Prediction of Coronary Heart Diseases Using Supervised Machine Learning Algorithms

by

Nazia Binte Salam
18301156
Samiha Raisa
18301150
Rahela Atia Rashid
18301080
Asmita Noor
19101640
Sin-Sumbil Binte Obaed
18301092

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2022

Declaration

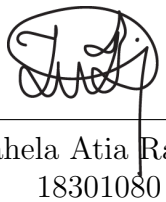
It is hereby declared that

1. The thesis submitted is my/our own original work while completing a degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nazia Binte Salam
18301156

Samiha Raisa
18301150



Rahela Atia Rashid
18301080

Asmita Noor
19101640

Sin-Sumbil Binte Obaed
18301092

Approval

The thesis/project titled "Prediction of Coronary Heart Diseases using Supervised Machine Learning Algorithms" submitted by

1. Nazia Binte Salam - 18301156
2. Samiha Raisa - 18301150
3. Rahela Atia Rashid - 18301080
4. Asmita Noor - 19101640
5. Sin-Sumbil Binte Obaed - 18301092

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 26, 2022.

Examining Committee:

Supervisor:
(Member)

Najeefa Nikhat Choudhury
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Ahanaf Hassan Hawlader
Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Cardiovascular disease is a leading cause of death worldwide. According to the Centers for Disease Control and Prevention, one person dies from heart disease every 36 seconds in the United States. In 2019, an estimated 17.9 million people died from CVD worldwide. High blood pressure, an unhealthy diet, high cholesterol, diabetes, air pollution, obesity, tobacco use, kidney disease, physical inactivity, harmful alcohol use, and stress can all contribute to it. Family history, ethnic background, sex, and age are some other contributing factors to a person's risk of heart disease. This paper seeks to predict heart diseases using a dataset that has factors like age, sex, the number of cigarettes smoked, etc. This prediction will be done by analyzing different parameters like blood pressure, oxygen level, hemoglobin count, etc. which are the major deciding factors to measure heart risks. The research will use supervised Machine Learning (ML) algorithms such as decision tree (a classification algorithm that works on categorical as well as numerical data), K-Nearest Neighbor (K-NN), Random forest algorithm, etc. to provide an accurate prediction. After applying ML on medical data, the outcome will be used to conduct a comparative analysis to measure the efficiency of different ML algorithms in predicting cardiovascular diseases. Furthermore, the major objective of this research is to use the algorithms and process in Bangladeshi dataset and explore the result outcome and newer possibilities.

Dedication

We dedicate this research to all those who lost their loved ones due to coronary heart disease and not being able to detect it at a primary stage. This thesis is our least but genuine effort to make a contribution to progress regarding coronary heart disease prediction so that prevention can be done beforehand through the path of prediction.

Acknowledgement

Firstly, glory be to Allah, the most Merciful and Benevolent, who ensured that our thesis was completed on schedule without any major drawbacks.
We would like to offer our heartfelt appreciation to Najeefa Nikhat Chowdhury, our distinguished mentor and advisor, for her advice and assistance during our work. We would also like to thank our co-advisor and mentor Ahnaf Hassan Hawlader for her assistance. Finally, we would like to express our gratitude to our most deserving parents. We might not have been able to accomplish this work without their relentless support. We are approaching graduation as a result of their goodwill and efforts.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

Alpha

Beta

Epsilon

Mu

Sigma

ANN Artificial Neural Network

DNN Deep Neural Network

GDP Gross Domestic Product

KNN K-Nearest Neighbor

ML Machine Learning

$RMSE$ Root Mean Square Error

SVC Support Vector Classifier

Chapter 1

Introduction

Heart disease is a term used to describe a group of conditions that affects the cardiovascular system. It could be a vessel or muscle problem, or it could be a congenital disorder. The most common type of heart disease is coronary artery disease (CAD), which reduces blood flow to the heart and causes heart attacks. Heart disorders can be "silent," with no symptoms or warnings until a person suffers from a heart attack or heart failure [13]. A heart attack is the unforeseen severe coronary thrombosis, which may cause angina, blood flow obstruction throughout the body, and severe myocardial damage. Chest pain or discomfort, heartburn, shortness of breath, nausea, and other symptoms are common for heart attacks.

According to the World Health Organization (2007), for the last 10 years, one of the main causes of death throughout the world were heart diseases. Because heart diseases are associated with a wide range of symptoms, it is difficult to detect them early and before they advance to a serious stage. Researchers have been looking for various data mining techniques and ML Algorithms to develop a heart disease predicting system that can accurately predict the severity of the problem in a short period of time. Bangladesh is no different. CVD is one of the leading causes of death in our country [3]. Bangladesh has a long history of poverty and genetic diseases due to its status as a developing country. CVD contributes the most to mortality in our country, accounting for 17 percent of all non-communicable diseases [3]. According to the most recent WHO statistics, 118,287 people died from Coronary Heart Disease in Bangladesh in 2018, accounting for 15.23% of all fatalities [19]. In Bangladesh, it has been observed that most people ignore their heart conditions until it is too late, which motivated us to work in this sector in order to contribute to the preliminary stage of heart conditions, which is disease detection. We decided to focus our efforts in the future on detecting the disease in the first place.

For people of all races, genders and ethnicities in the United States, heart disease is one of the main reasons for death. Heart disease kills approximately 655,000 Americans every year. The main cause of these fatal deaths was that they were not diagnosed in the early stages; when they went to the doctor, it took too long to diagnose them, and sometimes only at the very end. So, the motivation comes from this, to develop a prediction system with higher accuracy in order to ensure seeking medical help at an earlier stage.

1.1 Motivation

Coronary Heart Disease (CHD) is the main cause of death worldwide. Smart watches or fitness bands can now track our heart rates and calorie burn. These devices assist us in keeping track of our daily activities and guiding us in making healthy lifestyle decisions. Despite these advances in the health sector, some people are still not aware of the symptoms and risks that come with chronic diseases. As a result, forecasting such diseases is critical for humanity. [1]

CHD prediction models will help reach a wider user base, as the doctors can use this model with readily-available feature parameters that are mostly available to them, saving valuable time. Since CHD tests are costly, such a prediction model will help reduce the cost and give the patient/doctor primary information about the coronary heart condition of a certain individual without going for the expenses of medical tests in the first hand. Our model would be affordable and more accessible for patients of all socio-economic backgrounds which is necessary for a country like Bangladesh. Since the doctor to patient ratio is quite low, having a preliminary prediction of possible heart diseases will allow doctors to attend to the most vulnerable patients. As most people are not willing to spend much on tests due to negligence, ignorance, and affordability unless a doctor strictly prescribes those tests.

1.2 What is CHD?

One of the common types of heart disease in Bangladesh is CHD. Heart disease includes many types of conditions that affect the structure and function of the heart. CHD is a type of heart disease that occurs when the arteries of the heart are unable to supply the heart with enough oxygen-rich blood [18]. CHD is frequently caused by plaque, a waxy material that accumulates inside the lining of the larger coronary arteries. This deposit can restrict blood flow in the heart's main arteries partially or entirely. A disease or injury that changes the way the arteries in the heart operate might cause some kinds of this condition. Even though many people have the same type of CHD, the symptoms can vary from person to person. However, some of them may have no symptoms and are unaware they have CHD until they experience chest pain, a heart attack, or sudden cardiac arrest. Depending on the patient's condition, doctors may recommend healthy lifestyle changes, medications, or surgery to treat CHD.

In Bangladesh, there is a huge number of people who are suffering from this particular heart disease. According to the most recent WHO statistics, 118,287 people died from CHD in Bangladesh in 2018, accounting for 15.23 percent of all fatalities. Bangladesh ranks 115 in the world with an age-adjusted death rate of 109.32 per 100,000 population [19]. Thus, it can be said that CHD is one of the leading causes of mortality in Bangladesh.

1.3 Problems of Diagnosing CHD

CHD is typically diagnosed following a risk assessment and additional tests. These tests are costly as it includes electrocardiogram (ECG), exercise stress tests, X-rays, echocardiogram, blood tests, coronary angiography, radionuclide tests, MRI scans, CT scans etc [4]. All these tests are time consuming and many people might not be able to afford the tests. Also the patient needs to bring lifestyle changes and follow certain medical procedures, which many patients might find difficult to adapt to.

Furthermore, according to a McKinsey survey [24], chronic diseases are the leading cause of death in most countries, including China, and the US government spends approximately 2.7 trillion USD per year on the treatment of chronic heart diseases. A patient's life expectancy decreases after diagnosis. It causes stress, anxiety, depression, panic attacks and other mental disorders upon patients and their families. Hence, if an individual can predict the risk of chronic heart disease at an earlier stage then there is a lower chance for a patient to go through a diagnosis.

1.4 Aims and Objectives

Our goal is to develop a prediction system applying supervised ML algorithms to detect heart diseases with higher accuracy within a short period of time. At present most of the research is done on either Regression or Classification. In our paper, we aim to do Classification to see if we can improve the time complexity and accuracy. For the process, we will analyse the presence of CHD with Classification by using a Supervised ML algorithm. This classification will detect the presence of heart disease by analysing some features that contribute to disease detection, yielding a numerical output of either 0 or 1, where 0 indicates the absence of heart disease and 1 indicates the presence of heart disease.

1. The utmost goal of this research project is to predict a patient's heart disease using ML algorithms that has more accurate and efficient work process. The objective is to reach higher accuracy after surfing through different algorithms and designating the algorithm with the highest accuracy percentage.

2. The next objective is to categorise and to pre-process the datasets depending on various parameters and make it convenient for the research. The pre-processing includes representing the data in tabular form, removing null values and selecting only target columns for data modelling. Since most models assume that the data is numerical and has no missing values or variables, they need to be handled before putting them into the algorithm. Some data requires specific type format such as the height and weight data must be in float for the model to perform successfully. The dataset is splitted into training data and test data using a cross validation technique suitable for the model. With a filtered data set explored, an appropriate column should be decided to use as a target column for data modelling.

3. One of the primary goals of this study is to detect the presence of CHD using classification, which will aid in providing a binary value of the output.

4. Lastly, different ML algorithms will be applied to the model to indicate the appearance of heart disease and compare their output accuracy.

1.5 Research Problem

In underdeveloped and developing countries, the facilities of diagnosis are very expensive and unavailable. Socioeconomic changes, increased life expectancy, and an unhealthy lifestyle are all major risk factors contributing to the rapid rise in CHD in the majority of low- and middle-income countries [9]. CHD eliminated 7.3 million lives and 58 million disabilities globally in 2001. Below-average and average income nations accounted for 3 quarters of fatalities and 82% of total Disability Adjusted Life Years (DALYs) all over the world due to CHD. [10]. In 2019, an estimated 17.9 million people died from CVDs, accounting for 32% of all global deaths. Heart attacks and strokes were responsible for 85 percent of these deaths [11].

High blood pressure, high blood cholesterol, and smoking are key risk factors for heart diseases. Heart diseases can also be largely caused by various medical problems and lifestyle problems including diabetes, inactiveness, imbalanced diet, obesity, drug and alcohol abuse, etc. Study shows that some non-smokers, when exposed to Environmental Tobacco Smoke (ETS), have an increased risk of 20%, but this is still controversial [1]. Some researchers also emphasise the fact that lack of sleep can cause heart problems. A study on 43 candidates shows that both short sleep ($<$ 7 hours per night) and long sleep ($>$ 8 hours per night) were associated with a greater risk of all-cause mortality [1]. Our diet has a significant role in developing a healthy heart. As a result, when we do not intake proper food or nutrition apart from our body, our heart is also affected. According to a 2018 report published by the United Nations Children's Fund (UNICEF), malnutrition has an impact on cardiovascular health in both childhood and adulthood, and can lead to coronary artery disease, hypertension, and diabetes mellitus. As a result, a healthy diet appears to have a key role in the prevention of CHD and other factors like obesity, hypertension, and diabetes. Apart from a diet, drug abuse has also been proven to have negative consequences

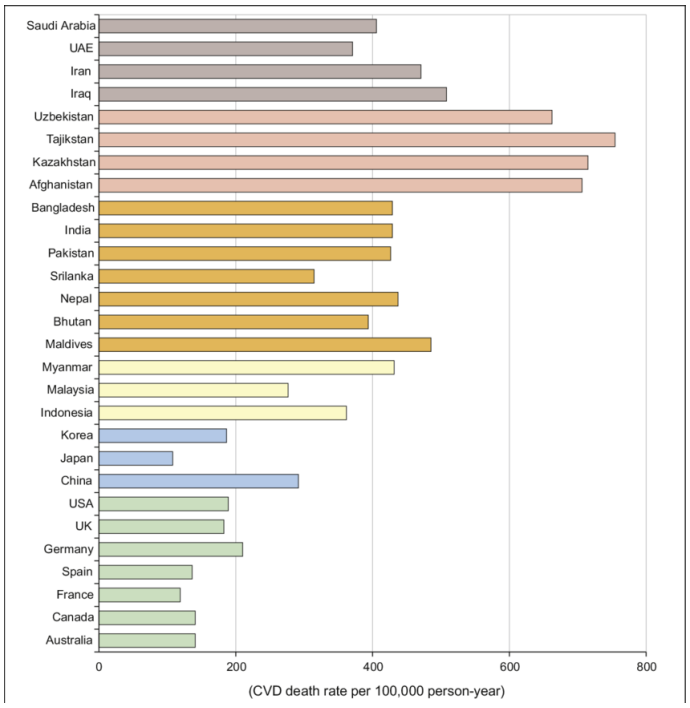


Figure 1.1: CVD statistics for different countries in the year 2002.

The graph above depicts various areas, each represented by a different colored bar. Here, purple represents Middle Eastern countries, pink Central Asian countries, orange South Asian countries, yellow South-East Asian countries, blue East Asian countries, and green Non-Asian countries. UAE is an abbreviation for the United Arab Emirates, USA is an abbreviation for the United States of America, and UK is an abbreviation for the United South Asian countries that have a higher rate of CVD death than other Asian and Western countries.

Chapter 2

Literature Review

2.1 Overview

Using data mining techniques to predict CHD has been a continuous effort for the past two decades. Most papers choose methods such as Decision Tree, Neural Network, Naive Bayes, kernel density, automatically defined groups, bagging algorithm, and SVM in the diagnosis of heart disease, which result in varying levels of accuracy. Furthermore, the outcome differs for different databases that include different parameters of heart disease attributes. In today's world, CHD has become one of the most concerning factors. So it goes without saying that predicting heart disease will be one of the most discussed and desired research topics among scientists and engineers all over the world.

2.2 Related Works

According to a study [12] conducted in India, data mining techniques were used, as were numerous variations of mining algorithms, and several potential paths were used to predict the detection of heart diseases. In this paper, they used four algorithms (Decision Table, Naive Bayes, SMO, Lazy K Star) in the dataset from UCI Library. After filtering the data and removing missing values, the algorithms were applied to achieve the highest accuracy. They also applied the Accuracy, Precision, and Recall metrics to compare the different results generated by the algorithms. Among the ML algorithms, the Naive Bayesian algorithm gives the most accurate result (231 out of 250 instances). On the other hand, the least accurate is the Lazy K Star (211 out of 250 instances).

The datasets used in this paper are classified in terms of medical parameters, and data mining classification techniques are used in another research paper [5] on "Prediction of Heart Disease Using ML Algorithms," which was also conducted in India. This paper implemented two supervised data mining algorithms, the Decision Tree Algorithm and the Naive Bayes Algorithm. To determine the best accuracy, these two algorithms were applied to the same dataset. In this paper, the DTC had an accuracy of 91% and the Naive Bayes classifier had an accuracy of 87%. As a result, the DTC had the highest accuracy for heart disease, at 91 percent.

A research paper proposes [7] a comparison between different supervised ML algorithms for disease prediction in general. This paper used two databases Scopus and PubMed. The paper includes the composition of ML and data mining procedures. For the final dataset, they nominated 48 selected articles in terms of the methods used, performance measures as well as the disease they targeted. The articles predicted a total of 49 diseases using ML algorithms to provide a higher acceptance rate. Artificial Neural Network (ANN), Decision Tree (DT), KNN, Linear Regression (LR), Naive Bayes (NB), Random Forest (RF), SVM algorithms were applied to find superior accuracy. Interestingly, SVM has been found the last time (only once) to give the superior result. On the contrary, SVM showed superior accuracy at most times for heart disease, which is our working sector, diabetes, and Parkinson's disease. The results showed that RF has the highest accuracy, while the SVM algorithm is used the most (in 29 studies), followed by the Naive Bayes algorithm (in 23 studies).

Deep learning was used to predict heart disease, according to the research paper [9]. In this study, a set of features derived directly from cardiac sounds was used as the input for a deep neural network to determine whether a cardiac sound belongs to a healthy person or a patient with cardiac disease. The dataset for the study was collected using a smartphone and the iStethoscope Pro mobile app. The iStethoscope Pro iPhone app4 is used to collect heartbeat sounds. The app takes advantage of modern mobile devices' audio capabilities, performing real-time filtering and amplification and allowing users to view Fast Fourier transform (FFT) spectrograms and email eight seconds of audio. The feature vector, design model, and study design structured the prediction procedure. The result was concluded with an accuracy of 0.98.

In the research paper [12] on "Improved Heart Disease Projection System Using Data-Mining Classification Techniques", the researchers used neural networks, decision trees, and Naive Bayes algorithms. After carrying out the research they found out that from the three, the neural network model predicts the most reliable heart attack, for predicting heart level risk using "Effective cardiac prediction system using data mining techniques." A predictive system is made using the neural network, and the readings showed that the test device developed would correctly forecast the chances of a heart attack. They carried out this research on 533 people who were suffering from cardiac disease.

A research paper [6] used the Hybrid RF with Linear Model (HRFLM). It is a proposed method that combines RF and Linear Method features (LM). Several feature combinations and well-known classification techniques were used to introduce the prediction model. By combining the hybrid RF with a linear model in the prediction model for heart disease, they achieved an improved performance level with an accuracy level of 88.7 percent HRFLM. To find heart disease risk factors on the UCI Cleveland dataset, HRFLM used a computational approach with three mining association rules, namely apriori, predictive, and Tertius. HRFLM used ANN with backpropagation as the input, along with 13 clinical features. The findings then were compared to those obtained using traditional methods. HRFLM was found to be quite effective in predicting heart disease.

Decision Tree classifier, Naive Bayes, Logistic Regression, Random forest, SVM and KNN models have been used in a research paper [10]. In that research, these classification techniques were applied on a cardiovascular dataset to predict the possibility of a heart disease. Among the models, the Decision Tree Classifier was shown to have the highest accuracy score of 73%. The other models, specifically Random Forest and KNN, did not have good accuracy scores. A possible reason for this could be the reduced number of dimensions in the dataset. Higher numbers of dimensions generally provide higher levels of accuracy.

DTC, Naive Bayes, Logistic Regression, RF, SVM and KNN models have been used in a research paper [10]. In that research, these classification techniques were applied on a cardiovascular dataset to predict the possibility of a heart disease. Among the models, the DTC was shown to have the highest accuracy score of 73%. The other models, specifically RF and KNN, did not have good accuracy scores. A possible reason for this could be the reduced number of dimensions in the dataset. Higher numbers of dimensions generally provide higher levels of accuracy.

Shashikant et al conducted research where they used the Heart Rate Variability (HRV) to predict the occurrence of heart disease in smokers [11]. HRV is a metric to identify the variance in time of the beating of a person's heart. For this research, a controlled environment was set up, to ensure a higher precision of the recorded time data. Using the 10-fold validation method to measure the performance of the system, they used Decision Tree, Logistic Regression and Random Forest as the classifiers. Decision Tree gave an accuracy of 92.59%, Logistic Regression had an accuracy of 89.7% and finally Random Forest had the highest accuracy score of 93.61%.

Shashikant et al conducted research where they used the Heart Rate Variability (HRV) to predict the occurrence of heart disease in smokers [11]. HRV is a metric to identify the variance in time of the beating of a person's heart. For this research, a controlled environment was set up, to ensure a higher precision of the recorded time data. Using the 10-fold validation method to measure the performance of the system, they used Decision Tree, Logistic Regression and RF as the classifiers. Decision Tree gave an accuracy of 92.59%, Logistic Regression had an accuracy of 89.7% and finally RF had the highest accuracy score of 93.61%

A SVM with radial basis function (SVM), L1-penalised logistic regression (LR), RF, decision tree (DT), Gaussian naive Bayes (NB), KNN, and extreme gradient boosting (XGB) was used in another study [15] to predict CVD. Model performance in the test set was evaluated in terms of discrimination, calibration, and clinical usefulness. RF was used to determine the variable importance of included variables. For model development, twenty-two variables were used, including medical history, sociodemographic characteristics, cytokines, and synthetic indices All seven models performed well in terms of discriminating (AUC varied from 0.770 to 0.872). The AUCs for LR and SVM were 0.872 (95 percent CI: 0.829– 0.907) and 0.868 (95 percent CI: 0.825– 0.904), respectively. LR had the most sensitivity and the lowest Brier score (0.078). (97.1 percent). SVM was shown to be somewhat superior to LR in decision curve analysis.

To predict patient survival in this paper, Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques [14], nine classification models are used: Decision Tree, Adaptive boosting classifier, Logistic Regression, stochastic Gradient classifier, RF, Gradient Boosting classifier, Extra Tree Classifier, Gaussian Naive Bayes classifier, and Support Vector Machine. The SMOTE is used to create balance in the dataset. Thus, ML models are trained using RF. The outcome is compared with the ML algorithms using all the features. It is found that the Extra Tree Classifier achieves the highest accuracy 0.9262 value with SMOTE in prediction of heart patient's survival.

Chapter 3

Research Methodology

The purpose of this study is to see if ML algorithms such as KNN, RF, Decision Tree, Naive Bayes, and Binary Logistic Regression (BLR) Model may be used to achieve higher accuracy and reduced error rates.

The "Cardiovascular Disease dataset," which contains 70000 records of patient data, 11 features, and targets obtained at the time of medical examination, will be used in this study. Svetlana Ulianova, a data science student from Toronto, has made this data set public via a web-based data-science environment [17]. Using the dataset as input and following the methods mentioned below, it is critical to make accurate predictions:

1. **Input data:** The dataset representing features for detecting heart disease is organised and formatted so that it can be input into the system to be preprocessed.
2. **Preprocessing:** The dataset has been classified based on numerous parameters. Before proceeding with the dataset, binary encoding and feature engineering must be used to preprocess the data.
3. **Dividing Datasets:** We used cross validation technique for dividing the dataset.
4. **Apply Model:** After the dataset has been separated into a training and testing set, BLR Model has to be applied. KNN, RF, Naive Bayes and Decision Tree must then be used to compare them with other models to find which gives a greater accuracy.
5. **Decision Function:** Based on the accuracy and standard deviation of the models employed, it can be determined which model is best to use to predict the presence of heart disease.

3.1 Machine Learning Models

A ML model is a declaration of a formula that searches through large amounts of data to find instances and generate predictions. ML has the ability to access data

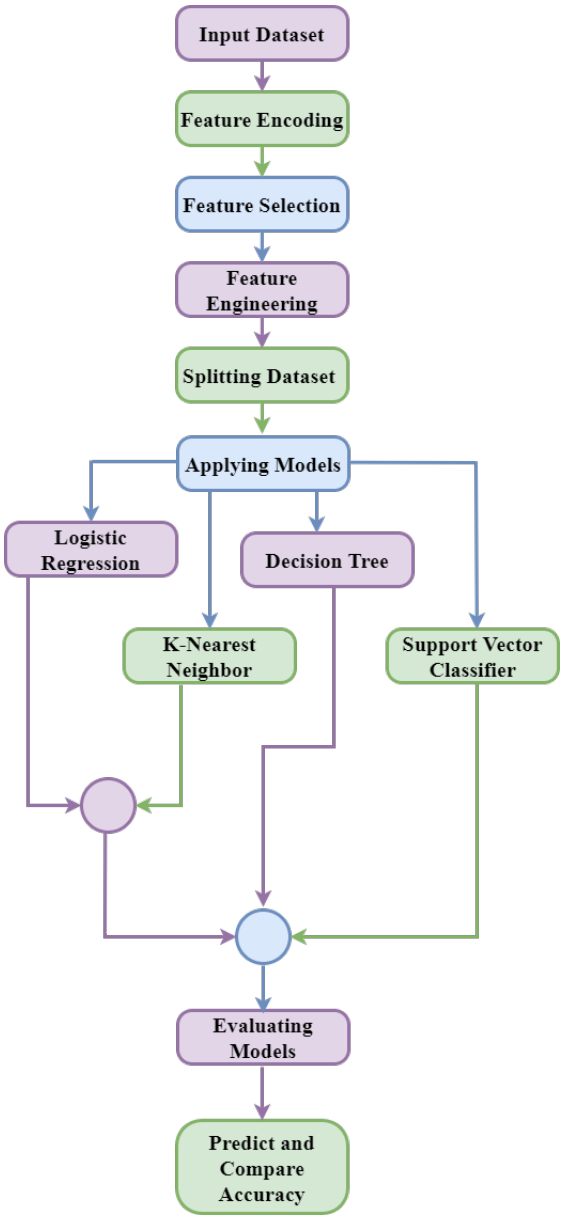


Figure 3.1: The proposed CHD model in flowchart.

and utilise it to create computer programs.Without human assistance, the computer learns from observations and can make better future predictions.Computers can grasp the meaning of a document using ML techniques, just like humans.

There are mainly four types of algorithms, namely supervised, semi-supervised, unsupervised and reinforcement. Supervised ML uses past data to predict future possibilities. By analysing the known training dataset, the learning algorithm produces a result to make predictions about the output values. After training is complete, the system will be able to produce possible results based on the training data. The model is also able to compare its actual output with the expected results to find any sort of discrepancies and make the necessary adjustments for better future predictions.

- **Cross Validation:** Cross-validation is a technique that we use to train our model using the subset of the dataset and then estimate using the complementary subset of the dataset. We use Cross Validation, in order to get an expected accuracy score with minimum and maximum range.
- **Confusion Matrix:** A confusion matrix is a summary of predicted results on a classification case. If we look at the table we can have a better understanding of the confusion matrix.

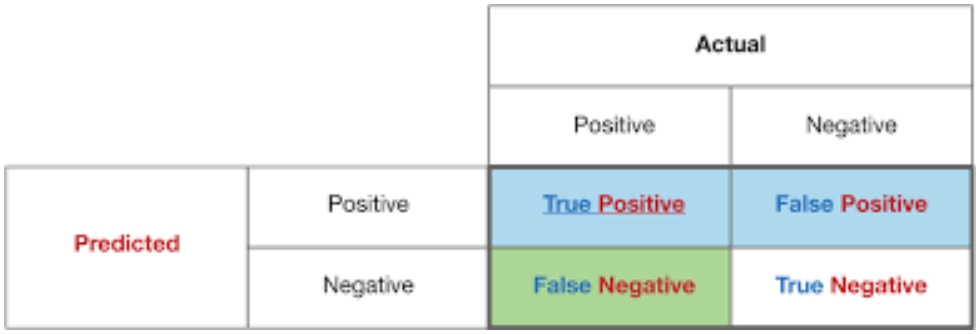


Figure 3.2: Confusion Matrix. [25]

In this case:

- **True Positive:** Cases Predicted with CHD and actually had CHD
- **True Negative:** Cases Predicted with no CHD and actually had no CHD
- **False Positive:** Cases Predicted with CHD and actually had no CHD
- **False Negative:** Cases Predicted with no CHD and actually had CHD

Here, for predicting the values over the training set, SVM Algorithm, KNN, DTC, Naive Bayes Algorithm, RF Algorithm and BLR have been used.

3.1.1 K-Nearest Neighbors (KNN)

KNN is a ML technique that solves classification and regression prediction issues using a supervised ML method.For large datasets, the KNN algorithm is effective.This algorithm saves all available data and categorises new data using the group that is the most homogenous to the new data. As a result, when new data is added, it can readily match the training set's point similarity.[21] The K-NN technique is commonly used for classification tasks.Using the KNN method, new data can be quickly categorised into a clearly stated category.
The K-NN algorithm is also known as a non-parametric algorithm. This means that the data isn't supposed to originate from pre-programmed models with a restricted set of variables. This algorithm is also known as a lazy learner because it steadily assimilates from the testing datasets rather it keeps the dataset and utilises it to classify. The KNN technique just stores the knowledge throughout the learning phase, and when it gets new data, it classifies it into a group that is relatively close

to the new data.

K-NN first determines the size of the neighbours which is represented by K. Then, the Euclidean distance is computed between each of the K neighbours. After that, the Euclidean distance is used to determine the K closest neighbours. Afterward, the number of data (K neighbour) in each class is counted and the class with the maximum number of data is considered as the preferred class so when new data comes it is set into that class.

One of the benefits of using K-NN is that it is simple to implement.It can endure a large amount of erratic learning dataset. It may be more successful if the learning dataset is large.There are several distance functions that can be used to calculate the distance, but Euclidean is the most generally used.

$$d(x;y) = \sqrt{\sum_{i=1}^{x} (x_i \quad y_i)^2}$$

3.1.2 Decision Tree Classi er (DTC)

Decision Trees are a type of Supervised ML, where the data is constantly broken down into its individual parameter. We can understand the structure of the tree by decision nodes and leaves. To get from observations about an item (presented in the branches) to inferences about the item's desired value, use a decision tree (as a predictive model) represented in the leaves. Decision Trees are of two types: Classification trees - where the outcome is either, Yes or No types. Here the decision variable is Categorical. And the other one is Regression trees - where the outcome variable is uninterrupted.

The main reason to use DTC is that it is flexible because it is no-linear. The second reason is that decision trees successfully communicate advanced processes. Another reason to use this algorithm is that the decision tree focuses more on probability and facts rather than sentiments. Apart from these, decision trees are capable of mapping different possibilities and ultimately determine which course of action has the highest likelihood of success. Among the most fundamental features of decision trees is that they can construct a series of splits, which divides the data into two purest groups. Decision trees compute the entropies of groups in order to calculate their purity. A decision tree with C classes has the following entropy:
Entropy:

$$\sum_{i} p_i \log_2 p_i$$

pi denotes the probability of selecting an element of class i at random. The entropy values vary from 0 to 1, with 1 representing maximum impure groupings and

0 representing entire pure groups. Information gain, or a decrease in entropy, is another statistical feature of decision trees. It calculates the difference in entropy between the dataset before and after splitting based on feature values. Information gain:

$$Entropy(before) \sum_{j=1}^{k} Entropy(j; after)$$

"Before" refers to the dataset prior to the division, "k" to the degree of subsets produced by the split, and "j, after" to subset j following the split.

3.1.3 Naive Bayes Algorithm

It is an algorithm based on the theorem of Naive Bayes that aims to decipher classification related issues. This algorithm uses Naive Bayes Classifier to assume that features are not codependent. Also this one is well known for its fast working capabilities and simple approach and it is also very well-suited for large datasets. The Gaussian probability density function can contribute to making predictions by replacing the parameters with the new input value of the variable and as a result, the Gaussian function will provide a new probability as an input. The Naive Bayes Classifier works on independent variables. It does not depend on another feature for output prediction. This algorithm works on training data to presume the parameters. For real-life scenarios, for simple design and application, this classifier is used.

According to research [2], R.Patil(2014) developed a heart disease prediction system using this algorithm along with Jelinek-mercer smoothing. In this research Naive Bayes was favoured as the data was huge, attributes had fewer similarities and a more accurate result was expected.

3.1.4 Random Forest Algorithm

RF, like its name inferred, has a huge number of independent decision trees. It can be implemented both on Classification and Regression models. RF is a concept of ensemble learning that ensures better predictive performance which is a method of merging several classifiers to tackle a complicated problem and to improve the performance of the model. It helps to avoid the problem of overfitting and the increased number of trees ensure a higher rate of accuracy. It's a comparatively faster algorithm.[23]

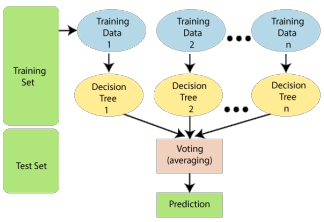


Figure 3.3: Working of RF Algorithm.[23]

3.1.5 Binary Logistic Regression

A predictive approach called logistic regression is used to analyse data and determine the connection between a dependent binary variable and one or more independent variables. This statistical analysis method uses dependent variables when the outcome will be either 1 or 0, this regression technique is close to linear regression and also used to predict the Probabilities for classification problems. The equation of the best fit line for linear regression is:

$$Y = {}_0 + {}_1 X +$$

0 is the y-intercept of the regression line, $_1$ is the gradient of the regression line, X is the explanatory variable, is the error of the model and Y is the dependent variable. For better classification, the output values of the regression are passed through a sigmoid function. The sigmoid function takes this input and displays the probability values of each output value. The logistic regression model, on the other hand, is better for limiting the range of data for the dependant variable. A graphical comparison between linear regression and logistic regression is shown below:

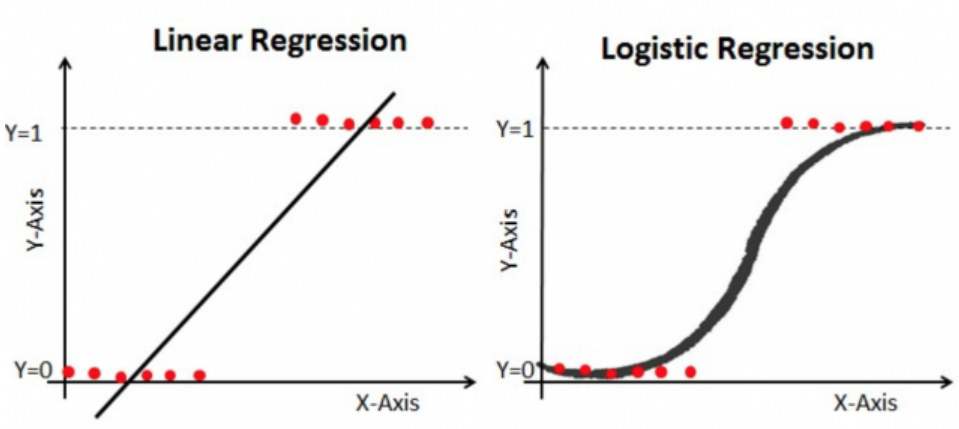


Figure 3.4: Contrasting linear to logistic regression.[22]

The following is the derivation of the logistic regression model's basic equation: where Y is substituted by a probability P, the probability range is considered to be between 1 and 0, and the odds of P are determined.

$$\text{or, } \frac{p}{1 \quad p} = \frac{{}_0 + \ {}_1 X}{{}_0 + \ {}_1 X}$$

$$\text{or, } \exp[\log \frac{p}{1 \quad p}] = \exp\,[\frac{{}_0 + \ {}_1 X}{{}_0 + \ {}_1 X}]$$

$$\text{or, } e^{\ln \frac{p}{1-p}} = e^{(\ {}_0 + \ {}_1 X)}$$

$$\text{or, } P = e^{(\ {}_0 + \ {}_1 X)} \quad P e^{(\ {}_0 + \ {}_1 X)}$$

$$\text{or, } P + P e^{(\ {}_0 + \ {}_1 X)} = e^{(\ {}_0 + \ {}_1 X)}$$

$$\text{or, } P[1 + e^{(\ {}_0 + \ {}_1 X)}] = e^{(\ {}_0 + \ {}_1 X)}$$

$$\text{or, } P = \frac{e^{(\ {}_0 + \ {}_1 X)}}{1 + e^{(\ {}_0 + \ {}_1 X)}}$$

This is the sigmoid function for the logistic regression model, to predict any binary dependent variable [16].

3.1.6 Hybrid Algorithms

According to an article [26], A hybrid algorithm is one that is basically a combination of two or more different algorithms that solve the same issue. The program either selects one (based on the data) or switches between while the algorithm is being executed. In general, this is performed to reach the optimal result by merging favoured attributes of each algorithm. It gives a better output than the individual components. Hybrid Algorithms are not used to just combine a few algorithms but the primary objective is to achieve more accurate results. As it works with different algorithms with various features, it can optimise the outperformance. In our dataset we have used a hybrid RF Classifier algorithm combining Logistic Regression, K-Neighbors Classifier and DTC. RF is a popular and highly used classification method which can include different types of decision tree algorithms and provide an optimised result. We have combined multiple decision trees into a RF and used the variety of the trees' to improve our resulting model. We have put Logistic Regression, K-Neighbors Classifier and DTC so that we can combine the features of each unique method and get an accurate result. We chose these algorithms as among all the other algorithms Logistic Regression, K-Neighbors Classifier and DTC gives the best outperformance. Thus, we decided to merge the best three algorithms to optimise the best result. The precision rate before using the hybrid algorithm was 72.33%. After combining the individual algorithms as a hybrid algorithm we got 71.22% as result which clearly shows that the model did not improve the accuracy. Since the model could not provide the highest accuracy on the Canadian dataset we decided to work on our Bangladeshi dataset, by applying SMOTE we tried to improve the accuracy of the model.

3.2 SMOTE Technique

3.2.1 What is SMOTE Technique

Synthetic Minority Oversampling Technique or SMOTE is an oversampling method [8]. Which means it is used to solve imbalance problems. As the name suggests it takes a minority class and adds new examples to it. It keeps adding instances to the dataset until the quantity of the two classes are equal. The process's main goal is to balance class distribution by randomly increasing minority class examples through replication.

3.2.2 Application of SMOTE Technique

SMOTE works by selecting examples near the feature space, drawing a line in the feature space between the examples, and then drawing a new sample in the space corresponding to that line. The chosen class must be an instance of a minority class [8].

Chapter 4

Dataset

4.1 Implementation

4.1.1 Source

The "Cardiovascular Disease dataset" recorded by a student of The Ryerson University in 2019, analysed and gathered the insights of coronary diseases. [17].

4.1.2 Dataset Description

The data is categorized into three types of input features that are narrated in the table below:

FEATURE TYPE		
Objective (unit)	Examination (unit)	Subjective (unit)
Age (days)	Systolic Blood Pressure (ap_hi)	Smoking (binary)
Height (cm)	Diastolic Blood Pressure (ap_lo)	Alcohol Intake (binary)
Weight (kg)	Cholesterol (scaled by 1,2,3)	Physical Activity (binary)
Gender (categorical code)	Glucose (scaled by 1,2,3)	0 or 1 prediction for CHD (binary)

Table 4.1: Feature Type.

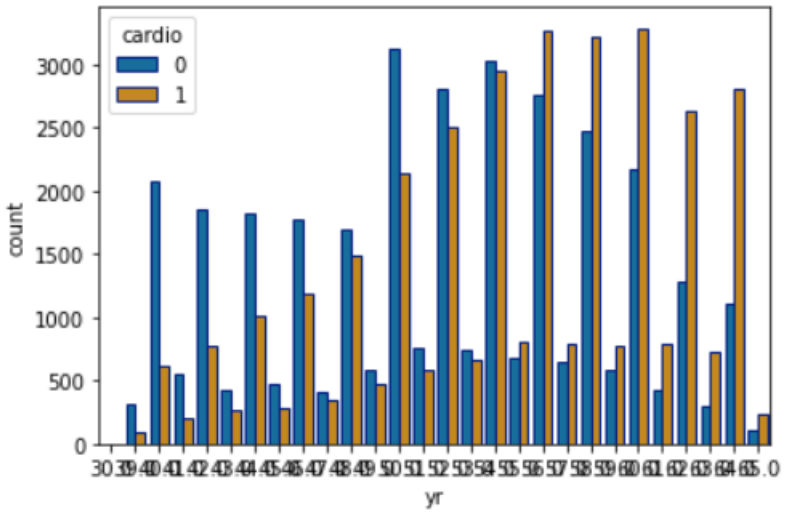


Figure 4.1: CHD risk graph based on age.

4.2 Preprocessing Dataset

4.2.1 Feature Observations

Feature observation implies the individual characteristics of the features and how the different parameters of such features affect our data results. For instance: the feature "Age" was transformed into years and after analysing , we observed that with the increasing age, the existence of cardio is increasing. And young people tend to have less possibility of heart diseases.

As per "Gender", male and female tend to have nearly the same risk of getting CHD. Here, male population is assigned by 1 and female by 2 here. As you can see the result varies in a closer margin . So, it can be said that CHD risks are quite high regardless of gender.

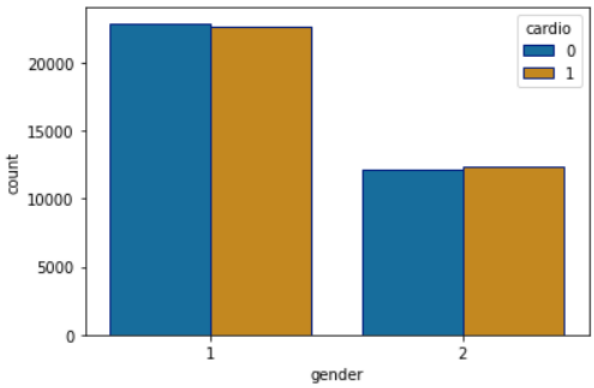


Figure 4.2: CHD risk graph based on gender.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

70000 rows × 13 columns

Figure 4.3: Raw dataset.

4.2.2 Feature Encoding

Since the majority of ML algorithms only accept numeric values,string values must be converted to binary values. Hence, the categorical values of the relevant attributes must be converted to numerical values, a process known as feature encoding. In our dataset, there was only one data type attribute present, and it was all numeric, so we didn't have to convert any data. However, in order to forecast the outcome, we must use Binary Encoding, where the existence of CHD is represented by 1 and the absence of CHD is displayed by 0.

4.2.3 Imputation

Imputation is the method of handling the missing value in the dataset ("null values") with some substitute values. Our dataset includes no missing values hence imputation was not implemented here and no substitute values were required.

```
age            0
gender         0
height         0
weight         0
ap_hi          0
ap_lo          0
cholesterol    0
gluc           0
smoke          0
alco           0
active         0
cardio         0
dtype: int64
```

Figure 4.4: Dataset without zero values

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	yr
age	1.000000	-0.022811	-0.081515	0.053684	0.020764	0.017647	0.154424	0.098703	-0.047633	-0.029723	-0.009927	0.238159	0.999090
gender	-0.022811	1.000000	0.499033	0.155406	0.006005	0.015254	-0.035821	-0.020491	0.338135	0.170966	0.005866	0.008109	-0.023017
height	-0.081515	0.499033	1.000000	0.290968	0.005488	0.006150	-0.050226	-0.018595	0.187989	0.094419	-0.006570	-0.010821	-0.081456
weight	0.053684	0.155406	0.290968	1.000000	0.030702	0.043710	0.141768	0.106857	0.067780	0.067113	-0.016867	0.181660	0.053661
ap_hi	0.020764	0.006005	0.005488	0.030702	1.000000	0.016086	0.023778	0.011841	-0.000922	0.001408	-0.000033	0.054475	0.020793
ap_lo	0.017647	0.015254	0.006150	0.043710	0.016086	1.000000	0.024019	0.010806	0.005186	0.010601	0.004780	0.065719	0.017754
cholesterol	0.154424	-0.035821	-0.050226	0.141768	0.023778	0.024019	1.000000	0.451578	0.010354	0.035760	0.009911	0.221147	0.154386
gluc	0.098703	-0.020491	-0.018595	0.106857	0.011841	0.010806	0.451578	1.000000	-0.004756	0.011246	-0.006770	0.089307	0.098596
smoke	-0.047633	0.338135	0.187989	0.067780	-0.000922	0.005186	0.010354	-0.004756	1.000000	0.340094	0.025858	-0.015486	-0.047884
alco	-0.029723	0.170966	0.094419	0.067113	0.001408	0.010601	0.035760	0.011246	0.340094	1.000000	0.025476	-0.007330	-0.029918
active	-0.009927	0.005866	-0.006570	-0.016867	-0.000033	0.004780	0.009911	-0.006770	0.025858	0.025476	1.000000	-0.035653	-0.009819
cardio	0.238159	0.008109	-0.010821	0.181660	0.054475	0.065719	0.221147	0.089307	-0.015486	-0.007330	-0.035653	1.000000	0.237749
yr	0.999090	-0.023017	-0.081456	0.053661	0.020793	0.017754	0.154386	0.098596	-0.047884	-0.029918	-0.009819	0.237749	1.000000

Figure 4.5: Dataset after feature selection.

4.2.4 Feature Selection

Feature selection in ML is a process which selects the significant features in the dataset that contributes most to the prediction variable. By reducing the number of input variables important for the prediction, it can improve the data accuracy, reduce the computational cost and train the model faster with the correlation between the features.
For this dataset , the id column was dropped, as it shows the index number only hence irrelevant to the prediction system. The correlation between the features are mentioned below:

4.2.5 Feature Engineering

Feature engineering is a procedure that enables us to analyze the sample data and specifically select the features that will make the ML model accurate. ML models like Linear and Logistic Regression, KNN, etc require the data to be scaled. So scaling beforehand is a measure requirement for such algorithms. On the other hand , tree-based algorithms, like DTC, is not affected by scaling, as its split on a feature does not change based on other features.

In order to make the dataset neutral towards all algorithms so that the results may not get any bias accuracy for any specific algorithm, Standard Scaler has been used. The data is scaled by centering them around the mean with a unit standard deviation. Method for standardisation can be defined as:

$$X = X \quad -$$

Where. = the mean of the feature values And, = the standard deviation of the feature values.
There are no restrictions on the range of the values. The dataset has been divided into a train and test set using cross validation. Then, the features have been scaled using the standard scaler.

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	yr
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729	0.499700	53.338686
std	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179	0.500003	6.765294
min	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	30.000000
25%	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000	48.000000
50%	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000	54.000000
75%	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000	58.000000
max	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000	65.000000

Figure 4.6: Feature selected for the data relations.

4.3 Programming Language

For this research, we have used Python language to test and implement our models. The models were tested with various computational calculations. For the large computational activities, Python along with Google Compute Engine Backend (16 GB RAM) has been used as the main programming language for the applied models as it can handle huge computations with required optimizations.

4.4 Application of the Models

For this research, we have used BLR, KNN and DTC , RF Algorithm, Naive Bayes Algorithm for predicting our data. The data has been split using cross validation techniques for training and testing the models. The classification has been done on 12 columns. Firstly, the models have been implemented on the whole dataset which consists of data from the "Cardiovascular Disease dataset" . Later, the same models have been implemented with a shorter data quantity of Bangladeshi Dataset and compare with the previous implementation.

Chapter 5

Comparative Feature Analysis with Bangladeshi Dataset

After the completion of implementing our model in the Canadian dataset, we approached the same model for CHD prediction in the Bangladeshi dataset that we collected from IPDI Foundation [20] and the comparison results regarding features like age, gender varied on a huge scale.
As per analysis on BD dataset, the feature "Age" was calculated in years and after analyzing, we observed that the presence of cardio is increasing and the highest around middle- aged people whereas it was highest for senior citizens in our first dataset.

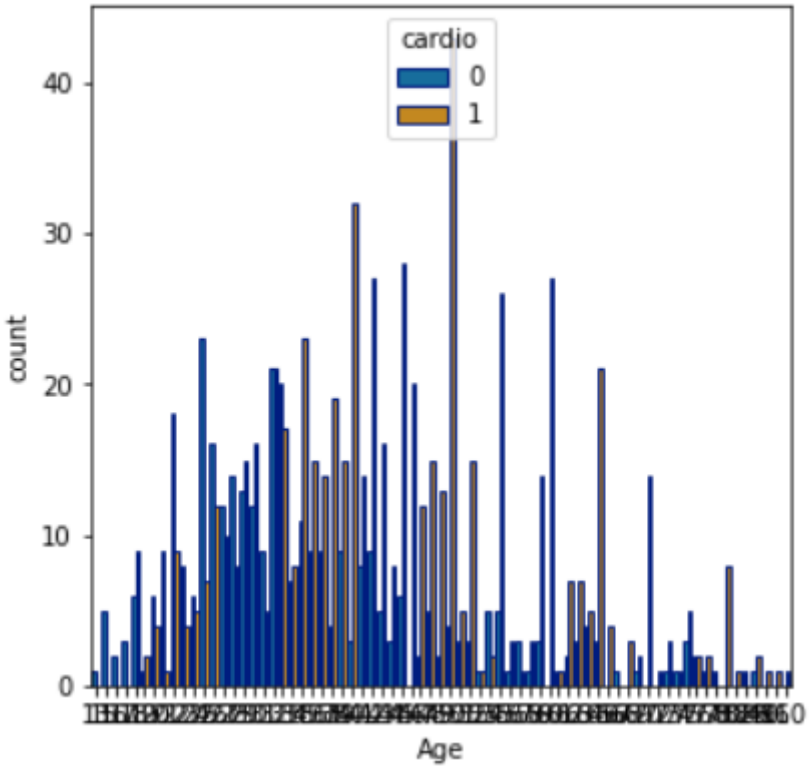


Figure 5.1: CHD risk graph based on age.

Secondly, analyzing the feature "Sex" in BD dataset, we can reach to the decision that Bangladeshi male has greater risk when it comes to CHD. On the contrary, in the Canadian dataset we observed that the result is not that affected by gender difference.

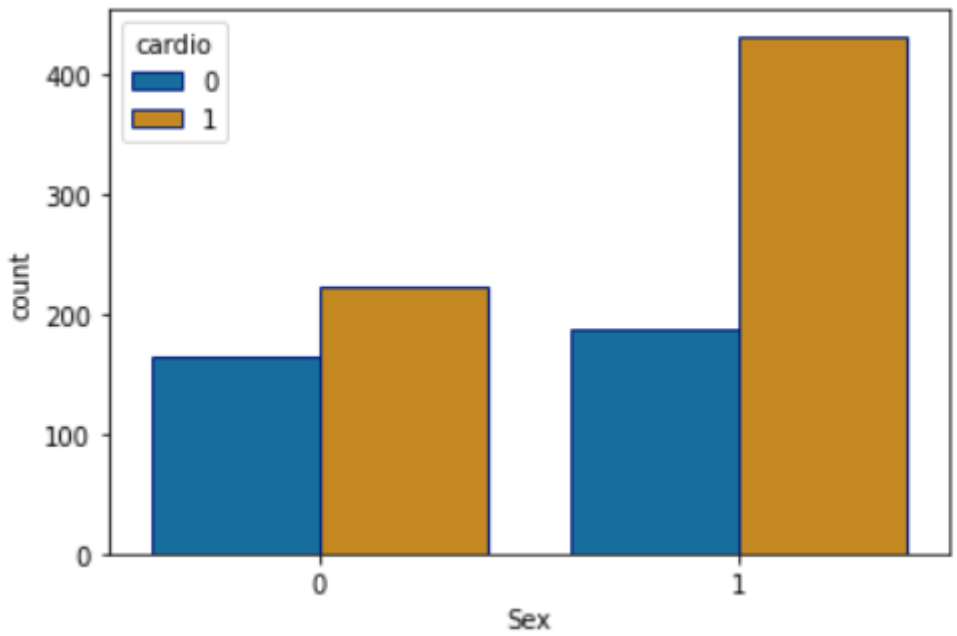


Figure 5.2: CHD risk graph based on gender.

Thus different results have been detected for the same features in our country dataset which will help us to work specifically on our model system for Bangladesh that we intend to develop in further research as this will be our further research perspective.

Chapter 6

Results and Discussion

6.1 Classification Result

For the prediction, five different types of classifiers have been used. The classifiers are - K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Naive Bayes Algorithms, Random Forest Algorithm and Binary Logistic Regression.

From the table, Binary Logistic Regression has the highest accuracy rate of 72.33% with a precision and recall score of 0.74 and 0.67 respectively.

After that, Random Forest Algorithm has an accuracy of 71.47% which is the highest after Binary Logistic Regression with a precision and recall score of 0.72 and 0.69 respectively.

Then, K-Nearest Neighbors (KNN) has an accuracy of 65.19% with a precision and recall score of 0.65 and 0.62 respectively.

Then, Decision Tree Classifier (DTC) has an accuracy of 63.62%, with equal precision and recall score of 0.63.

Machine Learning Algorithms	Accuracy	Precision	Recall
Binary Logistic Regression	72.33%(highest)	0.74	0.67
Random Forest Algorithm	71.47%	0.72	0.69
K-nearest neighbors (KNN)	65.19%	0.65	0.62
Decision Tree Classifier (DTC)	63.62%	0.63	0.63
Naive Bayes Algorithm	58.83%	0.72	0.28

Table 6.1: Classification Result Analysis.

	precision	recall	f1-score	support
0	0.70	0.77	0.74	7069
1	0.74	0.67	0.70	6931
accuracy			0.72	14000
macro avg	0.72	0.72	0.72	14000
weighted avg	0.72	0.72	0.72	14000

Figure 6.1: Classification Report of Binary Logistic Regression.

	precision	recall	f1-score	support
0	0.71	0.73	0.72	7069
1	0.72	0.69	0.71	6931
accuracy			0.71	14000
macro avg	0.71	0.71	0.71	14000
weighted avg	0.71	0.71	0.71	14000

Figure 6.2: Classification Report of Random Forest.

	precision	recall	f1-score	support
0	0.65	0.67	0.66	7069
1	0.65	0.62	0.64	6931
accuracy			0.65	14000
macro avg	0.65	0.65	0.65	14000
weighted avg	0.65	0.65	0.65	14000

Figure 6.3: Classification Report of KNN.

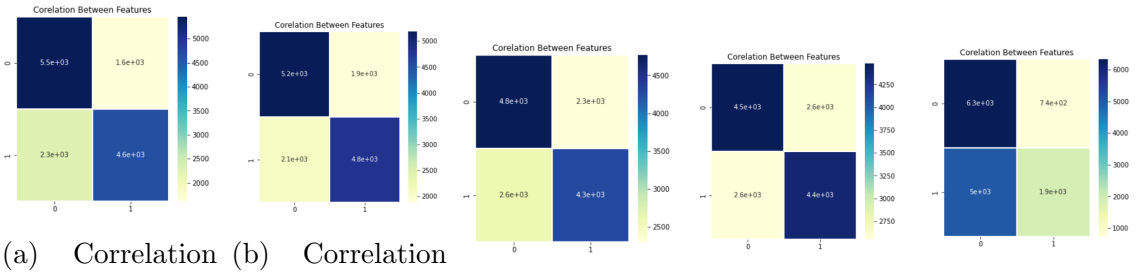
	precision	recall	f1-score	support
0	0.64	0.63	0.64	7069
1	0.63	0.63	0.63	6931
accuracy			0.63	14000
macro avg	0.63	0.63	0.63	14000
weighted avg	0.63	0.63	0.63	14000

Figure 6.4: Classification Report of DTC.

	precision	recall	f1-score	support
0	0.56	0.90	0.69	7069
1	0.72	0.28	0.40	6931
accuracy			0.59	14000
macro avg	0.64	0.59	0.54	14000
weighted avg	0.64	0.59	0.55	14000

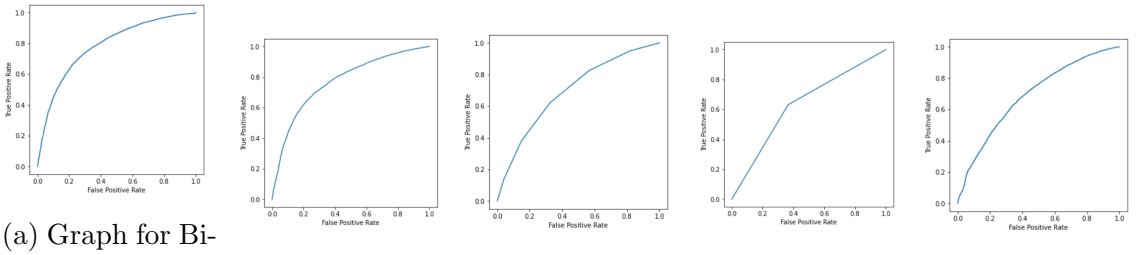
Figure 6.5: Classification Report of Naive Bayes.

Lastly, Naive Bayes has an accuracy of 58.83%, with a precision and recall score of 0.72 and 0.28 respectively.



(a) Correlation between features for Binary Logistic Regression. (b) Correlation between features for Random Forest. (c) Correlation between features for KNN. (d) Correlation between features for KNN. (e) Correlation between features.

Figure 6.6: Correlation between features all the applied algorithms.



(a) Graph for Binary Logistic Regression. (b) Graph for Random Forest. (c) Graph for KNN. (d) Graph for DTC. (e) Graph for Naive Bayes.

Figure 6.7: Graph for all the applied algorithms.

The hybrid algorithm gives an accuracy of 71.35% which does not improve the accuracy hence the implementation of this hybrid algorithm has been exempted.

6.2 Bangladeshi Dataset Result

While fetching the dataset for Bangladesh, the quantity of data was very insufficient present everywhere and the difficulties were not time but limitations in data observatory and data management. We collected the present data and our research needed more quantitative data to get better results so we decided to go for SMOTE technique which helped to increase the number of data without tampering it and help us to reach a proper and accurate conclusion. Since data collection is a barrier regarding research work in Bangladesh, applying this technique helped us and our research to a great extent . This could be an innovative approach towards research work based on ML in Bangladesh.

	SL	Age	Sex	Height_CM	Weight_Kg	Systolic_Upper	Diastolic_Lower	Smoking	HTN	S_Cholesterol	cardio
0	1	42	1	170.0	64.0	120	80	0	0	187	1
1	2	36	1	161.0	75.0	130	80	0	0	194	1
2	3	41	1	170.0	66.0	120	80	0	0	180	1
3	4	33	1	176.0	84.0	100	70	0	0	175	1
4	5	43	1	162.0	72.0	160	110	1	1	197	1
...	...	...	...	...	...	...	...	...	...	...	...
1003	1004	70	0	137.0	32.0	130	90	0	0	205	1
1004	1005	75	1	160.0	59.0	120	70	1	0	182	1
1005	1006	65	0	135.0	45.0	130	80	0	0	176	1
1006	1007	52	0	155.0	60.0	130	70	0	0	176	1
1007	1008	40	0	145.0	45.0	120	80	1	0	150	1

1008 rows × 11 columns

Figure 6.8: Bangladeshi Data set.

Machine Learning Algorithms	Accuracy
Binary Logistic Regression	76.23%
Random Forest Algorithm	95.54%(highest)
K-nearest neighbors (KNN)	82.67%
Decision Tree Classifier (DTC)	93.56%
Naive Bayes Algorithm	78.71%

Table 6.2: Classification Result Analysis on Bangladeshi Dataset.

The Bangladeshi dataset initially had 1000 samples which is very limited in quantity to apply in the prediction system . So we applied SMOTE to solve the oversampling problem and implemented the alogrithms. For the prediction, the same five types of classifiers have been used and same workflow has been followed for better comparison.The classifiers are - KNN, DTC, Naive Bayes Algorithms, RF Algorithm and BLR. The Bangladeshi dataset achieved higher results and RF Algorithm provides the most accuracy level.

Chapter 7

Conclusion

Cardiovascular disease is one of the top factors of mortality across the world. These deadly diseases are rapidly becoming the leading cause of morbidity and mortality on a global scale. As a result, researchers are working hard to develop a faster and more effective diagnosis process for this disease using various ML algorithms for the early detection and prevention of CHD. Indicators such as blood pressure, oxygen level, haemoglobin concentration, and other parameters that are useful in predicting cardiac risk are used to develop this evaluation.

In this paper, we wanted to do a comparative analysis of different existing algorithms such as K Nearest Neighbour, Decision Tree, Binary Logistic, RF and Naive Bayes on Cardiovascular Disease Dataset. We also compared different features of Bangladeshi Dataset to find better predictions and who are more likely to get affected by these diseases.

Since, these models change over the quantity of data, in this research, the models gave higher accuracy with a huge dataset that consists of 70000 data. Also, due to time constraints and insufficient presence of data in the system, we only could collect limited data for the Bangladeshi population. Hence,The future direction of this study is to collect the data in person through surveys by providing questionnaires among individuals to collect adequate data for comparative analysis of the aforementioned algorithms on Bangladeshi population and to develop an application based system that will detect the indication of CHD and alert the user to seek medical aid promptly.

As the phrase goes, “Prevention is better than cure” so it is better to predict the disease precisely as early in the process and seek appropriate medical guidance before it is too late. Our system will be an effective attempt at the stage of prevention by providing a model with better accuracy. Thus, the prediction at an early stage will save lives by making them aware of CHD beforehand and buy them sufficient time towards proper treatment and a higher life expectancy.

Bibliography

[1] T. A. Gaziano, A. Bitton, S. Anand, S. Abrahams-Gessel, and A. Murphy, "Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries," *Current problems in cardiology*, vol. 35, no. 2, pp. 72–115, Feb. 2010, issn: 0146-2806. doi: 10.1016/j.cpcardiol.2009.10.002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864143/ (visited on 05/22/2022).

[2] M. R. R. Patil, *Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing*, May 2014. [Online]. Available: https://ijarcce.com/wp-content/uploads/2012/03/IJARCCE9E-a-rupali-Heart-Disease-Prediction.pdf.

[3] A. M. Islam and T. Paul, *Cardiovascular Disease in Bangladesh: A Review*, Apr. 2017. [Online]. Available: https://www.researchgate.net/publication/316572340_Cardiovascular_Disease_in_Bangladesh_A_Review.

[4] *Coronary heart disease - Diagnosis*, en, Oct. 2018. [Online]. Available: https://www.nhs.uk/conditions/coronary-heart-disease/diagnosis/ (visited on 05/22/2022).

[5] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Apr. 2019, pp. 1–5. doi: 10.1109/ICIICT1.2019.8741465.

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81 542–81 554, 2019, issn: 2169-3536. doi: 10.1109/ACCESS.2019.2923707.

[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 281, Dec. 2019, issn: 1472-6947. doi: 10.1186/s12911-019-1004-8. [Online]. Available: https://doi.org/10.1186/s12911-019-1004-8 (visited on 05/22/2022).

[8] J. Brownlee, *SMOTE for Imbalanced Classi cation with Python*, en-US, Jan. 2020. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (visited on 05/22/2022).

[9] L. Brunese, F. Martinelli, F. Mercaldo, and A. Santone, "Deep learning for heart disease detection through cardiac sounds," en, *Procedia Computer Science*, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, vol. 176, pp. 2202–2211, Jan. 2020, issn: 1877-0509. doi: 10.1016/j.procs.2020.09.257. [Online].

Available: https://www.sciencedirect.com/science/article/pii/S187705092032161X (visited on 05/22/2022).

[10] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan, and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 570–575. doi: 10.1109/ICICCS48265.2020.9121169.

[11] R. Shashikant and P. Chetankumar, "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter," *Applied Computing and Informatics*, vol. ahead-of-print, no. ahead-of-print, Jan. 2020, issn: 2210-8327. doi: 10.1016/j.aci.2019.06.002. [Online]. Available: https://doi.org/10.1016/j.aci.2019.06.002 (visited on 05/22/2022).

[12] V. D. Soni, *Detection Of Heart Disease Using Machine Learning Techniques*, Aug. 2020. [Online]. Available: https://www.researchgate.net/publication/343813156_Detection_Of_Heart_Disease_Using_Machine_Learning_Techniques.

[13] CDC, *Heart Disease Resources | cdc.gov*, en-us, Sep. 2021. [Online]. Available: https://www.cdc.gov/heartdisease/about.htm (visited on 05/22/2022).

[14] A. Ishaq, S. Sadiq, M. Umer, *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, no. 9370099, pp. 39 707–39 716, 2021, issn: 2169-3536. doi: 10.1109/ACCESS.2021.3064084. [Online]. Available: http://www.scopus.com/inward/record.url?scp=85103755876&partnerID=8YFLogxK (visited on 05/22/2022).

[15] Y. Jiang, X. Zhang, R. Ma, *et al.*, "¡p¿Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China¡/p¿," English, *Clinical Epidemiology*, vol. 13, pp. 417–428, Jun. 2021. doi: 10.2147/CLEP.S313343. [Online]. Available: https://www.dovepress.com/cardiovascular-disease-prediction-by-machine-learning-algorithms-based-peer-reviewed-fulltext-article-CLEP (visited on 05/22/2022).

[16] *Logistic Regression | What is Logistic Regression and Why do we need it?* en, Aug. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/ (visited on 05/22/2022).

[17] *Cardiovascular diseases (CVDs)*, en. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (visited on 05/22/2022).

[18] *Coronary Heart Disease - What Is Coronary Heart Disease? | NHLBI, NIH*. [Online]. Available: https://www.nhlbi.nih.gov/health/coronary-heart-disease (visited on 05/22/2022).

[19] *Coronary Heart Disease in Bangladesh*, en. [Online]. Available: https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease (visited on 05/22/2022).

[20] *Healthy Heart - Happy Feet*. [Online]. Available: http://helo.health/ipdi.php (visited on 05/22/2022).

[21] *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*, en. [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (visited on 05/22/2022).

[22] *Logistic Regression with Python*. [Online]. Available: https://odsc.medium.com/logistic-regression-with-python-ede39f8573c7%E2%80%8C.

[23] *Machine Learning Random Forest Algorithm - Javatpoint*, en. [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm (visited on 05/22/2022).

[24] *The big-data revolution in US health care: Accelerating value and innovation | McKinsey*. [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care (visited on 05/22/2022).

[25] *Using Binary Classi cation Metrics to Maximize Enterprise AI's Potential*, en-US. [Online]. Available: https://c3.ai/blog/using-binary-classification-metrics-to-maximize-enterprise-ais-potential/ (visited on 05/22/2022).

[26] *What is Hybrid Algorithm | IGI Global*. [Online]. Available: https://www.igi-global.com/dictionary/particle-swarm-optimization-algorithm-its/13449 (visited on 05/22/2022).