggested to add more diseases with the prediction of the risk factors patients are suffering.

(Chauhan et al, 2018) had a study on using evolutionary rules in cardiovascular disease prediction. They used mining rules algorithm for rule extraction on cardiovascular disease dataset. Dataset from Cleveland database was utilised. With the application of data mining techniques, a strong association rules were generated to create more accurate heart disease predictive system.

(Sowmiya and Sumitra, 2017) conducted a study on application of classification technique of analysing research on heart disease diagnosis. In the study, several classification methods are important in data mining in cardiovascular disease were analysed. The study showed that classification-based techniques added high efficacy and acquire good accuracy when compared to the previous methods.

(Purushttam et al, 2015) had a study on efficient forecast of cardiovascular disease with decision tree algorithm, in which they collected their dataset from Cleveland with 303 instances and 14 features. The algorithm came out with performance evaluation of 86.7%.

CHAPTER 3

3. METHODOLOGY

The heart disease prediction study is proposed to be a quantitative research method, with a scientific approach. The dataset to be use are in numbers (numeric) with 14 features. Two approaches would be implemented in the study for model fitting to compare the performances and to see how reliable and accurate the predicting system would be. The first approach would be addressed with all the 14 features in the dataset, while the second approach would be based on feature selection and the feature would be selected based on how important the features are in the dataset before fitting with the eight algorithm classifiers and checking for performances evaluation. The hybrid ensemble method has been proposed for improve performance of the heart disease predicting system.

Heart disease prediction framework.

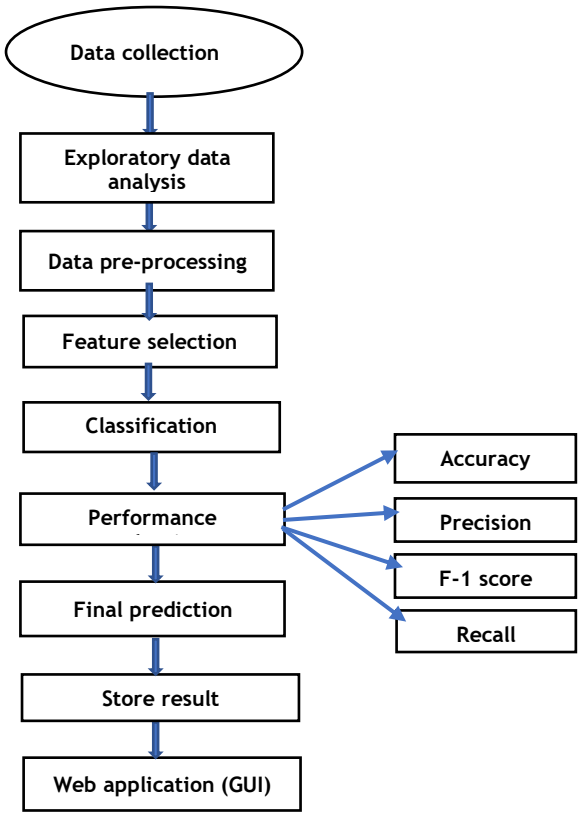

Figure 3.1. Heart disease prediction framework.

3.1 APPROACH TO LITERATURE REVIEW

A systematic review was the method used for literature review approach. PRISMA means preferred reporting items for systematic reviews and meta-analysis, designed to review transparency report systematically. **(Selçuk 2019)**

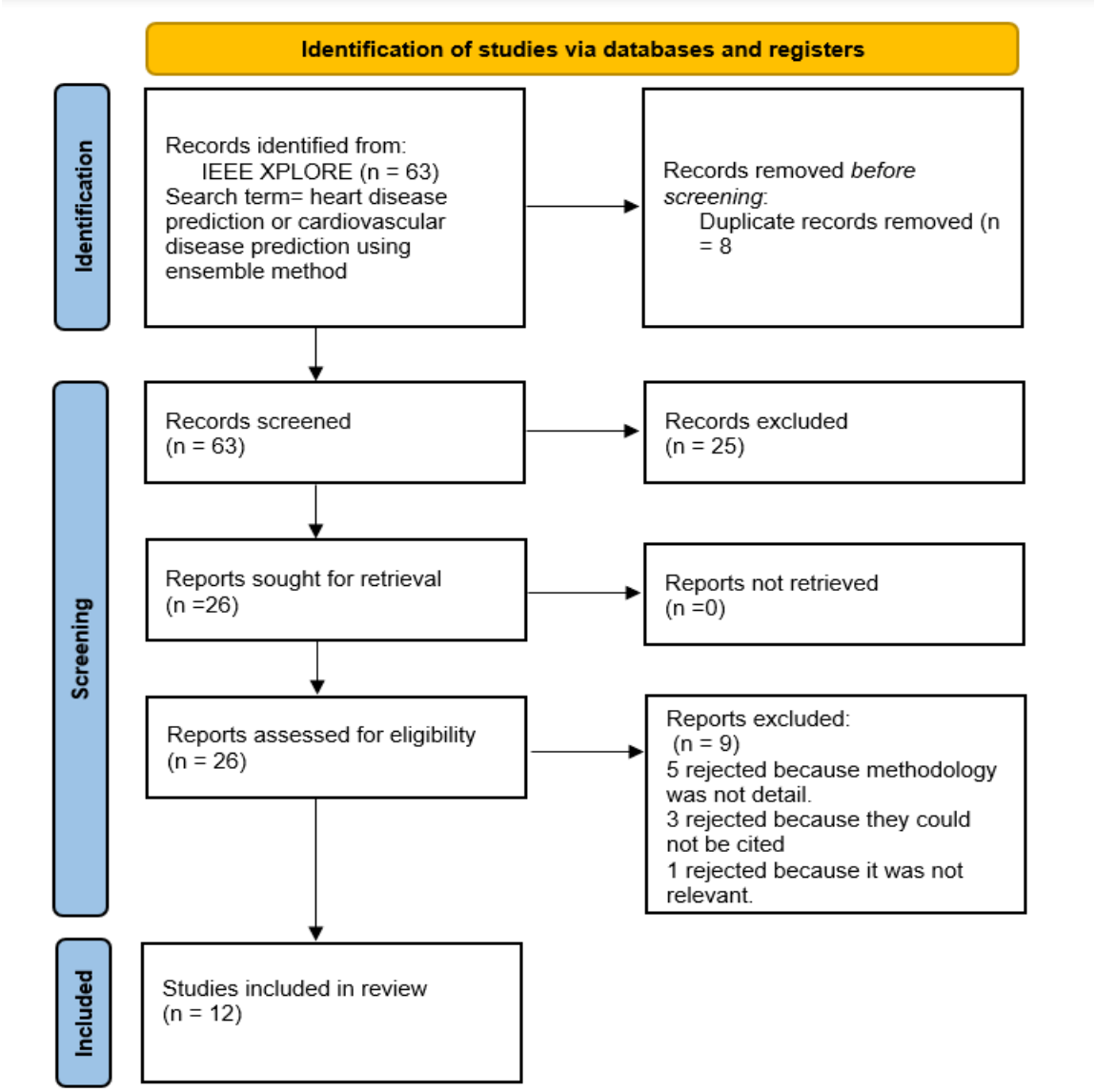

Figure 3.2 PRISMA flowchart of results from systematic literature search (PRISMA 2021)

As above, this literature review has identified 63 records from IEEE Xplore on cardiovascular disease, 8 duplicates taken out, 25 records were excluded, 26 records were assessed for eligibility, 5 records were rejected because the method used was not detailed, 3 were rejected for lack of citation, 1 report was rejected because it was irrelevant, and 12 studies were included in the review.

3.1.1 Ethics

Data was collected through secondary open data source, which is a collection of data from four data bases Cleveland, Hungry, Switzerland and Long beach V. Ethics clearance was applied, and approval was issued in compliance with Solent University's Ethics Policy, Appendix A.

3.2 Data Collection

Data collection was through secondary source because of limited time frame for the execution of the thesis. The dataset was dated from 1988 and was collected from four data base sources which are the Cleveland, Switzerland, Long Beach V and Hungary. The original dataset contains 76 attributes which include the target variable, the 76 attributes was reduced to 14 by categorising the relevant from the irrelevant attributes using feature selection (John et al, 1994.). While target variable indicates the patients with cardiovascular disease and those that has no heart disease.

The features or risk factors are collected from hospital database based on patients' medical records. The dataset consists of 14 columns (features) and 1025 rows (observation) which are stored in csv format for machine learning. For this study, seven algorithms will be implemented while the best three algorithms with the best performance will be ensembled for development of an improved heart disease predicting system.

3.2.1 Data insight

s/no	Variable	Values	Data type	Variable type
1	**Sex**	**Gender (male=1, female=0)**	Categorical numeric Nominal Int64	Independent
2	**Age (years)**	**The patients age in years**	Numerical continuous Ratio Int64	Independent
3	**Serum Cholesterol**	**Cholesterol measurement in mg/dl**	Numerical Int64	Independent
4	**Fasting blood sugar**	**Patients fast blood sugar (>120mg/dl, 1=true; 0=false)**	Numerical nominal Binary Int64	Independent
5	**Chest pain type**	**Value 0= asymptomatic** **Value1= atypical angina** **Value2= non-anginal pain** **Value3= typical angina**	Categorical Numeric Nominal Int64	Independent
6	**Resting bps**	**The patients resting blood pressure**	Numerical Int64	Independent
7	**Resting electrocardiographic results**	**Resting ECG results**	Numerical Nominal Int64	Independent
8	**Thalach**	**Maximum heart rate (range 60-202)**	Numerical Nominal Int64	Independent
9	**Exang/ Exercise induced angina**	**Pains caused by activities or stress.**	Numerical Nominal	Independent
10	**Old peak**	**Electrocardiogram result of old peak**	Numerical Float64	Independent
11	**Thalassemia**	**It is a blood disorder value 1= fixed defect** **Value 2= normal blood circulation** **Value 3= reversible defect**	Nominal Numerical Int64	Independent
12	**Slope**	**Slope of the peak ST segment 0: downslope;1: flat;2: upslope 0: downslope;1: flat; 2; upslope**	Nominal Numerical Int64	Independent
13	**Ca**	**The number of major vessels (0-3)**	Categorical	Independent
14	**Target**	**Heart disease (1=no, 0=yes)**	Nominal Binary Int64	Dependent

Table 3.1: Details of heart disease dataset

3.3 Exploratory data analysis (EDA)

Exploratory data analysis is a phase in machine learning that helps data analysts or data scientists to carry out investigation on a data, for easy detection of dataset patterns, and to identify anomalies, to spot hypothesis and to observe assumptions through the help of statistical summaries and graphical visualisation. Exploratory data analysis gives better insight of the dataset **(Greiff 2000)**.

The types of datasets were checked as shown in fig 3, to know the data type of the attributes we are dealing with, and to address them according to their datatypes or nature.

The exploratory data analysis could be:

3.3.1 Univariate Analysis

In univariate data analysis, only one variable is analysed. This means one feature, or a column is analysed. Univariate analysis can be graphically executed by histograms, pie chart, box plots, etc.

3.3.2 Multivariate data analysis

Multivariate data analysis can be the process of analysing the relationship between two or more variables. Multivariate analysis can be displayed using scatterplot, heat map etc. The data analysis will be carried out using both the univariate analysis and multivariate analysis.

3.3.3 STATISTICAL ANALYSIS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.00
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.75
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.03
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.00
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.00
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.00
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.00
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.00

Fig 3.3 Data description

In figure 3.3, the count, mean, standard deviation, minimum values, and percentiles (25%, 50%, and 75%) were checked for the statistical information.

3.3.3.1 Variable distribution

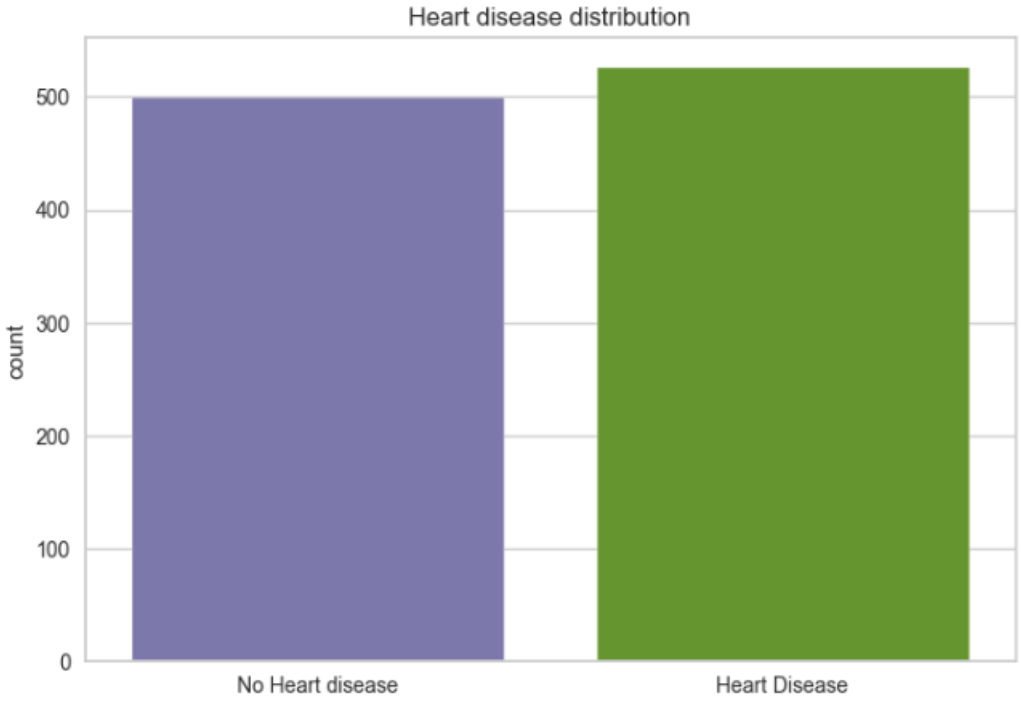

Figure 3.4: Target variable distribution

From the target variable distribution in figure 3.4, the visual display showed that the number of people with heart disease are slightly more than those without disease based on the data when counted. Considering the target variable distribution, the data are almost having the same distribution.

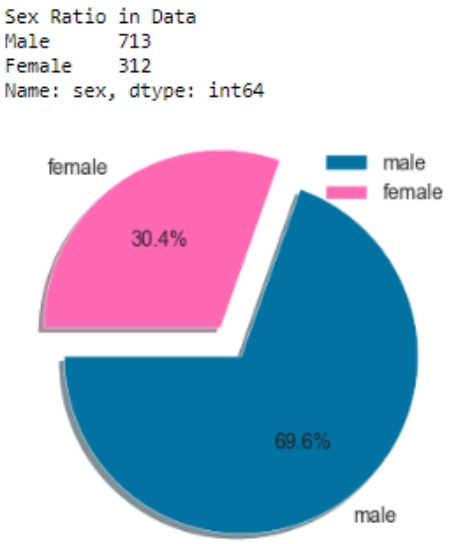

Fig 3.5 Sex distribution

Figure 3.5 pie chart graphical display showed that the males outnumbered the females. The distribution for sex variable is imbalance, the percentage of males is much more than the females.

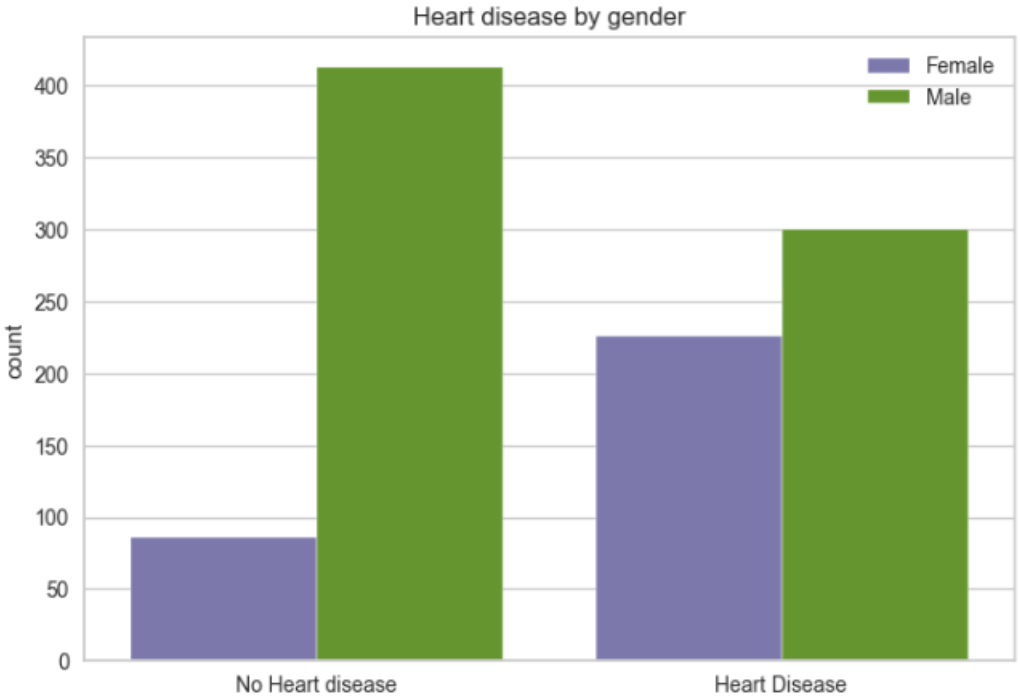

Figure 3.6 heart disease status distribution with gender

Considering the visual display in fig 3.6, the male outnumbered the female and based on this, there is imbalance in the distribution of “no heart disease”.

3.3.3.2 Correlation matrix

Correlation can be defined as reviewing statistical measure of the strength of the relationship between two variables in a dataset. We have the positive and negative

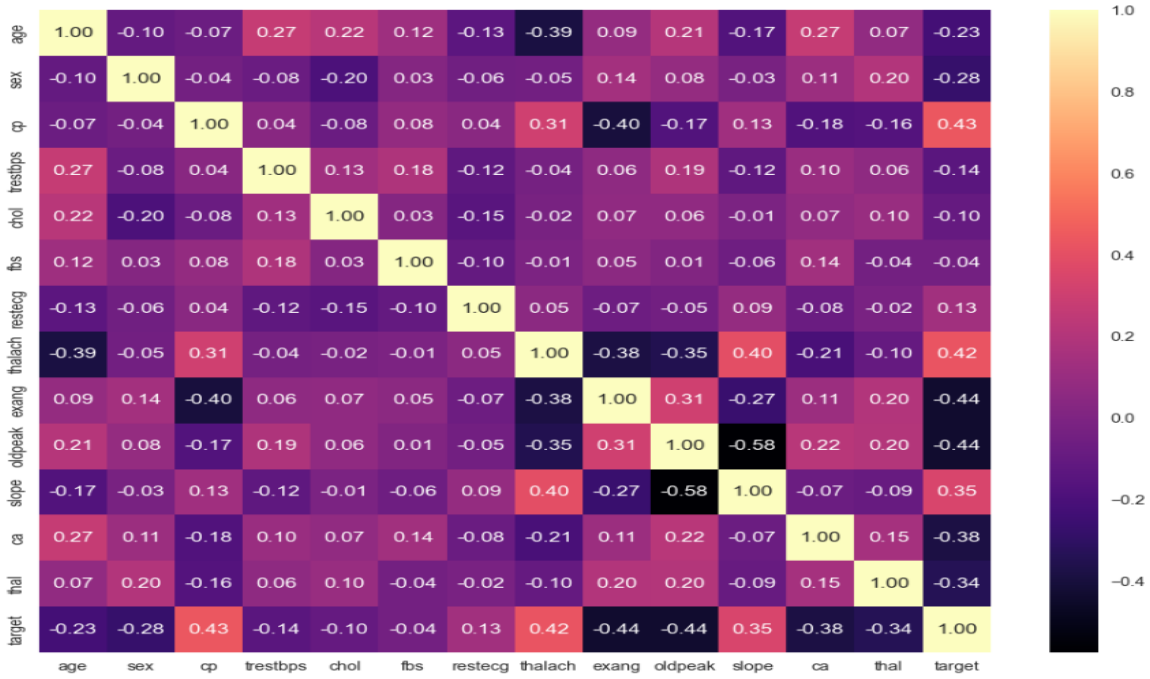

Figure 3.7 `heatmap between the 14 features

Figure 3.7 shows that the correlation between the features is not strong. The values with positive shows they portray a potential positive correlation (increase) while a negative value indicates a decrease. The heatmap gives good insight knowledge of the data.

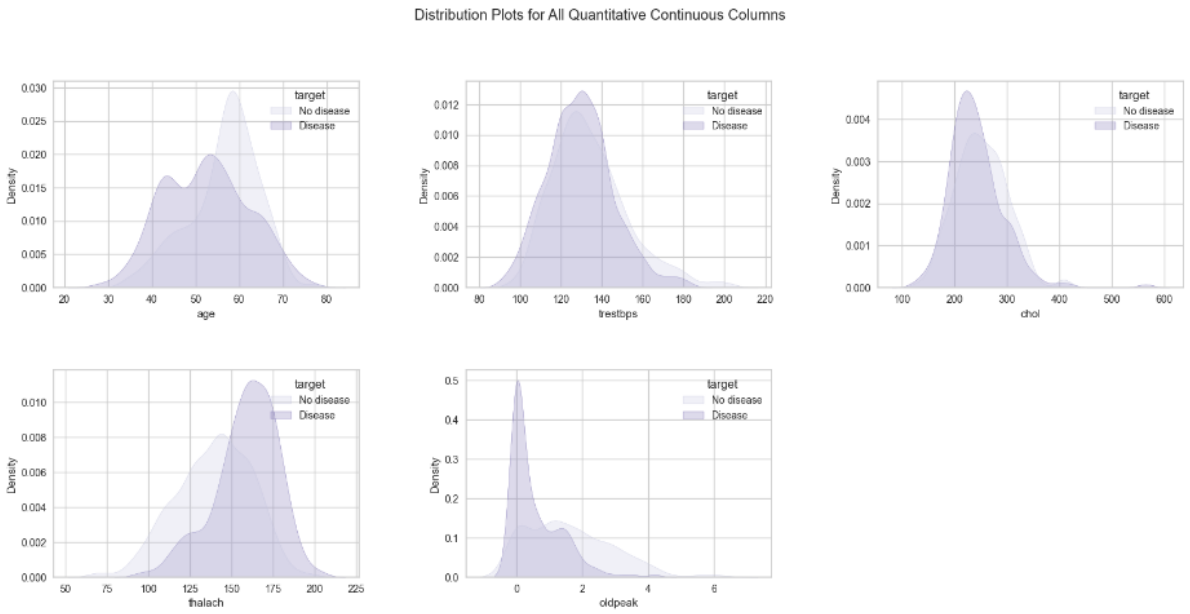

Figure 3.8 Distribution of the quantitative continuous variables

Figure 3.8 have shown that age, resting blood pressure, and maximum heart rate have normal distribution while cholesterol and old peak are not well distributed (skewed). The old peak has mesokurtic distribution, in the process of visualisation, some attributes where not well distributed and a boxplot was used for outliers' detection as shown in fig 3.9, which shows outliers in the dataset.

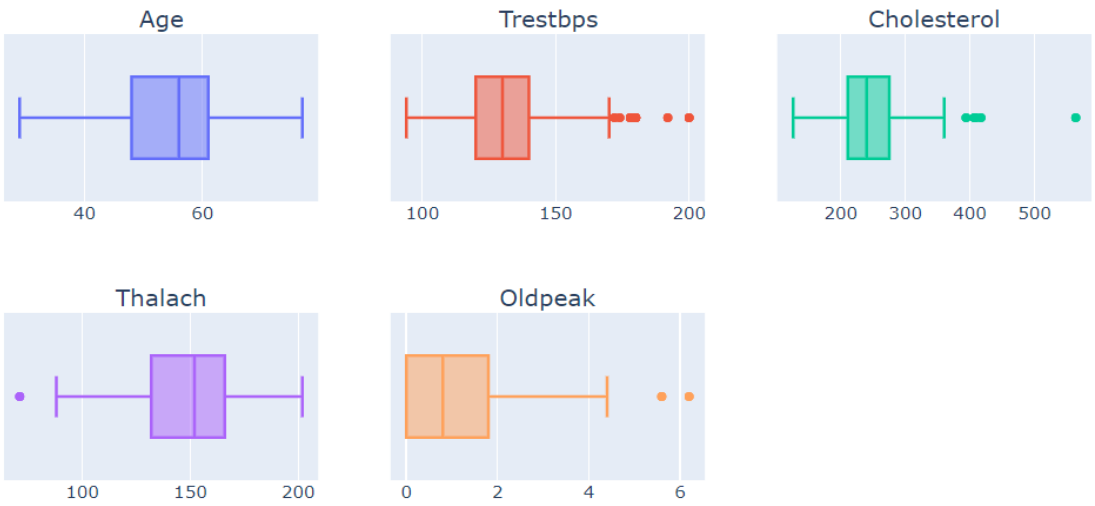

Figure 3.9: Visualising the outliers

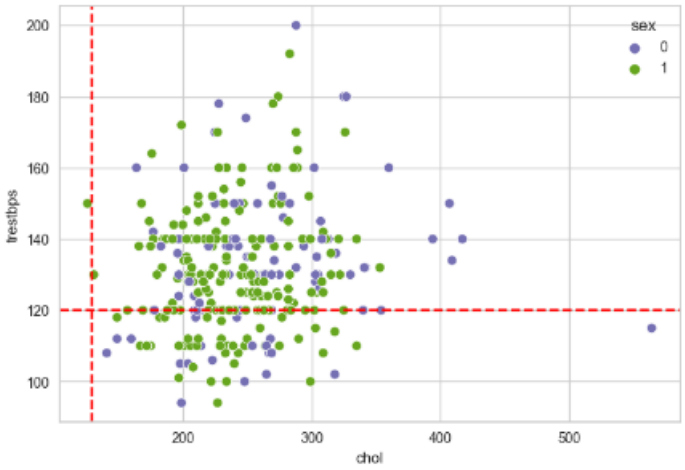

Figure 3.10 Scatterplot of resting blood pressure and cholesterol

In figure 3.10, we could see scatter plot displaying the correlation and ranges of the datasets of resting blood pressure and cholesterol in respect to sex or gender.

3.4 Data pre-processing

Data pre-processing is an important phase in data mining where statistical analysis is observed because of inconsistency in the real-world data. Data pre-processing phase involves the transformation of raw or unprocessed data into an understandable data format. This is the stage where fixing or removal of incorrect, corrupt, duplicated, empty and incomplete values are carried out for best performance of the models and improved accuracy. Datasets are known to be messy, and they need to be well examined for enhancement of algorithms and outcome. The following steps were taken in the pre-processing phase:

- Import the libraries: Important libraries were imported for the machine learning study such as pandas, seaborn, numpy, scikit learn, matplotlib etc.
- Import and load dataset: The heart disease csv dataset was imported for reading using appropriate command.
- Checking and handling of missing data or values: Dataset must be checked if there are missing data and must be well handle. It is either removed or replace. There was no missing value recorded, which means all values were imputed as shown in figure 3.12 using missingno library.

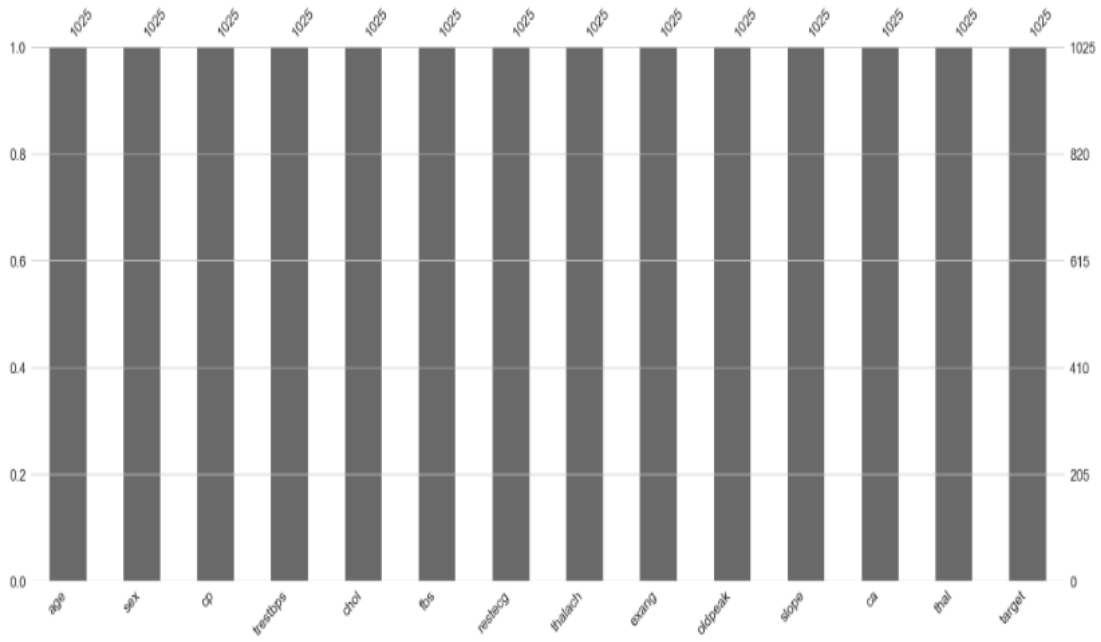

Figure 3.11: Checking for missing values

- Checking for duplicates: Duplicate is when a particular information or data is being repeated or same data appearing multiple times. Therefore, duplicates must be observed in dataset to avoid observation from same records of a particular record. The duplicates observed in the dataset were removed appropriately.
- Checking and handling of outliers: Outliers are the values or observations that lies abnormally to other values in distance and could occur because of data input error, sampling problems, and natural variation. Some outliers demand immediate removal while some could be left as they will not affect performance. (Smith and Martinez, 2011) in their research study on classification improvement in increasing accuracy by detecting and removing observations that are misclassified. The outliers in the dataset were appropriately removed using the interquartile range for effective performance of the models.
- Check for categorical values in the data (one-hot encoding): This is the phase where all categorical data were converted to numerical values by applying one hot encoding imported from sklearn pre-processing library for the machine to read. By so doing, machine models can now make their decisions on how the labels can be operated.

- Feature scaling of the data: This is the final pre-processing stage. Feature scaling is one of the important methods used in limiting the range of variables, so each variable can be evaluated on common grounds, no bias. It was applied to independent variables. It was used for normalising the data within a certain range because the dataset is not in gaussian distribution (skewness) observed figure 3.8. It also helped to speed up the calculations in the proposed algorithms **(Saranya and Manikandan 2013)**. Both the trained and test sets were scaled using min max scaler imported from scikit learn library. The scaling output was in an array format which was converted to data frame with the use of panda's data frame method.
- Split dataset into train and test set: Splitting of data is very important to avoid biasness when evaluating predictions. The dataset was split into train and test set, where the trained set was 70% and test was 30%. We had the following parameters for the train and test sets.

  X train which happened to be the trained part of the matrix of features.

  X test which is the test part of matrix features.

  y train was the training part of the dependent variable that is associated to X train.

  y test was the test part of the dependent variable that was associated to the X test.
- **Feature importance/selection:** Feature importance is necessary for identification of how relevant or influential an attribute is in the data. It is used in dropping irrelevant features for dimensionality reduction and to improve the performance of models **(Fisher, Rudin and Dominici 2019)**. The feature importance was checked using extreme gradient boost classifier (XGBoost Classifier), and the feature importance figure 3.12 showed that chest pain (CP) has the highest level of importance, then thalassemia, ca, exang and fast blood sugar has the least relevance as shown in the figure 3.12.

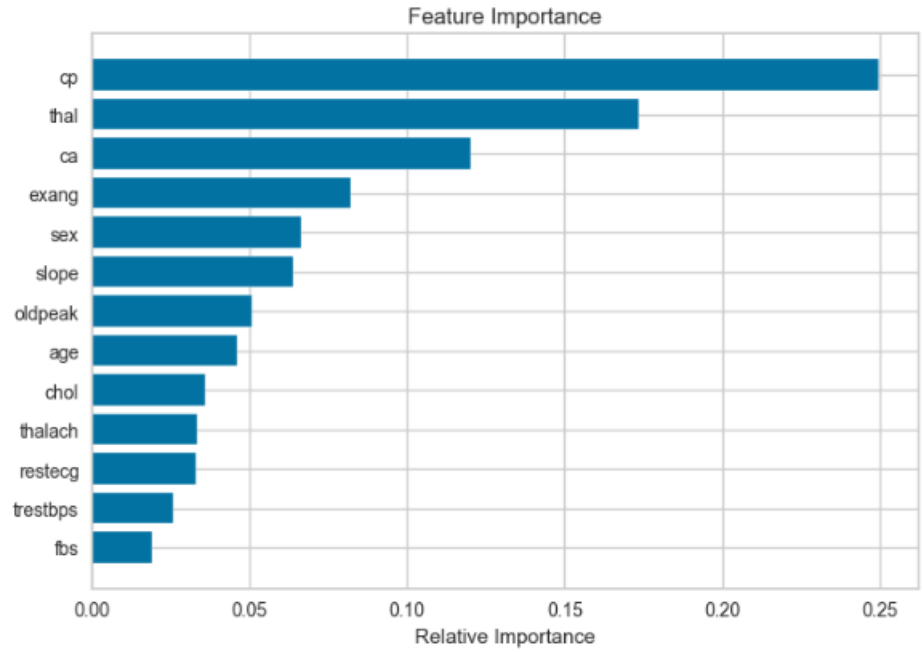

Figure 3.12 Feature importance

3.5 MODEL DEVELOPMENT

3.5.1 ALGORITHM SELECTION

Model development is the next stage after dataset have been put to good shape or condition for efficient performance of the models. The algorithms were selected based on the problem to solve. The data is a supervised machine learning, and it is based on classification indicating if heart disease is present or not. It is a binary classification problem with dependent (target), and independent variables. The rules of selecting machine learning algorithms for classification were considered **(Quinlan, J. Ross 2014)**, likewise for the proposed dataset rules have been generated based on different priority **(Quinlan, J. Ross 1995)**. Based on algorithm selection rules **(Quinlan 2014)**, K- Nearest Neighbour, Naïve Bayes, Logistic Regression, Decision Trees, Random Forest, Multilayer Perceptron and Extreme Gradient Boosting classifiers have been selected to test their performance by comparing their accuracies, precisions, recalls and F-1 scores to choose the three best performing algorithms for ensemble technique for the final modelling. The selected machine learning algorithms classifiers will be discussed in the next session.

K- NEAREST NEIGHBOUR (KNN)

K-Nearest neighbour can be described as a method of problem solving according to solutions to the known problems. K-nearest neighbour learning system requires setting parameters:

- Measuring the similarities between problems or data entries based on distance function. This technique is required to measure the closest neighbours to the new problem.
- Neighbour's numbers are considered when the new problem is being addressed.
- Increasing prediction and learning quality by a weighting function to activate further quantification of found neighbour.
- Creation of evaluation techniques that can describe functions on ways of approach to problem solving on the found neighbour.

K- nearest neighbour (KNN) does not need a computation to be implemented on dataset before query is given to the system. This technique contrasts with eager learning technique such as Decision trees, which try to implement the structuring of information before queries are received **(Ting, Vijayakumar and Schaal 2010)**.

NAÏVE BAYES CLASSIFIER

Naïve Bayes also known as Bayesian networks is made of nodes which connections are directed between the nodes which indicates dependencies between the nodes. The Bayesian networks are known to be probabilistic directed acyclic graphical models. They are one of Bayesian network simplest models in combination with kernel density and are scalable with required number of linear parameters present in the variables. The known data is used in estimation of the dependency between the features and the class labels whereas this information is extracted to calculate the probabilities of different outcomes of future events that could possibly be generated. Bayes' theorem is automatically applied to complex problems which makes it easy to achieve knowledge about the feature state and their dependencies. Mathematical expression:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \text{ ---------- } (1)$$

Where:

- A and B are the events
- P(A) is known as the probability that A event will occur.
- P(A/B) signifies the probability that B is to be true provided A event occurred **(Rosenblad 2011)**

Each node is labelled with distribution based on probability which defines the parent node effect on the child node **(Cleophas and Zwinderman 2013)**

DECISION TREES

A Decision Tree is one of the common learning methods of classification methods that focuses on easy representation forms that are understandable. A Decision Tree is built by splitting data iteratively on features, also into possible existing classes a certain criterion for stop is reached. The representation gives the user an overview of the dataset.

The Iterative Dichotomise 3 (ID3) and its successor the C4.5 algorithm are one of the first algorithms concerning the Decision Trees training developed by Ross Quinlan in 1986 and 1993 **(Quinlan, J. R. 1986) (Salzberg 1994)**. Decision Trees are direct trees created for the support of decision tools, which represents the rules of decisions and showed success in decisions. Root node, inner node and end nodes which are also known as leaf are the branches of decision trees node.

Root nodes: They are the beginning of decision support process with no incoming edges.

The inner nodes: They have one incoming edge and minimum of two outgoing edges. The test is based on the feature of the data set **(Geisler and Quix 2020)**

Training a Decision Trees is basically a classification technique. The main goal is to predict the value of target attribute in relation to number of input attributes.

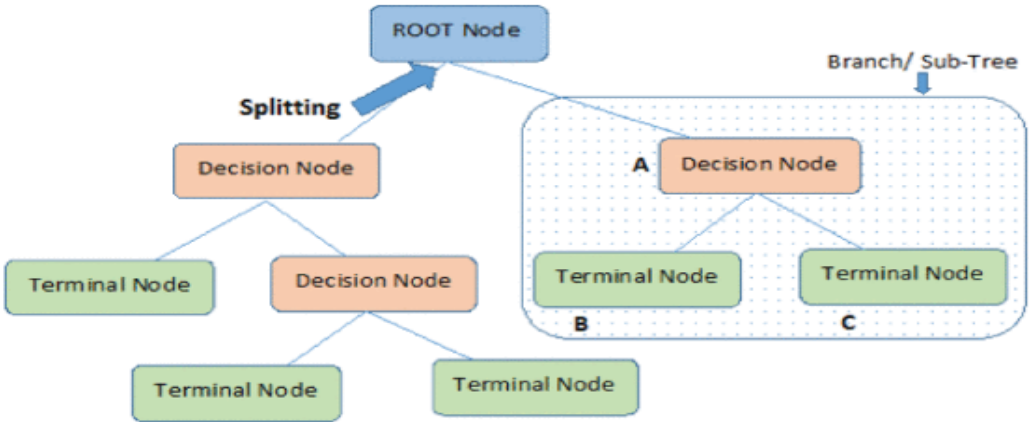

Figure 3.13 Decision Tree of heart disease dataset **(F. B. Chioson *et al*. 2018)**

RANDOM FOREST

Random Forest a bagging algorithm modification that is constructed dynamically with one fitting procedure. Random Forest is known to be a meta estimator used in the fitting of numbers of tree classifiers on various sub samples of datasets while prediction accuracy is improved and as well rectifies overfitting using the average in summing the values. In Random Forest classifier, generalising error could be because of capability of trees separated in the forest and the trees correlation. Random forest classifier main advantage is the enhancement of accuracy, and it is also prone to overfitting **(S. Pachange, B. Joglekar and P. Kulkarni 2015)**

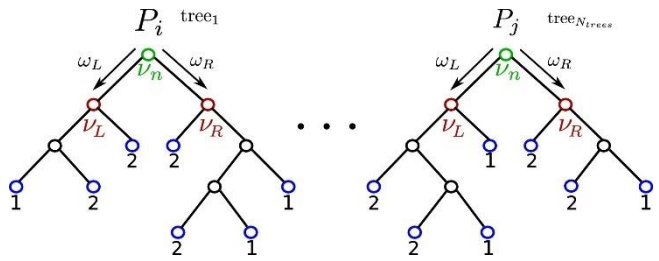

Figure 3.14. Schematic representation of Random Forest **(Izquierdo-Verdiguier and Zurita-Milla 2020)**

LOGISTIC REGRESSION

Logistic Regression algorithm is known to be supervised algorithm that can be use in predicting for discrete outputs (yes or no, true, or false) **(Golande and Pavan Kumar 2019)**. The logistic regression predicts values that are between 0 and 1 and its function perform with S shaped sigmoid function.

Mathematical expression:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))----------(2)

Where:

x= value input

y= predicted output

b0= intercept term

b1= coefficient for single input value (x)

MULTILAYER PERCEPTRON

A multilayer perceptron is a feedforward artificial neural network of fully connected class. Multilayer perceptron can also be called vanilla neural network most times when they possess single hidden layer **(Hastie, Tibshirani and Friedman 2009)** Multilayer perceptron has minimum of three layers, the input layer, the hidden layer, and the output layer. MLP make use of backpropagation for training **(Van Der Malsburg 1986)**

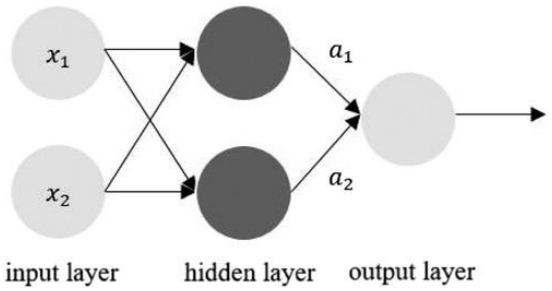

Figure 3.15 Multilayer Perceptron **(Tang, Zhang and Ding 2019)**

EXTREME GRADIENT BOOSTING CLASSIFIER (XGBOOST CLASSIFIER)

Extreme gradient boosting is known to be an ensemble machine learning model created by Tianqi Chen in 2004. It can be used for solving both classification and regression problems and has been proven by its performances and efficacy machine learning competitions to be one of the best algorithms. It has various parameters for regularisation that reduces overfitting and helps to improve overall machine learning performance. XGBoost enhance speed as well as performance by the implementation of gradient boosting decision trees **(Chen and Guestrin Aug 13, 2016).**

3.5.2 MODEL FITTING

Model fitting is the phase in which machine learning models are measures on how models generalise to similar data on which they were trained. A well fitted model performs better and generates accurate outcome.

Appropriate data pre-processing (missing values checked, outliers checked and handled, duplicate checked and removed, data skewness considered) was carried out and dataset was put into good condition for better performance, also machine learning rules were considered to get the best of our model performance **(Quinlan 2014)** All the steps for data pre-processing and data splitting were applied to all the seven proposed algorithms, and dataset was splits into training and test sets (X train, X test, y train, y test respectively) at 70% to 30% ratio, where 70% train and 30% test. Models were all fitted based on individual standard fitting codes.

In the model fitting phase, two approaches were considered for model fitting; the first approach, models were fitted with all the 14 features while in the second approach we selected 8 features to be fitted with the proposed algorithms based on feature importance fig 3.12. For model fitting, seven algorithm classifiers (K Nearest Neighbour, Logistic Regression, Naïve Baye, Multilayer Perceptron, Decision Trees, Random Forest, and Extreme Gradient Boosting) were implemented.

3.5.3 MODEL COMPLEXITY

Model complexity can be defined as the function of complexity that is to be learned, which could be like a polynomial degree. Model complexity level can be determined in respect to the nature of the training data. Failure in spreading the amount of data or the entire data uniformly throughout different possible scenarios, then there should be consideration in the model complexity reduction because when there is high model complexity there would be overfitting on a small number of data points. Overfitting is then process of training models that suitably fit the trained data but fails to generalise to other data **(Schneider, Xhafa and Linke 2002).**

(M. Pidd 1996) in his study on "The five principles of simulating models", stated a simple model is preferable to complex ones, simplicity is the essence of simulation. **(Ward 1989)** stated model simplicity relating it to transparency (relating it to understanding) and constructive simplicity (relating to the model itself) and reinforced the idea of model simplicity in his study.

Model complexity was seriously put into consideration to avoid poor performance of proposed models. Models were tuned to correct errors that might affect performance due to overfitting.

(Bergstra and Klop 1982) stated that the use of column subsampling is much more effective than conventional row subsampling when preventing overfitting. The subsampling of rows was done using hyper-parameter subsample, colsample_bytree while the tree structure was established calculating the leaf scores, regularisation, and the objective functions at each level. The tree structure was used over again in subsequent iterations which will eventually reduce model complexity. Extreme gradient boosting hyper-parameters are many and the hyperparameters can be implemented in carrying out some tasks as the model desires. They have their default hyperparameter for optimisation if not set at the initial, but the parameters can be inputted as desired by chosen model.

3.5.3.1 HYPER-PARAMETER OPTIMISATION

Hyperparameter optimisation or tuning in machine learning is the process by which a set of optimal hyperparameters are chosen for a learning algorithm, the values are basically used to control the learning process of algorithms to find the best performing algorithm when evaluating based on validation set **(Feurer and Hutter 2018)**.

There four common hyperparameter optimisation.

- Manual search
- Grid search
- Random search
- Bayesian optimisation

Random search is the process by which random combinations of various hyperparameters are utilised in getting the solution for the model built, while the manual hyperparameters are set manually and tuned for model optimisation. The random search optimisation was adopted for this study because of its ability to run the train-predict-evaluate cycle that can be done automatically in a loop of hyperparameter on a pre-decided grid, also the manual search was used for optimisation of some models. The selected parameter for optimisation is shown in figure 3.16 for hyperparameter tuning. Generalisation performance was estimated using cross validation **(Bergstra and Klop 1982)**

```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.4, gamma=0.1, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.1, max_delta_step=0, max_depth=10,
              min_child_weight=1, missing=nan, monotone_constraints='()',
              n_estimators=100, n_jobs=0, num_parallel_tree=1, random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

Figure 3.16 Random search result for best parameters used

3.5.3.2 CROSS VALIDATION

Cross validation is also known to be rotation estimation is a model validation technique that is used in evaluating how statistical analysis result will generalise to data that are independent. Cross validation is known as a resampling technique which utilises different portions of the data that uses different iterations to test and train a model **(Kohavi and John 1995)**. The main goal of cross validation is to test the ability of the model to predict new data that has not been used for estimation, so that overfitting or bias selection can be flagged **(Cawley 2010)**. Cross validation was implemented on the models to improve and evaluate the performance of the model, if over-fitted, underfitted or accurate.

3.5.4 ENSEMBLE CLASSIFICATION APPROACH

The ensemble classification technique is the combination of several models to improve the result of a machine learning, this technique influences the production of better predictive performance when compared to single model **(Ho, Hull and Srihari 1994)**. **(Dietterich 2002) in his study on sequential data in machine learning proved that ensembles technique can overcome three problems, which are:**

1. Statistical problem: Statistical problem happens when the amount of available data has large hypothesis spaces.
2. Computational problem: this occurs because of inability of a learning algorithm to guarantee finding the best hypothesis.
3. Representational problem: The representational problem can be caused

when hypothesis space fail to have any good approximation of classes that are target.

For the proposed of this study, two ensembled classifiers were implemented to evaluate their accuracies and individual ensemble technique performance. The two adopted ensemble techniques are:

1. The stacking classifier approach
2. The voting classifier

The two ensemble classifier techniques were used for comparison of results, to evaluate the performance of the two techniques. The ensemble technique was adopted to improve the accuracy and performance of the models. Finally for the purpose of this study three best performing models from the proposed algorithms were ensembled using the stacking classifier and the voting classifier. The three models ensembled were the Decision Trees Classifier, the Random Forest Classifier, and Extreme Gradient Boosting Classifier.

3.5.5 MODEL EVALUATION

Model evaluation is the process by which machine learning model performance, strength and weaknesses are being checked using different evaluation metrics **(Hossin and Sulaiman 2015)**. Confusion matrix, accuracy score, precision, recall, AUC and F-1 score were used for the model evaluation.

CONFUSION MATRIX

Confusion matrix is a structured table that contains the true values and predicted values known as true positive and true negative, it can be use in the evaluation of performance of classification models **(Vujovic 2021)**. The target variable comprises of two values, which are positive and negative.

The columns are the actual values of target variables while the rows are the predicted values of our target variables. The confusion matrix table is divided into four parts, two columns and two rows respectively.

- True positive (TP): Here the values are identified as true because they are indeed true.
- False positive (FP): The values are predicted as true/ positive in the division, but they are false/ negative.
- False negative (FN): The values were predicted as false/ negative, but they

are true/ positive.

- True negative (TN): The predicted value is negative while the actual value is negative.

ACTUAL VALUE			
POSITIVE	NEGATIVE		
TP	FP	POSITIVE	PREDICTED VALUE
FN	TN	NEGATIVE	PREDICTED VALUE

Table 3.2 Confusion matrix

The column is the actual values while the row is predicted values.

PRECISION AND RECALL

Precision explains the number of correctly predicted cases, which eventually turned out to be positive.

Precision mathematical expression:

$$\text{Precision} = \text{TP}/ (\text{TP} + \text{FP}) \text{ ----------(3)}$$

This is designed to know if our model is reliable or not.

Recall is based on the exact number of the actual positive cases that was able to be predicted correctly with our models.

Recall mathematical expression:

$$\text{Recall} = \text{TP}/ (\text{TP} + \text{FN}) \text{ ----------(4)}$$

Precision and recall for our models can easily be calculated by inputting the values into the precision and recall equations.

F1 SCORE

F1 score can be defined as the harmonic mean between the recall and the precision.

F1 score can be used for statistical measure used in rating performance.

Mathematical equation for F1 score:

$$\text{Fbeta} = \frac{(1 + \beta^2)\ \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \text{ ----------(5)}$$

ACCURACY

Accuracy is one of the evaluating metrics which measures the number of observations for both positive and negative, that were classified correctly.

Accuracy equation:

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \quad \text{----------(6)}$$

ROC AUC

AUC simply means area under the curve. ROC curve must be defined first. ROC visualises the trade-off between true positive rate (TPR), and false positive rate. The receiver operating characteristics (ROC) is one of the best ways of visualising model classifier performance to select the best and suitable operating point, or decision threshold. It can also be used for cross validation of the classifier's overall performance **(Bradley 1997).**

3.5.6 SOFTWARE DEVELOPMENT

A web application was implemented for the purpose of making the heart disease predicting system easy to use and friendly. The graphical user interface was created from the python-based libraries called stream lit. Stream lit is an open-source application framework from python language, which is use for the creation of web applications in data science and for machine learning project in a very short time. Stream lit works well with the major python libraries such as numpy, scikit-learn, matplotlib, pandas etc.

The web application (GUI) has been developed with twelve features which are displayed on the graphical user interface for service users to enter the information. The values of the features are to be entered based on users' medical information for diagnostic purpose. The web application developed has been trained by model saved in the job lib, the voting ensembled classification algorithm has been trained to predict user input.

The web application contains the predicting system name, the user input features lists and the predicting phase. The web application is shown below.

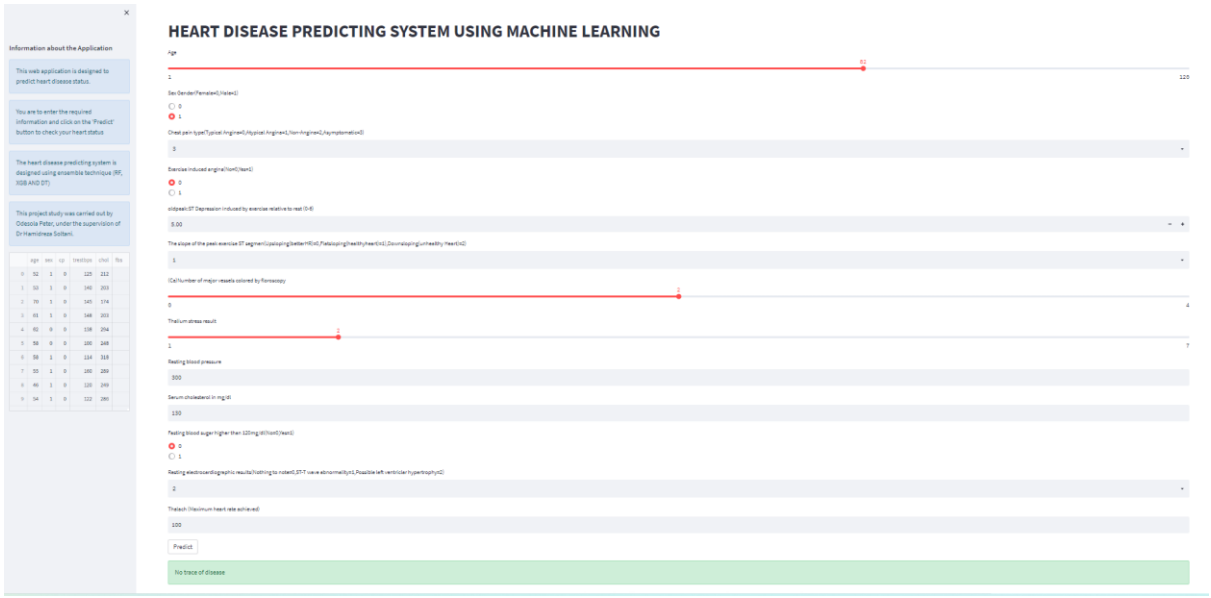

Figure 3.17. Graphical user interface

CHAPTER 4

4.0 RESULT AND DISCUSSION

4.1 RESULTS

First approach results using all the 14 features

The values that were obtained for different algorithms using confusion matrix

Algorithms	True Positive	False Positive	True Negative	False Negative
Logistic Regression	115	29	138	9
K Nearest Neighbour	107	37	109	38
Decision Trees	134	10	139	8
Multilayer perceptron	77	67	124	23
Naïve Bayes	111	33	132	15
Random Forest	125	10	144	3
Extreme Gradient Boosting	141	3	147	0
Stacking Classifier	141	3	147	0
Voting classifier	141	3	147	0

Table 4.1 Confusion matrix summary table

Evaluation table for testing set based on accuracy, precision, recall, and F1 score for first approach

Algorithms	Accuracy	Precision		Recall		F1 score	
Heart disease status		0	1	0	1	0	1
Logistic Regression	87%	93%	83%	80%	94%	86%	88%
K Nearest Neighbour	74%	74%	75%	74%	74%	74%	74%
Decision Trees	94%	94%	93%	93%	95%	94%	94%
Multilayer perceptron	69%	77%	65%	53%	84%	63%	73%
Naïve Bayes	84%	88%	80%	77%	90%	82%	85%
Random Forest	92%	98%	88%	87%	98%	92%	93%
Extreme Gradient Boosting	98%	100%	97%	98%	100%	98%	98%
Stacking Classifier	99%	100%	98%	98%	100%	99%	99%
Voting classifier	99%	100%	98%	98%	100%	99%	99%

Table 4.2 Accuracy, precision, recall, and F1 score summary table

Second approach using just 8 features displaying only the results of the proposed models

The values obtained for different algorithms using confusion matrix

Algorithms	True Positive	False Positive	True Negative	False Negative
Decision Trees	139	17	138	14
Random Forest	138	18	147	5
Extreme Gradient Boosting	143	13	142	10
Stacking Classifier	153	3	152	0
Voting classifier	153	3	152	0

Table 4.3 Confusion matrix summary table

Evaluation table for testing set based on accuracy, precision, recall, and F1 score for second approach

Algorithms	Accuracy		Precision		Recall		F1 score	
Heart disease status			0	1	0	1	0	1
Decision Trees	90%		91%	89%	89%	91%	90%	90%
Random Forest	93%		97%	89%	88%	97%	92%	93%
Extreme Gradient Boosting	92%		93%	92%	92%	93%	93%	93%
Stacking Classifier	99%		100%	98%	98%	100%	99%	99%
Voting classifier	99%		100%	98%	98%	100%	99%	99%

Table 4.4 Accuracy, precision, recall, and F1 score summary table

The explanations of how the precision, accuracy, F1 score, and recall are calculated using confusion matrix are shown in chapter 3 model evaluation section.

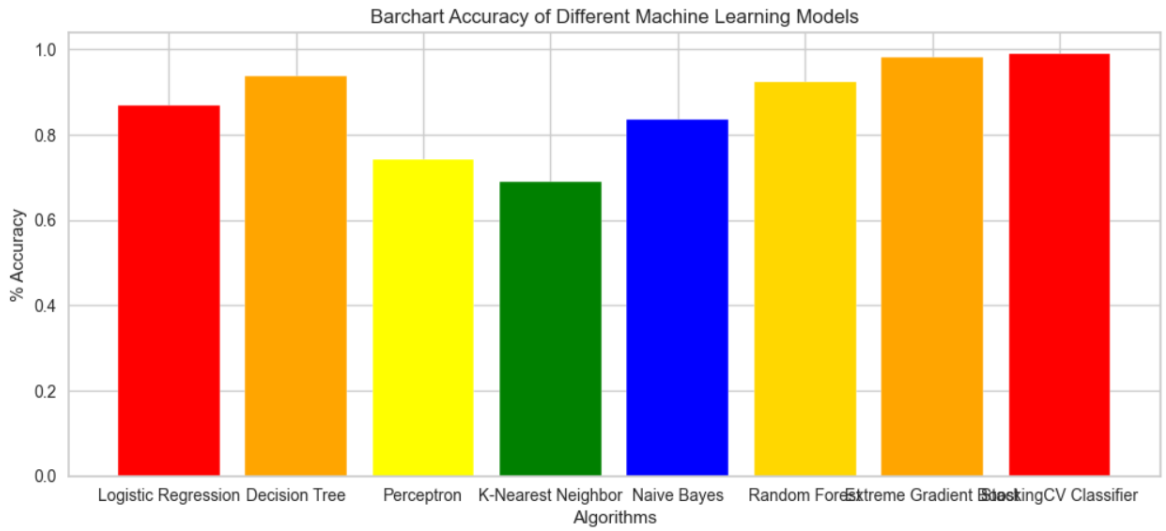

Figure 4.1 Bar chart displaying model accuracy

The bar chart in figure 18 display the accuracy of the models, the two-ensemble classifier have the best performing accuracy.

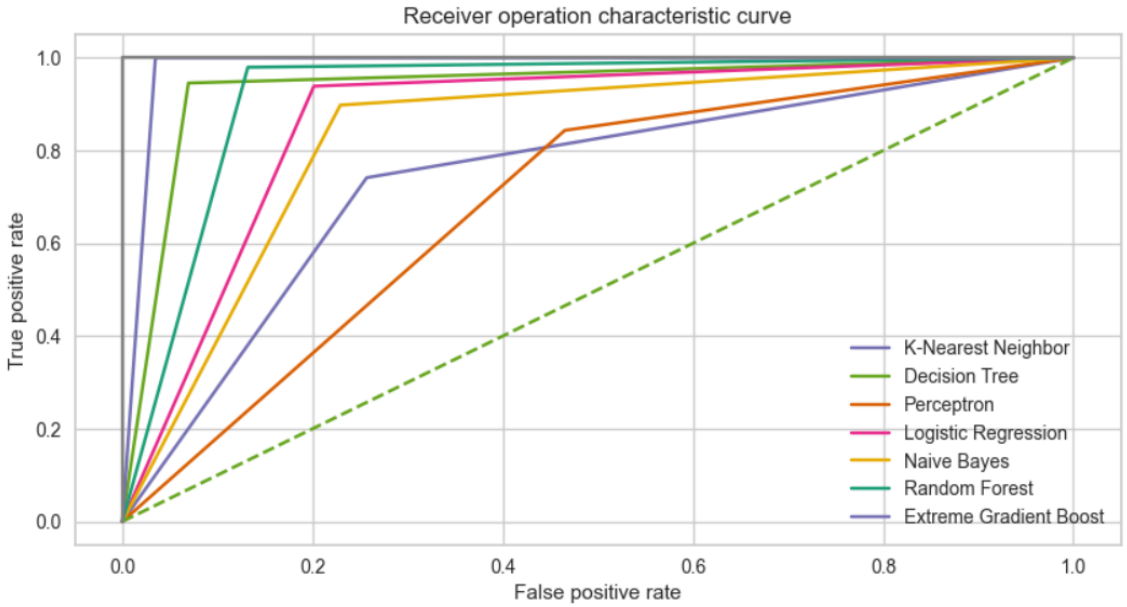

Figure 4.2 ROC curve

The ROC curve in figure 20 shows a threshold for performance and the performances of each model.

4.2 DISCUSION

In the proposed research work, the ensembled technique was adopted based on performances evaluation. Six machine learning algorithms (Logistic Regression, Naïve Bayes Classifier, K- Nearest neighbour Classifier, Random Forest Classifier, Decision Tree Classifier and Extreme Gradient Boosting Classifier) and one deep learning algorithm (Multilayer Perceptron) were implemented to compare individual performance in the classification problem to select the best three to be ensembled. The pre-processed dataset was used in carrying out the experiments for the two approaches and the above listed algorithms for classification were applied.

MODEL TUNING

This is the process by which the performance of model is being optimised by setting hyperparameters for the model. The values are basically used to control the learning process of algorithms to find the best performing algorithm when evaluating based on validation set **(Feurer and Hutter 2018).**

Hyperparameters

The parameters for each model were implemented through random search and manual search with 10 folds cv, the parameters for the three models for ensemble learning are stated below.

Random Forest Classifier: n-estimators= 20, random state= 2, max depth= 5

Decision Trees Classifier: max depth= 6, random state= 1.

Extreme Gradient Boosting Classifier: base score= 0.5, booster= 'gbtree', colsample-bylevel=1, colsample-bynode=1, colsample-bytree=0.4, gamma= 0.1, learning rate=0.1, max depth= 10, n-estimators=100, n-jobs=0, random state= 0, reg lamba=1, subsample=1, validate parameters=1.

Three best algorithms (Random Forest Classifier, Decision Tree Classifier and Extreme Gradient Boosting Classifier) were adopted to be ensembled based on performance evaluation in tables 4.1, 4.2, 4.3, and 4.4, while other algorithms had a promising result as well but only the best three were ensembled figure 4.1. Two approaches were developed for implementation of models to compare their performances. The first approach was implemented using all the 14 features in the dataset, while the second approach was implemented using 8 selected features based on feature importance to see if there would be variations in the results and

performances based on numbers of features. The two approaches had no significant differences in the yielded general results for the ensembled techniques, but there were slight changes in the results of individual algorithms due to dimensionality reduction tables 4.2 and 4.4. But ensembled approach compensated for the performances **(Dietterich 2002)**.Therefore any of the ensembled methods can be used for prediction and development of web application. The hyperparameters for the algorithms were optimised for effective performance of the models using manual search and random search figure 17 respectively **(Feurer and Hutter 2018)**.

Table 4.1 and table 4.2 results are for the first approach used for acquiring our results, whereby the outliers were removed, and all 14 features were fit, which shows a good and promising results. Random Forest, Decision Trees, and Extreme Gradient Boosting Classifier had a good evaluating performance. The proposed ensemble classifiers came out with an outstanding result based on the model evaluation metrics **(Hossin and Sulaiman 2015).**

ROC

Fig 20 shows the receiver operation characteristic curve (ROC) displaying the performance of each model. Extreme Gradient Boosting Classifier had the highest performing accuracy in the ROC, followed by Random Forest Classifier and Decision Tree Classifier, they are the best three best models as indicated in the ROC. The ROC shows the threshold line in broken green lines, any model that performs below the threshold indicates poor performance **(Bradley 1997).**

Looking into the confusion matrix of all the models, we would see that the confusion matrix of the proposed model (ensemble classifier) came out with outstanding results, high rate of true positive and true negative, no record of false negative and three record of false positive which looks promising for the model performances **(Chicco and Jurman 2020)**.

4.3 PERFORMANCE COMPARISON AGAINST BENCHMARK STUDIES

Comparing the performance of this study with benchmark studies models proposed by Kavitha et al. 2021, **(M. Kavitha *et al*. 2021)**, and **(A. Lakshmanarao, A. Srisaila and T. S. R. Kiran 2021)** because they had similar study on heart disease prediction using the ensemble technique.

(M. Kavitha *et al*. 2021) had a research study on heart disease prediction using

ensemble, in which a novel method of machine learning was developed. Three machine learning algorithms were implemented which were, Random Forest, Decision Tree, and Hybrid model. Random Forest and Decision Trees were ensembled and the accuracy was 88.7%. Decision Tree was 79% accuracy, while Random Forest yielded 81% accuracy. The model was evaluated using MSE, MAE, RMSE and accuracy.

(D. R. Krithika and K. Rohini 2021) researched on predicting heart disease with ensemble technique. The algorithms implemented were XGBoost with 74% accuracy, Decision Tree with 72% accuracy, KNN (71%), SVM (72%), Logistic Regression (70%), Naïve Bayes (58%), Random Forest (73%), ANN (71%) and Hyperparameter tuned Random Forest Classifier (96%).

(A. Lakshmanarao, A. Srisaila and T. S. R. Kiran 2021) in his study on heart disease prediction using ensemble learning approached his study was through feature selection using ANOVA F-value and Mutual information. The best features were selected using the selection approach, and three techniques were applied which are random over sampling, synthetic minority oversampling and adaptive synthetic sampling approach. Two datasets were used for implementation, one from UCL (Dataset 1) and the other from Kaggle (Dataset 2). The results of the 2 datasets are shown below.

Algorithm	Precision	Recall	Accuracy
Logistic regression	86%	88%	85%
KNN	69%	65%	64%
SVM	68%	88%	70%
Decision Tree	84%	79%	80%
Naïve Bayes	83%	85%	81%
Random Forest	81%	76%	77%
Adaboost	86%	88%	85%
Stacking classifier	85%	84%	85%
Voting classifier	89%	90%	90%

Table 4.5 Lakshmanarao 2021 dataset 1 result

Algorithm	Precision	Recall	Accuracy
Logistic regression	62%	61%	60%
KNN	74%	93%	79%
SVM	66%	64%	64%
Decision Tree	86%	99%	91%
Naïve Bayes	72%	36%	60%
Random Forest	93%	100%	96%
Adaboost	65%	68%	64%
Stacking classifier	100%	99%	99%
Voting classifier	79%	73%	76%

Table 4.7 Lakshmanarao 2021. dataset 2 result

Comparing the results of the three-benchmark study by Kavitha et al. 2021, **(M. Kavitha *et al*. 2021)**, and **(A. Lakshmanarao, A. Srisaila and T. S. R. Kiran 2021)** with the result of our research study (tables 4.2 and 4.4, figure 4.1 and 4.2) we would observe the improvement in the accuracy of our models. Comparing this study and the benchmark studies, the hyperparameters for this study was duly observed for the models, likewise the model evaluation was put to consideration effectively and a cross validation was implemented to give insight on how model can generalise to an independent dataset which has helped in the improvement of results and performance. A friendly graphical user interface was developed for easy use figure 3.17: https://github.com/peterodesola/Heart-disease-classification.git.

4.4 LIMITATIONS

The time frame for the execution of the thesis was a short one, and due to the limited time caped for the completion of the research work, there were several limitations to put into consideration.

- Sample size of dataset: The dataset sample size was 1025 with 14 features, which is not large enough to implement for deep learning algorithms and as a result we could only implement just one deep learning algorithm and the result was not impressive but promising.
- The main participants record in the dataset were from the US, UK, Hungarian and Switzerland, hence it is not a global represented dataset which means

the outcome of the result might not favour other races because of genetic variation.

- Model evaluation: Due to limited time allocated for the completion of the thesis, we will limit the success metrics of our model evaluation to accuracy, precision, F1 score, ROC, recall and implementing a cross validation to generalise model performance.
- Dataset features: The dataset used only identifies 14 features as risk factors for cardiovascular disease whereas there are more important risk factors such as excessive alcohol intake, smoking, obesity, unhealthy diet, family history etc that should be considered.

CHAPTER 5

5 CONCLUSION AND FUTURE WORK

An ensemble classifier model was proposed in this thesis for heart disease prediction, in which the data set used was a combination of four repository database (Cleveland, Switzerland, Long Beach V and Hungary). Two approaches were implemented, first approach was implemented using 14 features while feature selection was adopted in the second approach and eight features were selected based on feature importance. In the process, various classifier algorithms were used to compare the performances of individual algorithms in which Extreme Gradient Boosting classifier had the best performance with 98% accuracy, then Decision Trees Classifier with 94% accuracy and Random Forest Classifier with 92% accuracy (table 4) before ensemble classifier was implemented using stacking classifier and voting classifier, which gave 99% accuracy. Table 4.2 and table 4.4 shows the results for the two approaches and are both promising. The experimental results for the two techniques show that the proposed model was efficient and had a good performance over the benchmark study.

In the future, the ensemble technique will be applied using large dataset and more of the predictions will be done using deep learning algorithms. Also, the proposed algorithms will be implemented for other dataset with different features.

6. References

A. LAKSHMANARAO, A. SRISAILA and T. S. R. KIRAN, 2021. Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques. - *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. pp.994-998

Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: a survey. Artif Intell Rev. 2017;47(3):313–40.

A. M. Alqudah, "Fuzzy expert system for coronary heart disease diagnosis in Jordan", Health Technol*, vol. 7, 2017; pp. 215-222.*

American Heart Association, Classes of Heart Failure, American Heart Association, Chicago, IL, USA, 2020, https:// www.heart.org/en/health-topics/heart-failure/what-is-heartfailure/classes-of-heart-f

Amin MS, Chiam YK, Varathan KD Identification of significant features and data mining techniques in predicting heart disease. Telem Inform. 2019;36;82–93.

B. E. K. GÜZEL and D. ÖNDER, 2018. Performance comparision of boosting classifiers on breast termography images. - 2018 26th Signal Processing and Communications Applications Conference (SIU). pp.1-4

BERGSTRA, J.A. and J.W. KLOP, 1982. Algebraic specifications for parametrized data types with minimal parameter and target algebras. *Automata, Languages and Programming*, 140, 23-34

BRADLEY, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145-1159

CAWLEY, G.C., 2010. Over-Fitting in Model Selection and Its Avoidance. *Advances in Intelligent Data Analysis XI*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.1

C. D. ANISHA and N. ARULANAND, 2020. Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers. - 2020 International Conference on Innovative Trends in Information Technology (ICITIIT). pp.1-6

CHEN, T. and C. GUESTRIN, Aug 13, 2016. XGBoost. Ithaca: ACM, pp.785-794

CHICCO, D. and G. JURMAN, 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6

CLEOPHAS, T.J. and A.H. ZWINDERMAN, 2013. Introduction to Machine Learning. In: T. J. CLEOPHAS and A. H. ZWINDERMAN, eds. *Machine Learning in Medicine*. Dordrecht: Springer Netherlands, pp.1-15

D. R. KRITHIKA and K. ROHINI, 2021. Ensemble Based Prediction of Cardiovascular Disease Using Bigdata analytics. - *2021 International Conference on Computing Sciences (ICCS).* pp.42-46

DIETTERICH, T.G., 2002. Machine Learning for Sequential Data: A Review. In: T. CAELLI*, et al.,* ed. *Structural, Syntactic, and Statistical Pattern.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp.15-30

F. B. CHIOSON *et al.,* 2018. Classification and Determination of pH Value: A Decision Tree Learning Approach. - *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM).* pp.1-4

FEURER, M. and F. HUTTER, 2018. Towards further automation in automl. *ICML AutoML workshop.* pp.13

FISHER, A., C. RUDIN and F. DOMINICI, 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J.Mach.Learn.Res.,* 20(177), 1-81

GEISLER, S. and C. QUIX, 2020. Database Management Systems (DBMS). In: L. A. SCHINTLER and C. L. MCNEELY, eds. *Encyclopedia of Big Data.* Cham: Springer International Publishing, pp.1-6

GOLANDE, A. and T. PAVAN KUMAR, 2019. Heart disease prediction using effective machine learning techniques. *International Journal of Recent Technology and Engineering,* 8(1), 944-950

GREIFF, W.R., 2000. The use of Exploratory Data Analysis in Information Retrieval Research. In: W. B. CROFT, ed. *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval.* Boston, MA: Springer US, pp.37-72

HASTIE, T., R. TIBSHIRANI and J.H. FRIEDMAN, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer

HO, T.K., J.J. HULL and S.N. SRIHARI, 1994. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 16(1), 66-75

HOSSIN, M. and M.N. SULAIMAN, 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process,* 5(2), 1

IZQUIERDO-VERDIGUIER, E. and R. ZURITA-MILLA, 2020. An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation,* 88, 102051

Kannan AG, Castro TARVC, BalaSubramanian R. A comprehensive study on various

association rule mining techniques; 2018.

K, B., 2021. Heart Disease Prediction Using Machine Learning. Revista GEINTEC, 11(4), 3694-3702

K. G. DINESH et al., 2018. Prediction of Cardiovascular Disease Using Machine Learning Algorithms. - 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). pp.1-7

KOHAVI, R. and G.H. JOHN, 1995. *Wrappers for feature subset selection.* Elsevier BV, pp.273

Maji S, Arora S. Decision tree algorithms for prediction of heart disease. In Information and communication technology for competitive strategies. Springer, Singapore; 2019, pp. 447–454.

M. KAVITHA *et al.,* 2021. Heart Disease Prediction using Hybrid machine Learning Model. *- 2021 6th International Conference on Inventive Computation Technologies (ICICT).* pp.1329-1333

M. McClellan, N. Brown, R. M. Califf and J. J. Warner, "Call to Action: Urgent Challenges in Cardiovascular Disease: A Presidential Advisory from the American Heart Association", Circulation*, vol. 139, pp. e44-e54, 2019.*

Mohammed KI, Zaidan AA, Zaidan BB, Albahri OS, Albahri AS, Alsalem MA, Mohsin AH. Novel technique for reorganisation of opinion order to interval levels for solving several instances representing prioritisation in patients with multiple chronic diseases. Comput Methods Programs Biomed. 2020;185:105151.

M. PIDD, 1996. Five simple principles of modelling. *- Proceedings Winter Simulation Conference.* pp.721-728

QUINLAN, J.R., 1995. MDL and categorical theories (continued). *Machine Learning Proceedings 1995.* Elsevier, pp.464-470

QUINLAN, J.R., 2014. *C4. 5: programs for machine learning.* Elsevier

QUINLAN, J.R., 1986. Induction of decision trees. *Machine Learning,* 1(1), 81-106

ROSENBLAD, A., 2011. *The Concise Encyclopedia of Statistics.* Taylor & Francis, pp.867-868Available from: https://www.tandfonline.com/doi/abs/10.1080/02664760903075614

Roth GA, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736–88.

S. ISLAM, N. JAHAN and M. E. KHATUN, 2020. Cardiovascular Disease Forecast using Machine Learning Paradigms. - 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). pp.487-490

S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning," From 4eory to Algorithms, Cambridge University Press, Cambridge, UK, 2020.

S. PACHANGE, B. JOGLEKAR and P. KULKARNI, 2015. An ensemble classifier approach for disease diagnosis using Random Forest. *- 2015 Annual IEEE India Conference (INDICON).* pp.1-5

S. PAWAR et al., 2021. Detection of Breast Cancer using Machine Learning Classifier. - 2021 Asian Conference on Innovation in Technology (ASIANCON). pp.1-5

S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease", In 2017 IEEE symposium on computers and communications (ISCC)*, pp. 204-207.*

SALZBERG, S.L., 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning,* 16(3), 235-240

SARANYA, C. and G. MANIKANDAN, 2013. A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET),* 5(3), 2701-2704

SCHNEIDER, P., F. XHAFA and P. LINKE, 2002. *Engineering Privacy-Preserving Machine Learning Protocols.* ACM

SELÇUK, A.A., 2019. A guide for systematic reviews: PRISMA. *Turkish archives of otorhinolaryngology,* 57(1), 57

SURI, J.S. et al., 2022. A Powerful Paradigm for Cardiovascular Risk Stratification Using Multiclass, Multi-Label, and Ensemble-Based Machine Learning Paradigms: A Narrative Review. Diagnostics (Basel), 12(3), 722

TANG, X., L. ZHANG and X. DING, 2019. SAR image despeckling with a multilayer perceptron neural network. *null,* 12(3), 354-374

TING, J., S. VIJAYAKUMAR and S. SCHAAL, 2010. Locally Weighted Regression for Control. In: C. SAMMUT and G. I. WEBB, eds. *Encyclopedia of Machine Learning.* Boston, MA: Springer US, pp.613-624

UDDIN, M.N. and R.K. HALDER, 2021. An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach. Informatics in Medicine Unlocked, 24, 100584

VAN DER MALSBURG, C., 1986. Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. In: G. PALM and A. AERTSEN, eds. *Brain.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp.245-248

VUJOVIC, ŽĐ, 2021. Classification Model Evaluation Metrics. *International journal of advanced computer science & applications,* 12(6),

WARD, S.C., 1989. Arguments for Constructively Simple Models. *The Journal of the Operational Research Society,* 40(2), 141-153

World Health Organization. Global action plan for the prevention and control of non-communicable diseases 214–2020. ISBN 978 92 4 150623 6. Geneva 2013; 2013.

World Health Organization, Cardiovascular Diseases, WHO, Geneva, Switzerland, 2021, https://www.who.int/healthtopics/cardiovascular-diseases/#t

Wu, Ching-seh Mike, Mustafa Badshah and Vishwa Bhagwat, "Heart Disease Prediction Using Data Mining Techniques", *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 2019, pp. 7-11.

7 Appendices

7.1 Appendix A: Ethics Approval

Ethical clearance for research and innovation projects

Project status

Status

Approved

Actions

Date	Who	Action	Comments
21:08:00 24 June 2022	Femi Isiaq	Supervisor approved	Take note that you will need to adhere to the use of secondary data as you have mentioned for the research work. A new ethical application will be required should there be a change in the research process particularly, data collection process.
18:53:00 24 June 2022	Peter Odesola	Principal investigator submitted	
17:07:00 24 June 2022	Peter Odesola	Principal investigator saved	

Get Help

7.2 Appendix B: GitHub repository

Follow the link to appendices and artefacts codes:

https://github.com/peterodesola/Heart-disease-classification.git