

**PREDICTING THE RISK OF DEVELOPING CORONARY HEART DISEASE BASED ON  
LIFESTYLE AND HEALTH FACTORS**

**LONGDI XIAN**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2023**

**PREDICTING THE RISK OF DEVELOPING CORONARY HEART  
DISEASE BASED ON LIFESTYLE AND HEALTH FACTORS**

**LONGDI XIAN**

**[DISSERTATION] SUBMITTED IN [FULFILMENT] OF THE REQUIREMENTS  
FOR THE DEGREE OF [MASTER OF DATA SCIENCE]**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY  
UNIVERSITY OF MALAYA  
KUALA LUMPUR**

**2023**

**UNIVERSITY OF MALAYA**  
**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: LONGDI XIAN (I.C/Passport No: EJ5449088 )

Matric No: S2172650

Name of Degree: Master Of Data Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Predicting the risk of developing coronary heart disease based on lifestyle and health factors

Field of Study:

Machine Learning

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes, and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature 威龙迪

Date: 31/03/2023

Subscribed and solemnly declared before,

Witness's Signature

Date:

Name:

Designation:

**UNIVERSITI MALAYA**  
**PERAKUAN KEASLIAN PENULISAN**

Nama: LONGDI XIAN (No. K.P/Pasport: EJ5449088)

No. Matrik: S2172650

Nama Ijazah: Guru Sains Data

Tajuk Kertas Projek/Laporan Penyelidikan/Disertasi/Tesis (“Hasil Kerja ini”):

Memandangkan risiko untuk mengembangkan penyakit jantung koronari berdasarkan gaya hidup dan faktor kesehatan

Bidang Penyelidikan: Pembelajaran Mesin

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

- (1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
- (2) Hasil Kerja ini adalah asli;
- (3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
- (4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
- (5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya (“UM”) yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
- (6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon

威龙迪

Tarikh: 31/03/2023

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi

Tarikh:

Nama:

Jawatan:

**PREDICTING THE RISK OF DEVELOPING CORONARY HEART DISEASE BASED ON LIFESTYLE AND  
HEALTH FACTORS**

**ABSTRACT**

Until now, there were many terrible and fatal diseases in addition to newer ones, such as neo-coronary pneumonia and monkeypox. Coronary heart disease (CHD) is one of these high-risk, fatal diseases. Many lifestyle and health factors are arguably the most important causative environments leading to CHD. Despite the increasing availability of scientific and technological tools, many prediction systems are still unable to predict these diseases reasonably. The need for more effective prediction systems has not yet been adequately met. In order to further investigate the impact of lifestyle and health factors on coronary heart disease, as well as what health factors and lifestyles have a large impact on coronary heart disease and how to improve the accuracy of machine-learning models in predicting coronary heart disease and other important issues, this study aims to identify lifestyle and health factors, construct models that optimally predict coronary heart disease risk and develop a convenient web application for users and healthcare practitioners. The data for this study utilizes datasets on lifestyle and health factors (e.g., smoking, alcohol consumption, physical activity, hypertension, and cholesterol levels) collected from public websites. By using the XGBoost algorithm to define high-risk factors for CHD, it was finally obtained that hypertension, BMI, Difficult Walking, High Cholesterol, Income, Age, Stroke, MentHlth, Sex, GenHlth, Smoker, and Diabetes were the most important lifestyle factors. Based on the defined risk factors and using five machine learning algorithms (AdaBoost, Random Forest, Decision Tree, KNN, and Naive Bayes) and based on the five models and using Smote balanced dataset to construct machine learning models. All the models were subjected to Recall, F1 score, Precision, Accuracy, and ROC curve values to

evaluate the accuracy of the machine learning algorithms. Finally, the best prediction model (AdaBoost+Smote) was found. This model can achieve more than 90% accuracy in predicting CHD risk. The best model was combined with Python and Dash to develop a very convenient web application, which was tested and evaluated by Unit Testing and Data Validation, and the final test and evaluation results showed that it met the expected normal operation goals.

Keywords: coronary heart disease, machine learning algorithms, prediction accuracy



**MERAMALKAN RISIKO UNTUK MENGEMBANGKAN PENYAKIT JANTUNG KORONARI BERDASARKAN  
GAYA HIDUP DAN FAKTOR KESEHATAN**

**ABSTRAK**

Sehingga sekarang, terdapat banyak penyakit yang mengerikan dan mematikan selain daripada penyakit yang lebih baru, seperti pneumonia neokoronari dan cacing monyet. Penyakit jantung koronari (CHD) adalah salah satu penyakit yang berbahaya tinggi ini. Banyak faktor gaya hidup dan kesihatan mungkin adalah persekitaran penyebab yang paling penting yang menyebabkan CHD. Walaupun kemampuan alat saintifik dan teknologi meningkat, ramai sistem ramalan masih tidak dapat meramalkan penyakit ini secara rasional. Perlukan sistem ramalan yang lebih berkesan belum dipenuhi dengan cukup. Untuk menyelidiki lebih lanjut kesan gaya hidup dan faktor kesihatan pada penyakit jantung koronari, serta apa faktor kesihatan dan gaya hidup mempunyai kesan besar pada penyakit jantung koronari dan bagaimana untuk meningkatkan ketepatan model pembelajaran mesin dalam meramalkan penyakit jantung koronari dan masalah penting lain, kajian ini bertujuan untuk mengenalpasti gaya hidup dan faktor kesihatan, - membina model yang secara optimal meramalkan risiko penyakit jantung koronari dan mengembangkan aplikasi web yang sesuai untuk pengguna dan doktor rawatan kesihatan. Data untuk kajian ini menggunakan set data tentang gaya hidup dan faktor kesihatan (cth., merokok, konsumsi alkohol, aktiviti fizik, hipertensi, dan aras kolesterol) yang dikumpulkan dari laman web awam. Dengan menggunakan algoritma XGBoost untuk menentukan faktor risiko tinggi untuk CHD, ia akhirnya mendapat bahawa tekanan tinggi, BMI, Perjalanan Sulit, Kolesterol Tinggi, Income, Age, Stroke, MentHlth, Sex, GenHlth, Smoker, Diabetes adalah faktor gaya hidup yang paling penting. Berdasarkan faktor risiko yang ditakrif dan menggunakan lima algoritma pembelajaran mesin (AdaBoost, Random Forest, Decision Tree, KNN, dan Naive Bayes) dan berdasarkan lima model dan

menggunakan set data yang seimbang Smote untuk membina model pembelajaran mesin. Semua model telah ditakdirkan mengingat, skor F1, Tepat, Tepat, dan nilai lengkung ROC untuk menilai ketepatan algoritma pembelajaran mesin. Akhirnya, model ramalan terbaik (AdaBoost+Smote) ditemui. Model ini boleh mencapai lebih dari 90% ketepatan dalam meramalkan risiko CHD. Model terbaik digabungkan dengan Python dan Dash untuk mengembangkan aplikasi web yang sangat selesa, yang diuji dan diuji oleh Unit Ujian dan Pengesahan Data, dan hasil ujian akhir dan penilaian menunjukkan bahawa ia memenuhi tujuan operasi normal yang dijangka.

Keywords: Ramalan penyakit jantung, algoritma pembelajaran mesin, ketepatan ramalan

## ACKNOWLEDGEMENTS

I want to express my gratitude to those individuals and organizations that have contributed to collecting and analyzing data sets on "predicting the risk of coronary heart disease based on lifestyle and health factors."

First, we would like to thank my tutor for her detailed and thoughtful comments. Without her efforts, this research would not have been possible.

Secondly, I would like to thank the open-source Kaggle website, without which I could not download valuable datasets.

In addition, we would like to recognize the role of technology and statistical tools that enable us to analyze datasets and derive meaningful insights from them.

Finally, we would like to thank the broader scientific community for sharing their knowledge and contributing to research progress in cardiovascular disease.

In summary, we recognize the collaborative efforts and dedication of all those who have contributed to this study and their valuable contributions to improving our understanding of coronary heart disease risk factors.

## TABLE OF CONTENTS

<b>CHAPTER 1: INTRODUCTION</b>	<b>1</b>
1.1 Introduction	1
1.2 Research Background	3
1.3 Problem Statements	3
1.4 Research questions	4
1.5 Research Objectives	4
1.6 Research Significances	5
<b>CHAPTER 2: LITERATURE REVIEW</b>	<b>7</b>
2.1 Coronary Heart Disease Prediction Using Data Mining	7
2.1.1 Comparison Table For Coronary Heart Disease Prediction Using Data Mining	9
2.2 Coronary Heart Disease Prediction Using Machine Learning	10
2.2.1 Comparison Table For Coronary Heart Disease Prediction Using Machine Learning	12
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>14</b>
3.1 Research Design	14
3.2 Data Science Project Framework	14
3.3 Experiment	15
3.3.1 Data Collection	15
3.3.2 Exploratory Data Analysis	20
3.3.3 Data Preprocessing	25
3.2.3.1 Label Encoding	25
3.2.3.2 Strongly Risk Factors Identify And Dataset Regeneration	25
3.2.3.3 Partition Of Test Set and Training Set	28
3.2.3.4 Handling Imbalanced Data	28
3.2.3.5 K-Fold Cross Validation	30
3.3.4 Modeling	30
3.3.5 Developing Web Application	33
3.3.6 Testing And Evaluating Web Application	35
<b>CHAPTER 4: Result</b>	<b>38</b>
4.1 Identify The Lifestyle And Health Factors	38
4.1.1 Feature Importance Statistics And Analysis	38
4.1.2 Shape Figure Analysis	41
4.1.3 Summary	42
4.2 Comparison Of Model Performance	42
4.2.1 Accuracy	44
4.2.2 Precision	46
4.2.3 Recall	47

4.2.4 F1-Score	49
4.2.5 ROC Curve	51
4.2.6 Summary	52
4.3 Evaluating Web Application	53
4.3.1 Unit Testing	53
4.3.2 Data Validation	54
4.3.3 Cross-browser And Cross-device Testing	56
4.3.4 Error Handling And Logging	57
<b>CHAPTER 5: Conclusion</b>	<b>59</b>
<b>References</b>	<b>61</b>

## LIST OF FIGURES

Figure 3. 1 Data science OSEM framework	18
Figure 3. 2 Research Design Flowchart	19
Figure 3. 3 Summary statistics for each categorical field physical and mental health indicators	24
Figure 3. 4 Bar chart of the impact of factors on coronary heart disease	25
Figure 3. 5 Summary statistics for each numerical field in the raw data set.	26
Figure 3. 6 Relationship between each numerical variable in the dataset and the target variable.	27
Figure 3. 7 Correlation among features	28
Figure 3. 8 Original dataset with label encoding Yes or No	29
Figure 3. 9 Standardized dataset with label encoding 1 or 0	29
Figure 3. 10 Using XGBoost to calculate the correlation between CHD and other factors	30
Figure 3. 11 Code of Combining the average value with the feature importance value obtained three times	31
Figure 3. 12 Code of Merge Feature Factors	31
Figure 3. 13 Code of Visualization of top factors that associated with CHD	31
Figure 3. 14 Code of Generate a new dataset based on identified factors	32
Figure 3. 15 Split the dataset into train set and test set	32
Figure 3. 16 Target values Rate in Dataset	33
Figure 3. 17 Balance the dataset using Smote	33
Figure 3. 18 10-Fold Cross-Validation	34
Figure 3. 19 Integrated machine learning model	37
Figure 3. 20 Develop Web Application	38
Figure 4. 1 Bar chart of the degree of correlation between lifestyle and health factors related to coronary heart disease(Weight $\geq$ 1%,Gian $\geq$ 1%and Cover $\geq$ 1%)	42
Figure 4. 2 Overall SHAP explanations	43
Figure 4. 3 Confusion Matrix	45
Figure 4. 4 Accuracy For Different Model's Performance	46
Figure 4. 5 Precision For Different Model's Performance	48
Figure 4. 6 Recall-Score For Different Model's Performance	50
Figure 4. 7 F1-Score For Different Model's Performance	51
Figure 4. 8 The Receiver Operating Characteristic (ROC) Curve For Different Model's Performance	53
Figure 4. 9 Height and Weight Validation	56
Figure 4. 10 Mental Health and Income Validation	56
Figure 4. 11 Log files	59
Figure 4. 12 Error log	59

## LIST OF TABLES

Table 2. 1 Comparison Table For Coronary heart disease prediction using data mining	14
Table 2. 2 Comparison Table For Coronary heart disease prediction using machine learning	17
Table 4. 1 XGBoost model feature importance statistics.	40
Table 4. 2 Unit Testing	54
Table 4. 3 Cross-browser And Cross-device Testing Table	57

## CHAPTER 1:INTRODUCTION

### 1.1 Introduction

Coronary heart disease (CHD) holds the top position as the primary cause of global mortality and incapacitation. This ailment transpires when the arteries responsible for delivering blood to the heart experience constriction or obstruction due to the buildup of plaque. Consequently, this leads to reduced blood and oxygen supply to the heart. Although several modifiable CHD risk factors have been identified, accurately predicting an individual's risk of CHD based on these factors remains a challenge.

The first problem is Coronary heart disease (CHD) is a global epidemic and the leading cause of death, necessitating urgent efforts to understand the specific lifestyle and health factors that contribute to its development. While there is growing recognition that lifestyle and health factors play a crucial role in CHD, there is a lack of definitive knowledge regarding the specific factors most closely associated with its occurrence. This research gap poses a significant obstacle to developing targeted preventive measures and interventions. Therefore, there is an immediate need to address this gap and establish a comprehensive understanding of the lifestyle and health factors most strongly correlated with the development of coronary heart disease. Such knowledge will provide crucial insights for public health strategies and enable the implementation of effective preventive measures to reduce the burden of CHD worldwide.

The second problem is that While lifestyle and health factors have been recognized as critical determinants of CHD risk, there is a need to develop a robust machine learning model that can effectively integrate and analyze these factors to provide accurate risk predictions. Therefore, the problem addressed in this study is the lack of a comprehensive and reliable machine learning model that can predict the risk of developing coronary heart disease based on identified lifestyle and health factors.

The third problem is that While machine learning models have shown promise in predicting the risk of coronary heart disease (CHD) using lifestyle and health factors, there is a notable gap in the availability of accurate web application tools that effectively identify high-risk groups for CHD based on these factors.

Recent advances in machine learning algorithms have shown promise in predicting coronary heart disease risk based on lifestyle and health factors. These algorithms can analyze large and complex data sets, identify patterns and relationships among different risk



factors, and generate accurate predictions. However, only some studies have compared the performance of different machine learning algorithms in predicting coronary heart disease risk.

Therefore, this study aims to identify the lifestyle and health factors most closely related to the development of coronary heart disease and develop and evaluate machine learning algorithms based on these factors to predict the risk of coronary heart disease. The study also aims to compare the performance of different machine learning algorithms in predicting coronary heart disease risk, to develop more accurate and efficient prediction models. The findings of this study can provide information for targeted prevention strategies and help Healthcare providers identify high-risk populations with coronary heart disease. In addition, comparing the performance of different machine learning algorithms can help develop more accurate and effective CHD risk prediction models. Overall, the results of this study have the potential to improve cardiovascular health and reduce the burden of coronary heart disease worldwide.

## **1.2 Research Background**

Coronary heart disease (CHD) is a complex disease with a multifactorial etiology involving interactions between genetic, lifestyle, and environmental factors. Previous studies have identified several modifiable risk factors associated with CHD, including smoking, high blood pressure, high cholesterol, and an unhealthy diet. However, it is still challenging to accurately predict an risk of developing CHD based on these factors.

Recently, machine learning algorithms have shown promise in predicting CHD risk based on lifestyle and health factors. These algorithms can analyze large and complex datasets, identify patterns and relationships between risk factors, and generate accurate predictions. However, only some studies have compared the performance of different machine learning algorithms in predicting CHD risk.

This study aimed to identify the lifestyle and health factors most strongly associated with the development of CHD and to develop and evaluate machine learning algorithms that predict CHD risk based on these factors. The findings of this study can inform targeted prevention strategies and help healthcare providers identify individuals at high risk of developing CHD. Moreover, comparing the performance of different machine learning

algorithms can contribute to developing more accurate and efficient prediction models for CHD risk.

### **1.3 Problem Statements**

- Coronary artery disease stands as the primary global cause of mortality. It's been established that lifestyle and health elements significantly contribute to the onset of this condition. Nevertheless, a clear-cut declaration concerning the lifestyle and health factors with the strongest links to coronary heart disease development is yet to be established.(Nazli, Sukma Azureen, et al.,2021)
- Although many machine learning scholars who predict coronary heart disease have used machine learning models to predict coronary heart disease risk based on lifestyle and health factors, the accuracy of predicting CHD risk using machine learning models based on identified lifestyle and health factors is still low and needs further improvement.(Alaa, Ahmed M., et al.,2019)
- Using machine learning models to predict the risk of coronary heart disease using lifestyle and health factors has achieved numerous effective results. However, there is currently a lack of available and accurate Web application tools based on machine learning models to effectively identify high-risk groups of coronary heart disease using lifestyle and health factors.(Huang, Weiting, et al.,2022)

### **1.4 Research questions**

1. How to solve the problem of no definitive statement regarding the lifestyle and health factors most closely related to the development of coronary heart disease?
2. How to solve the problem of low accuracy in predicting CHD risk based on identified lifestyle and health factors in machine learning models, which still needs further improvement?
3. How to solve the lack of web applications that utilize machine learning models to predict the risk of coronary heart disease by utilizing lifestyle and health factors?

## **1.5 Research Objectives**

1. To identify the lifestyle and health factors most strongly associated with the development of coronary heart disease.
2. To determine an optimal machine learning model that can predict the risk of CHD based on lifestyle and health factors.
3. To develop a web application based on the machine learning model in predicting the risk of developing coronary heart disease based on lifestyle and health factors.

## **1.6 Research Significances**

This study aims to address significant challenges in coronary heart disease prediction and risk assessment using machine learning models and lifestyle and health factors. The research objectives have practical significance and potential benefits, summarized as follows:

Improve the accuracy of CHD risk prediction by enhancing machine learning models based on identified lifestyle and health factors. Identifying strongly related lifestyle and health factors helps to understand the root cause of the disease. This information can help medical professionals, researchers, and policymakers to design targeted interventions and prevention strategies to reduce the incidence rate of coronary heart disease. Healthcare professionals can obtain more reliable patient risk assessments by developing more accurate models. This study helps with early detection, timely intervention, and personalized treatment plans, ultimately improving patient prognosis and reducing medical costs. Developing a network application that utilizes machine learning models to predict the risk of coronary heart disease based on lifestyle and health factors is a significant contribution. This tool can be used by many users, including healthcare professionals, individuals concerned with heart health, and researchers. This network application can provide convenient and user-friendly risk assessment and increase awareness of coronary heart disease prevention and proactive measures.

Meanwhile, this study contributes to advancing machine learning technology in healthcare. Applying these models to coronary heart disease risk prediction expands the application scope of machine learning and demonstrates the potential for improving disease prevention and management. This study can inspire further research and innovation in applying machine learning to other cardiovascular diseases and medical conditions. And, it

help to reduce the burden of global coronary heart disease, improve patient prognosis, and promote positive methods for cardiovascular health.

## **CHAPTER 2:LITERATURE REVIEW**

### **2.1 Coronary Heart Disease Prediction Using Data Mining**

In recent times, the realm of forecasting coronary heart disease has garnered substantial attention from scholars. Given the escalating pervasiveness of this ailment, it has become imperative to discern dependable and precise prognostic techniques. This investigation strives to delve into and juxtapose an array of data excavation methodologies and algorithms, with the intent of ascertaining the utmost accuracy in forecasting coronary heart disease. Through a meticulous scrutiny of each approach's merits and limitations, the study aims to pinpoint the most trustworthy and efficacious avenues for prognostication. The core objective is to cultivate a dependable and precise model for foreseeing coronary heart disease. Such a model holds the potential to empower medical practitioners and healthcare experts in making premature diagnoses, ultimately fostering superior therapeutic results and ameliorated patient consequences. By harnessing the prowess of data excavation methodologies and algorithms, the research seeks to unveil pivotal risk determinants and premature signals of coronary heart disease. The outcomes of this inquiry bear substantial ramifications for the curbing and management of coronary heart disease, a paramount contributor to global mortality.

Several studies have been conducted in this area, including those by Singh et al. (2018), Meda & Bhogapathi (2018), Wan Zunaidi et al.(2018), Fadnavis et al. (2021) and K. al-Taie et al. To predict coronary heart disease, these studies used various data mining algorithms such as multilayer perceptron neural networks, fuzzy neurogenesis algorithms, decision tree algorithms, and Bayesian classifiers.

Singh et al. (2018) utilized a publicly available coronary heart disease database. They used a multilayer perceptron neural network with backpropagation as a training algorithm and achieved 80% accuracy in predicting coronary heart disease. Media and Bhogapathi (2018) proposed a method for predicting coronary heart disease using FNGA and different element selection strategies for effective prediction. They found that the fuzzy neurogenesis algorithm was more accurate than traditional methods in identifying coronary heart disease. Wan Zunaidi et al.(2018), using a coronary heart disease dataset collected from the Hungarian Institute of Cardiology, concluded that multilayer perceptual neural networks were the best choice among the algorithms they tested. Fadnavis et al.(2021) studied the Cleveland Coronary coronary heart disease dataset and found that the decision tree

classification model had an accuracy of 87.97% compared to 85.25% for the standard Bayesian classifier. al-Taie et al. (2021) analyzed 200 samples from a shared dataset obtained from Iraqi hospitals and found that Bayes Net had the highest classification accuracy of approximately 80.50%. Wang et al. (2021) randomly selected 100 patients with occult CHD admitted to public health hospitals from February 2016 to December 2020 (52 males and 48 females, aged 45 ~ 72 years) to experiment. They also explored the sensitivity and accuracy of the classification algorithm and traditional EGM in data mining to diagnose arrhythmias in patients with occult coronary artery disease. They found that the CART algorithm had the highest accuracy of 84% in detecting ICEGM.

The results of these studies have demonstrated that various algorithms, such as multilayer perceptron neural networks, fuzzy neurogenetic algorithms, decision tree algorithms, Bayesian algorithms, and random forests, can effectively predict coronary heart disease with varying degrees of accuracy. Data mining algorithms can become essential for healthcare professionals in diagnosing and treating coronary heart disease. Despite the progress made in this field, there is still room for further research to improve the accuracy of coronary heart disease prediction using data mining techniques. One area of focus could be exploring the combination of different algorithms to achieve higher accuracy rates. Additionally, more extensive datasets could be used to improve the prediction accuracy of these algorithms. Overall, data mining algorithms in predicting coronary heart disease hold great promise for improving the health outcomes of patients with this condition. By developing more reliable and accurate prediction models, healthcare professionals can provide earlier diagnoses and more effective treatments, ultimately improving patient outcomes and quality of life. Further research and development could pave the way for more significant coronary heart disease prediction and treatment advancements.

### **2.1.1 Comparison Table For Coronary Heart Disease Prediction Using Data Mining**

**Table 2. 1 Comparison Table For Coronary heart disease prediction using data mining**

Study	Dataset	Method	Accuracy (%)
-------	---------	--------	--------------

Singh et al. 2018	publicly available coronary heart disease database	MLPNN with backpropagation using Weka 3.6.11	80
Meda & Bhogapathi 2018	Andhra Pradesh population	Fuzzy neurogenetic algorithms (FNGA)	FNGA more accuracy
Wan Zunaidi et al. 2018	Hungarian Institute of Cardiology	Multilayer perceptron neural network (MLPNN)	MLPNN more accuracy
Fadnavis et al. 2021	Cleveland and Statlog datasets	Naïve Bayes and decision trees	87.97 (decision tree), 85.25 (Naïve Bayes)
K. AL-Taie et al. 2021	dataset obtained from Iraqi hospitals	SMO, MLP, Bayesian networks, random forests	71(SMO),80(MLP),84.50 (Bayesian networks), 83 (MLP)
Wang et al. 2021	Dataset from patients with occult CHD	Classification and Regression Tree (CART)	84 (CART)

## 2.2 Coronary Heart Disease Prediction Using Machine Learning

Coronary heart disease, a leading global fatality cause, underscores the significance of early identification and precise prognosis for enhanced treatment outcomes and life preservation. The domain of machine learning has exhibited encouraging outcomes in prognosticating coronary heart disease. However, the intricate task of discerning the most efficacious algorithms and methodologies remains. Consequently, numerous investigators have executed prognostic endeavors concerning machine learning algorithms in this realm. Their central investigative approach revolves around the exploration of diverse machine learning algorithms and strategies, encompassing decision trees, neural networks, and logical regression. The aim is to discern the paramount algorithms and configurations that excel in coronary heart disease prediction. Contemporary research in this machine

learning-driven anticipation of coronary heart disease reveals auspicious findings, fueled by the ailment's pervasive prevalence and the potential dividends of timely identification and intervention.

Furthermore, directly use patient information datasets to train and test machine learning models. The accuracy and performance of each model will be evaluated using indicators such as accuracy, recall, and F1 scores. The direction is to find a reliable and accurate way to predict coronary heart disease, which can lead to earlier diagnoses and better treatment outcomes.

Recent research has employed machine learning algorithms to anticipate cardiovascular incidents during exercise assessment among patients with coronary artery disease, and to scrutinize the efficacy of intelligent systems empowered by machine learning in prognosticating coronary heart disease. Shen et al. (2022) conducted an investigation into various machine learning techniques, including AUC, SVM, logistic regression, GBDT, and XGBoost, to prognosticate cardiovascular events using 16645 patients afflicted by coronary heart disease who underwent cardiopulmonary exercise testing (CPET) at Peking University Affiliated Hospital between January 2016 and September 2019. The outcomes indicated that machine learning methods, particularly XGBoost, displayed adeptness in accurately predicting cardiovascular incidents. In another study, Absar et al. (2022) employed four machine learning models, namely random forest (RF), decision tree (DT), AdaBoost (AB), and K-nearest neighbor (KNN), to predict coronary heart disease using datasets from Cleveland, Hungary, Switzerland, and Long Beach (CSLB). The findings revealed KNN as the most precise model for coronary heart disease prediction. Additionally, Hassan et al. (2022) harnessed machine learning classifiers encompassing gradient enhancement trees (GBTs), multi-layer perceptrons (MLPs), random forests (RF), and UCI-repository datasets to foresee the presence of cardiac issues. The results underscored the random forest model (RF) as the most accurate in prognosticating coronary heart disease. These investigations collectively underscore the potential of machine learning algorithms in predicting coronary heart disease, ultimately enhancing diagnostic and therapeutic precision.

Through a comprehensive examination of existing literature, it has come to light that machine learning algorithms have showcased remarkable efficacy in forecasting coronary heart disease with a notable degree of precision. Nevertheless, it is apparent that the selection of the algorithm and the dataset wield significant influence in determining the



predictive accuracy. The careful curation of pertinent attributes and the precise preprocessing of data emerge as pivotal in attaining precise prognostications. The fidelity of these predictions is subject to variables including the dataset's scale and caliber, the nature of chosen attributes, and the specific algorithm employed. Consequently, there exists a need for further exploration to delve into the most optimal methodologies and techniques for attaining heightened prognostic precision. The overarching objective of this inquiry remains the enhancement of treatment outcomes for individuals afflicted by coronary heart disease, as well as the provision of timely and precise diagnoses, culminating in a marked amelioration of their quality of life.

### 2.2.1 Comparison Table For Coronary Heart Disease Prediction Using Machine Learning

**Table 2. 2Comparison Table For Coronary heart disease prediction using machine learning**

Study	Dataset	Methods	Accuracy
Shen et al. (2022)	Pre-test clinical data and exercise data	AUC, SVM, logistic regression, GBDT, XGBoost	SVM: 68.6 % LR: 77.8% GBDT: 78.4% XGBoost: 79.4%
Absar et al. (2022)	CHSLB dataset (Cleveland, Hungary, Switzerland, Long Beach)	Random Forest (RF), Decision Tree (DT), AdaBoost (AB), K-Nearest Neighbour (KNN)	RF: 87.03% DT: 85.10% AB:87% KNN: 87.83%

Sarin et al. (2022)	BRFSS Dataset	Random Forest (RF), XGBoost, AdaBoost (AB), Naive Bayes (NB)	RF:86.6% XGBoost:77.54% AB:87.8% NB:86%
Hassan et al. (2022)	UCI repository dataset	Gradient Augmented Tree (GBT), Multilayer Perceptron (MLP), Random Forest (RF)	GBT: 79.5% MLP: 80% RF: 76%

## CHAPTER 3:RESEARCH METHODOLOGY

### 3.1 Research Design

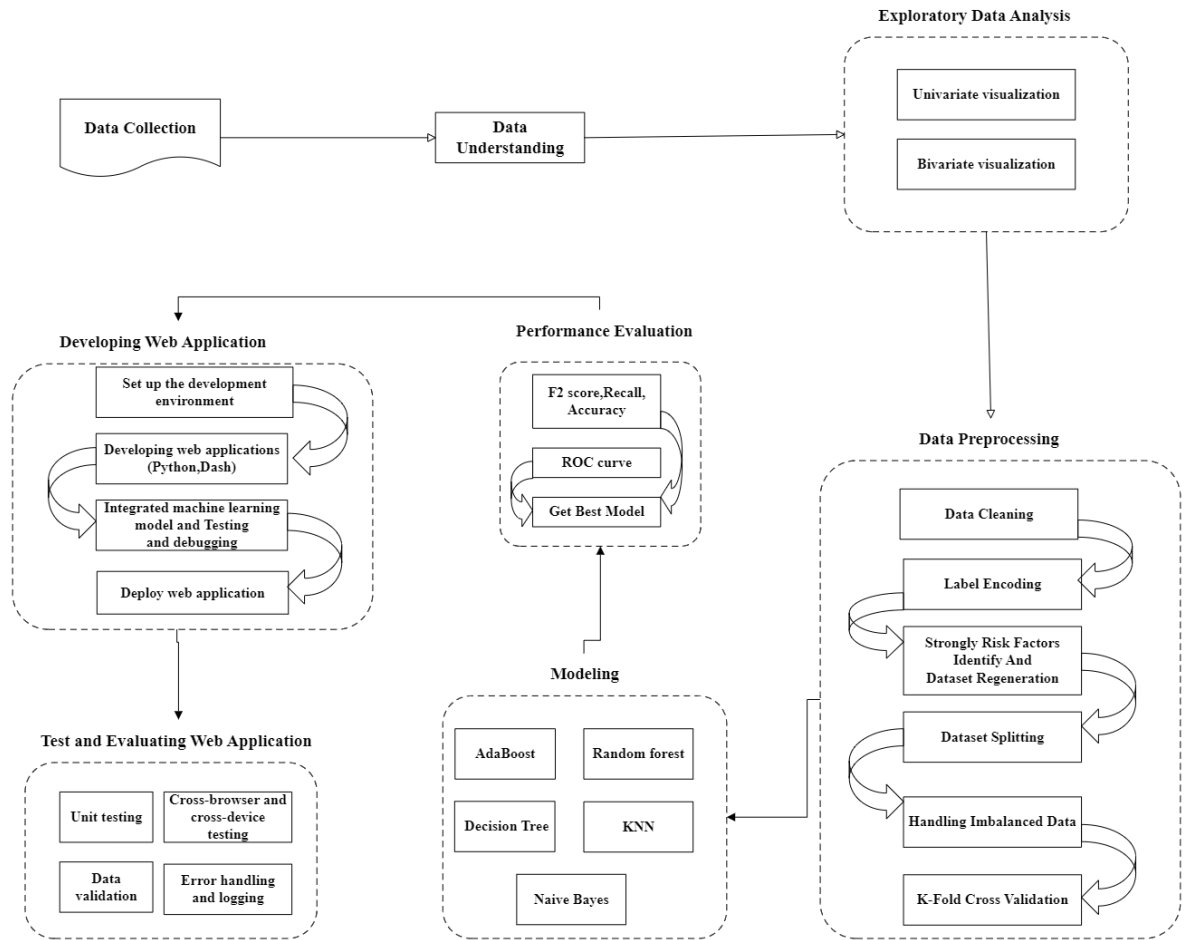
This research entails a quantitative investigation focused on identifying the primary factors carrying substantial risks for coronary heart disease. Employing machine learning models, the study aims to forecast the likelihood of developing this condition. The dataset employed was sourced from openly accessible online platforms.

### 3.2 Data Science Project Framework

Based on Figure 3.1 data science OSEM<sup>N</sup> framework(Kumari, etc.,2020), this study will be conducted in six steps. This study includes Data Collection, Data Understanding, Exploratory Data Analysis, Data Preprocessing, Modeling, Performance Evaluation, Developing a Web application, and Test and evaluating the Web application each stage plays a crucial role in achieving the overall goal of predicting the risk of developing coronary heart disease. Figure 3.2 shows the specific implementation process of this study.



Figure 3. 1 Data science OSEM<sup>N</sup> framework



**Figure 3. 2 Research Design Flowchart**

### 3.3 Experiment

#### 3.3.1 Data Collection

The prevalence of coronary artery disease was established through examination of medical records and subsequent follow-up surveys. Participants were categorized as individuals with coronary heart disease if they had a history encompassing myocardial infarction, angina pectoris, or undergone coronary artery bypass graft surgery. For a comprehensive investigation of coronary heart disease within the study populace, pertinent datasets were acquired directly from the widely-used Kaggle platform, recognized for hosting and disseminating diverse datasets pertaining to various diseases, including coronary heart disease. The dataset employed in this study is named "Personal Key Indicators of Heart Disease," encompassing a substantial 401,958 rows and 279 columns. The majority of these columns involve queries about respondents' health conditions,

encompassing inquiries like "Do you face substantial challenges in walking or ascending stairs?" or "Have you ever consumed a minimum of 100 cigarettes in your lifetime? [Note: 5 packs = 100 cigarettes]". The dataset encompasses numerous distinct factors or queries that exert a direct or indirect influence on coronary heart disease.

Dataset Description source:

**Table3. 1 Dataset Description source table**

id	Category	Label	Describe	Value
1	CHD	Have CHD or MI	Having previously indicated a history of Coronary Heart Disease (CHD) or myocardial infarction (MI) at any point.	Yes/No
2	HighBP	hypertension	Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional	Yes/No
3	BMI	Computed body mass index	Computed body mass index	Float[1-9999]
4	HighChol	High cholesterol	Having High cholesterol or not	Yes/No
5	Smoker	Smoked at Least 100 Cigarettes	Have you consumed a minimum of 100 cigarettes throughout your lifetime? [Please note: 100 cigarettes are equivalent to 5 packs.]	Yes/No

6	CholCheck	Cholesterol Check	Having cholesterol Check or Not	Yes/No
7	HvyAlcohol Consump	Heavy Alcohol Consumption Calculated Variable	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	-Yes -No
8	Stroke	Ever Diagnosed with a Stroke	(Ever told) (you had) a stroke.	Yes/No
9	PhysHlth	Number of Days Physical Health Not Good	Consider your overall physical well-being, encompassing instances of physical ailments and injuries. Over the previous month, how many days did you experience subpar physical health?	Number of days [1-30]
10	MentHlth	Number of Days Mental Health Not Good	Consider your psychological well-being, encompassing stress, depression, and emotional challenges. How many days out of the last 30 days did you experience poor mental health?	number of days [1-30]

11	DiffWalk	Difficulty Walking or Climbing Stairs	Are you facing significant challenges when it comes to walking or ascending stairs?	Yes/No
12	Sex	Are you male or female?	Are you male or female?	Male/Female
13	Age	Age	Real Age	Age
14	Fruits	Eating Fruit times per day	Eating Fruit 1 or more times per day	Yes/No
15	Diabetic	(Ever told) you had diabetes	Have you ever been informed about having diabetes? In case the answer is 'Yes' and the person responding is female, inquire whether this was specifically during pregnancy. If the respondent mentions pre-diabetes or borderline diabetes, utilize response code 4.	-Yes/No/No , borderline diabetes/Yes (during pregnancy)

16	PhysActivity	Exercise in Past 30 Days	In the previous month, apart from your usual work, did you engage in any physical pursuits or workouts like jogging, bodyweight exercises, golfing, tending to gardens, or purposeful walking?	Yes/No
17	GenHealth	General Health	Do you consider your overall health, on the whole, to be:	Excellent/ Very good/Good/ Fair/Poor
18	Education	Education level	What is the most advanced level of education you have finished?	number of level[1-6]
19	Income	Income level	Does your yearly household earnings encompass all origins of income? (In case the participant declines to provide income information at any level, mark it as "Refused.")	number of degree[1-10]
20	Veggies	Vegetables	Integrate vegetables into your diet at least once a day. Within the last year, did you experience a situation where you required medical attention but refrained due to financial constraints?	Yes/No



21	NoDocbcCost	Needed to see a doctor but could not because of cost	Integrate vegetables into your diet at least once a day. Within the last year, did you experience a situation where you required medical attention but refrained due to financial constraints?	Yes/No
22	AnyHealthcare	Healthcare	Do you possess any form of healthcare protection, encompassing health insurance, prepaid schemes like HMOs, or governmental programs such as Medicare or the Indian Health Service?	Yes/No

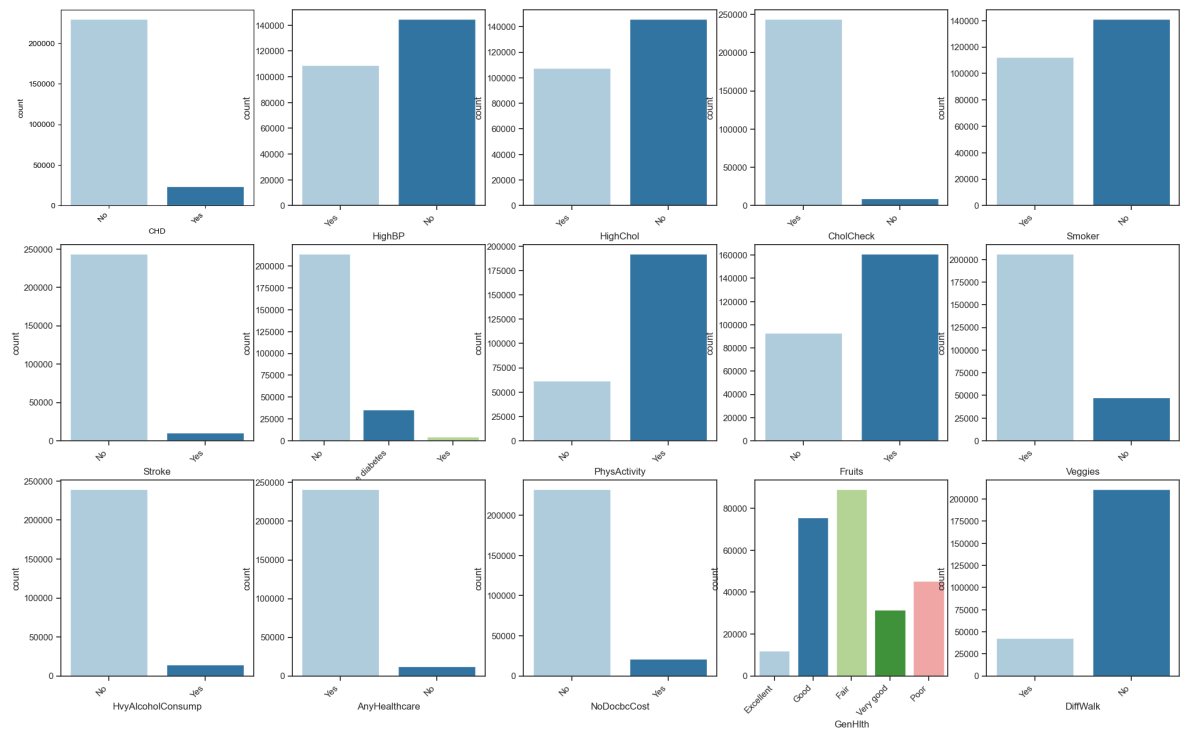
Before this dataset can be studied, it must be cleaned and analyzed. The purpose of cleaning the dataset is to remove any duplicates, missing values, and outliers.

At this stage, duplicate data can be identified by examining identical records, while outliers can be identified using statistical methods (e.g., scatter plots).

For missing data, identification is done by looking at blank or null values in the data set. Two methods for dealing with missing values are generally used for resolution. The first is a complete case study, which involves deleting any records with missing data. However, if the missing data is not random, this approach can lead to a loss of statistical power and may introduce bias. The second is speculation: filling in missing values with estimated values. There are several ways to interpolate missing data, including mean interpolation, regression interpolation, and multiple interpolation. Mean interpolation involves replacing missing values with the mean of the non-missing values for that variable. Regression interpolation includes predicting the missing values using a regression model based on the non-missing data. Multiple interpolations involve creating multiple interpolated datasets and analyzing them separately.

The dataset consists of several lifestyle and health factors associated with the risk of developing coronary heart diseases, such as smoking, alcohol consumption, and BMI, which can be explored by exploring the effects of these factors on the outcome variables. The dataset also contains two physical and mental health indicators, which may provide essential insights into the relationship between overall health and the risk of developing coronary heart disease. The age distribution is divided into three groups, which may limit the analysis of age as a potential risk factor for coronary heart disease. A continuous age variable would provide more information, and Diffwalking may be an essential characteristic to highlight as a possible indicator of mobility and physical activity limitations. There was also a slight bias towards female participants in this dataset. The fact that there were more women than men in the dataset may be due to the higher prevalence of coronary heart disease in men or due to gender differences in participation rates.

### 3.3.2 Exploratory Data Analysis

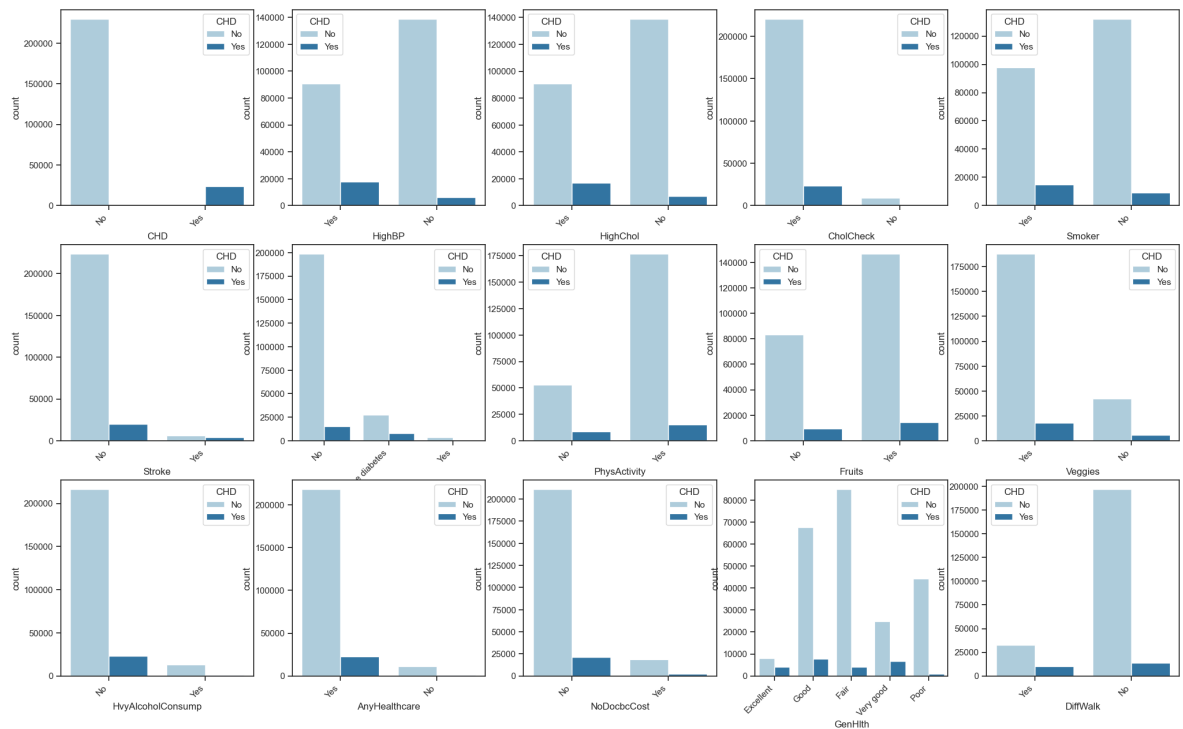


**Figure 3. 3 Summary Statistics For Each Categorical Field Physical And Mental Health Indicators**

Based on several lifestyle and health factors that can predict the risk of developing coronary heart disease (CHD). Out of the total respondents, only 9% reported having CHD or myocardial infarction (MI), while the remaining 91% did not report any such conditions.

High BP (hypertensive), HighChol(high cholesterol), and Body Mass Index (BMI) are essential indicators of CHD risk. The dataset reveals that 41% of respondents have a history of smoking at least 100 cigarettes, and 7% are heavy drinkers. Furthermore, only 4% of respondents reported a history of stroke, a known CHD risk factor. Physical and mental health also play a crucial role in determining CHD risk. The dataset provides information on the number of days respondents reported poor physical and mental health in the past 30 days. 14% of respondents reported having difficulty walking or climbing stairs, and the dataset shows a nearly equal distribution of male and female respondents. The age category of the respondents is also an essential factor to consider, with 11% falling in the age groups of 60-64 and 65-69 and the remaining 78% falling in other age categories.

Overall, the dataset provides valuable insights into the various lifestyle and health factors that can influence the risk of developing CHD. The findings suggest that smoking, heavy drinking, poor physical and mental health, and difficulty walking can increase the risk of CHD. Moreover, High BP (hypertensive), HighChol(high cholesterol), age and BMI are also essential factors to consider when predicting the risk of developing CHD.

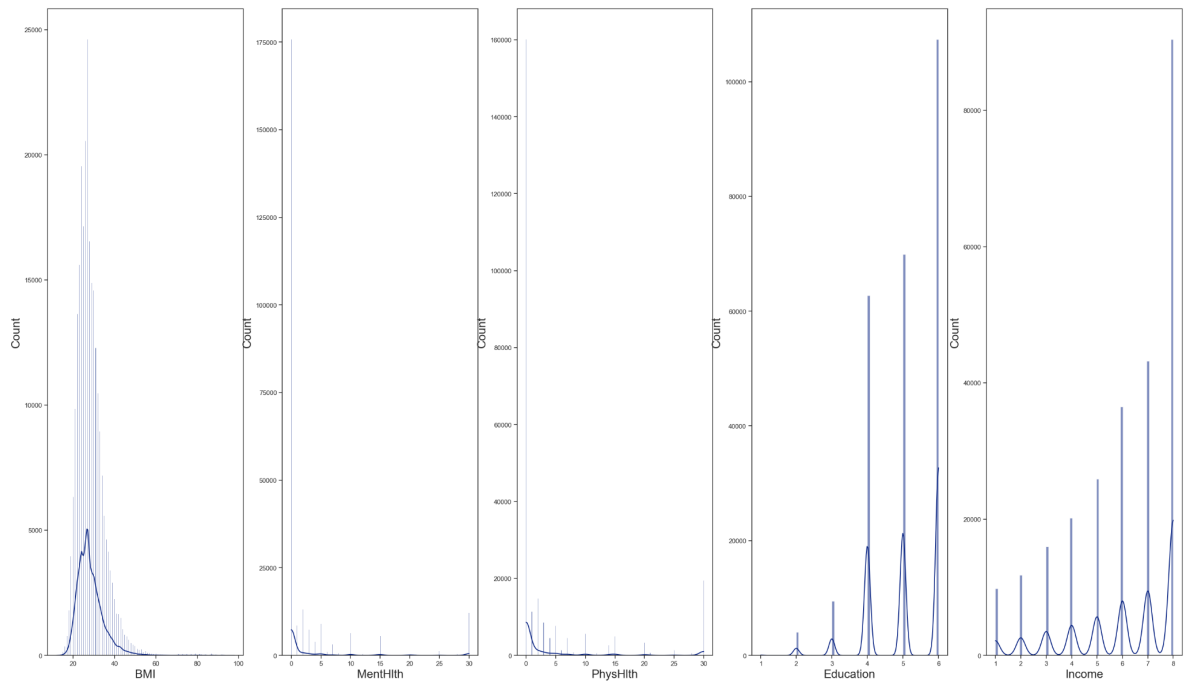


**Figure 3.4 Bar chart of the impact of factors on coronary heart disease**

The given statement discusses the impact of various factors on coronary heart disease, which has been depicted in a bar chart. The chart shows that HighBP(hydertensive) and

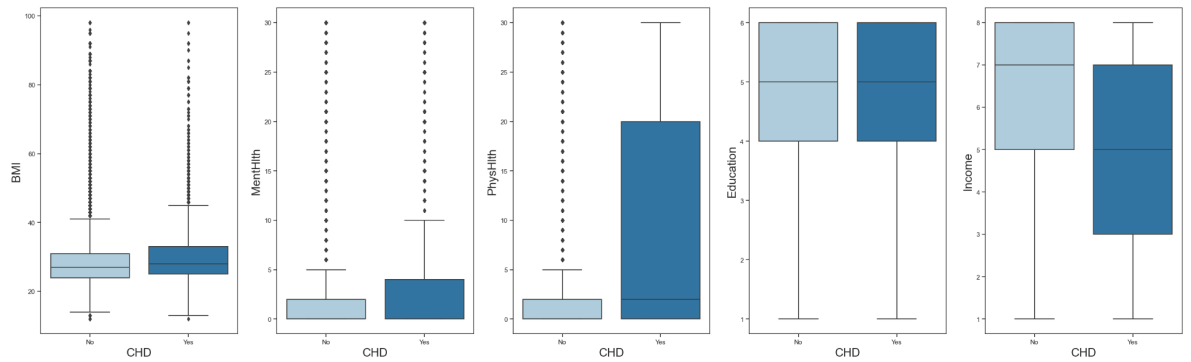
HighChol(High Cholesterol) have the highest impact on coronary heart disease among the given factors. This means that people who engage in normal blood pressure and cholesterol have lower risk of developing coronary heart disease than those who lead a sedentary lifestyle.

The bar chart also indicates that Smoking and Diffusion Walking has the second-highest impact on coronary heart disease. This suggests that individuals who smoke or have difficulty walking are more likely to develop coronary heart disease than those who do not smoke or do not have trouble walking.



**Figure 3.5 Summary statistics for each numerical field in the raw data set.**

According to the data in Figure 3.5, it is evident that people with a body mass index (BMI) of around 25 have the highest representativeness. Many people have a health index of 0, and those with a maximum health index of 30 exhibits a zero value state. From the mental health perspective, the number of people with a mental health index of 0 is the highest, and people with an index of 30 are also in a zero state. Regarding education level, the highest value is concentrated at 6, the lowest at 2, and a value of 1 is almost nonexistent. For income, the number of people with an income level of 8 is the highest, while the number of people with an income level of 1 is the lowest.



**Figure 3. 6 Relationship between each numerical variable in the dataset and the target variable.**

In Figure 3.5, the relationship between coronary heart disease and various numerical factors becomes apparent, shedding light on the key contributors to this condition. One striking observation is that the incidence rate of coronary heart disease is highest within the BMI range of 20 to 40, consistent with the previous analysis involving 25 individuals. Moreover, most incidence and non-incidence rates are concentrated among individuals aged 20 to 40. These trends underscore the importance of weight indicators in determining the risk of coronary heart disease, suggesting that maintaining a healthy BMI is crucial in preventing its onset.

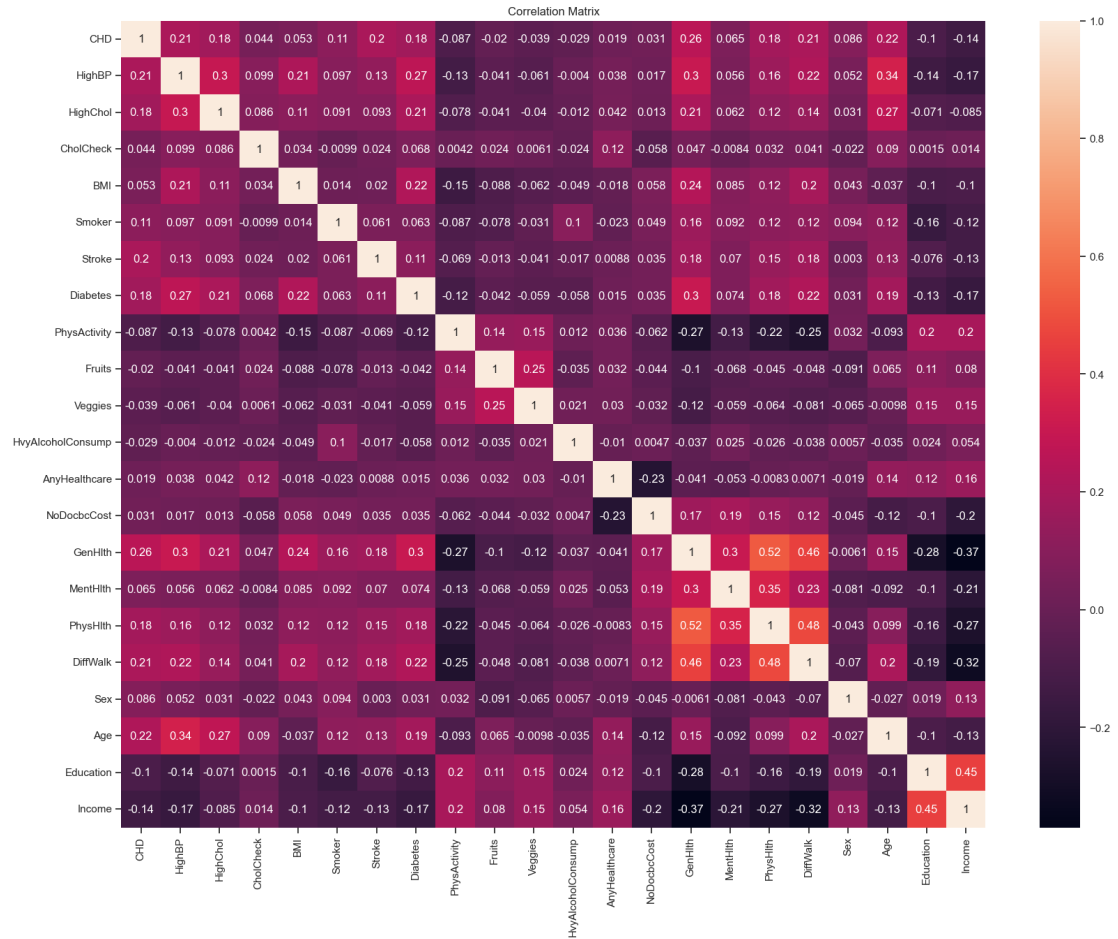
Further analysis of the data reveals compelling associations between mental health and the prevalence of coronary heart disease. Individuals reporting a mental health indicator below 5 exhibits a significantly higher prevalence rate than those without such an indicator, nearly doubling the non-prevalence rate. This finding strongly indicates that mental health plays a vital role in influencing the incidence of coronary heart disease. Addressing mental well-being should be considered an integral component of any comprehensive strategy to reduce the risk of this cardiovascular ailment.

In addition to mental health, physical activity also emerges as a noteworthy factor in coronary heart disease. The incidence rate of individuals engaging in physical exercise for

less than 20 days per month is remarkably high, suggesting a potential link between a sedentary lifestyle and disease development. Therefore, regular physical activity becomes paramount in promoting overall health and preventing coronary heart disease.

The analysis of income levels also yields interesting insights. Notably, a significant disparity is observed in the number of people affected by coronary heart disease in different income brackets. Specifically, individuals with incomes ranging from 3 to 7 experience more than twice the incidence rate of those without illnesses. This highlights the influence of income on health outcomes and emphasizes the need for targeted interventions to address health disparities among different socioeconomic groups.

In conclusion, the findings in Figure 3.5 emphasize the importance of various factors in maintaining overall health and reducing the risk of coronary heart disease. Weight indicators, mental health, income, and physical activity are pivotal in determining the likelihood of developing this condition. As such, public health initiatives should focus on promoting healthy lifestyle choices, addressing mental health concerns, and tackling socioeconomic inequalities to effectively combat coronary heart disease and enhance the well-being of individuals across diverse populations.



**Figure 3. 7 Correlation among features**

Figure 3.7 shows high correlations between HighBP(hypertensive), HighChol(High Cholesterol), Stroke, GenHealth, physical health, Smoking, age, and coronary heart disease. In contrast, Fruit, Vegetable, Education, and Income correlate poorly with Coronary heart disease.

### 3.3.3 Data Preprocessing

#### 3.2.3.1 Data Cleaning

Data cleaning is an important process in data analysis. In the context of the chosen dataset for lifestyle and health factors related to coronary heart disease (CHD), data cleaning becomes particularly important due to the presence of numerous missing values and errors. This process involves identifying and addressing these issues to create a clean dataset that can be used for further analysis. Missing values are filled in using appropriate techniques such as imputation or deletion, while errors are corrected by applying validation

checks and cross-referencing with reliable sources. By performing data cleaning, researchers can enhance the quality of their dataset, minimizing biases and ensuring the integrity of subsequent analyses and conclusions.

### 3.2.3.2 Label Encoding

When working on certain datasets, it was observed that certain features are categorical. If these categorical features are directly inputted into our model, the model will not be able to comprehend the variables associated with those features. It is well-known that machines are incapable of understanding categorical data.

Example:

Original: categorical\_features['Smoking'] = [Yes, No]

Encoded: categorical\_features['Smoking'] = [1, 0]

	CHD	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Physi
0	No	Yes	Yes	Yes	40	Yes	No	No	No	No	...	Yes	No	Excellent		18
1	No	No	No	No	25	Yes	No	No	Yes	No	...	No	Yes	Good		0
2	No	Yes	Yes	Yes	28	No	No	No	No	Yes	...	Yes	Yes	Excellent		30
3	No	Yes	No	Yes	27	No	No	No	Yes	Yes	...	Yes	No	Fair		0
4	No	Yes	Yes	Yes	24	No	No	No	Yes	Yes	...	Yes	No	Fair		3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	No	Yes	Yes	Yes	45	No	No	No	No	Yes	...	Yes	No	Good		0
253676	No	Yes	Yes	Yes	18	No	No	No, borderline diabetes	No	No	...	Yes	No	Very good		0

**Figure 3. 8 Original dataset with label encoding Yes or No**

HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	...	Education	Income	Diabetes_No	Diabetes_bordi dial
1	1	1	40	1	0	0	0	1	0	...	4	3	1	
0	0	0	25	1	0	1	0	0	0	...	6	1	1	
1	1	1	28	0	0	0	1	0	0	...	4	8	1	
1	0	1	27	0	0	1	1	1	0	...	3	6	1	
1	1	1	24	0	0	1	1	1	0	...	5	4	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

**Figure 3. 9 Standardized dataset with label encoding 1 or 0**

### 3.2.3.3 Strongly Risk Factors Identify And Dataset Regeneration

Based on the results of Huang AA and Huang SY's (2023) article, the machine learning model XGBoost is a highly accurate algorithm for identifying feature importance relationships between CHD and other health factors, and its predictive accuracy in healthcare prediction has also been improved (Huang AA, Huang SY, 2023). And this algorithm is also the most commonly used machine learning algorithms for identifying



feature importance relationships between diseases and other health factors (Zhong, X. et al., 2022). So this study will utilize XGBoost machine algorithms for implementation.

```
import pandas as pd
from xgboost import XGBClassifier

X = df.drop('CHD', axis=1) # Features
y = df['CHD'] # Target variable

clf = XGBClassifier()
clf.fit(X, y)

importance_dict = clf.get_booster().get_score(importance_type='total_gain')
total_gain_sum = sum(importance_dict.values())
importance_dict1 = clf.get_booster().get_score(importance_type='total_cover')
total_cover_sum = sum(importance_dict1.values())
importance_dict2 = clf.get_booster().get_score(importance_type='weight')
total_weight_sum = sum(importance_dict2.values())

importances_df = pd.DataFrame(importance_dict.items(), columns=['Feature', 'Gain'])
importances_df['Weight'] = importances_df['Feature'].apply(lambda x: clf.get_booster().get_score().get(x, 0))
importances_df['Cover'] = importances_df['Feature'].apply(lambda x: clf.get_booster().get_score(importance_type='cover').get(x, 0))

# Convert importance values to percentages
importances_df['Gain'] = (importances_df['Gain'] / total_gain_sum) * 100
importances_df['Weight'] = (importances_df['Weight'] / total_weight_sum) * 100
importances_df['Cover'] = (importances_df['Cover'] / total_cover_sum) * 100

importances_df.sort_values('Gain', ascending=False, inplace=True)
importances_df.reset_index(drop=True, inplace=True)

average_importances = importances_df.mean()

print(importances_df)
print(average_importances)
```

**Figure 3. 10 Using XGBoost to calculate the correlation between CHD and other factors**

Figure 3.10 shows the Python code for calculating the correlation between CHD and other factors using the XGBoost machine learning model. Based on this code, XGBoost's "feature\_importances\_" was used to calculate the important eigenvalue correlation coefficient (Gain, Cover, Weight) between CHD and other factors in this study (Chen, T., 2016).

**Cover:** Cover refers to the average coverage of a feature across the trees in a model. It measures the frequency of a feature being used to split the data. Higher cover values indicate that the feature has a broader influence on the model's decisions (Chen, T., 2016). For example, in predicting coronary heart disease (CHD) based on lifestyle and health factors, if a feature related to smoking has a high cover value, it suggests that smoking has a significant impact on CHD predictions. Below are detailed explanations of Gain, Cover, and Weight.

**Gain:** Gain represents the average gain of a feature when it is used for splitting in the model's trees. It measures the improvement in the model's objective function attributed to that specific feature. Higher gain values indicate that the feature contributes more to enhancing the model's performance (Chen, T., 2016). For instance, if a feature related to exercise has a high gain value, it implies that exercise is a crucial factor in predicting CHD accurately.

Weight: Weight denotes the number of times a feature is selected as a split criterion across all the trees in the model. It represents the occurrence frequency of a feature. Higher weight values suggest that the feature is frequently used for decision-making within the model (Chen, T., 2016). For example, if a feature related to cholesterol levels has a high weight value, it indicates that cholesterol levels are frequently considered when predicting CHD.

At the same time, to reduce errors, this study also set up to execute the XGBoost algorithm calculate feature importance (Gain, Cover, Weight). Then it used the mean function to calculate the average of three feature importance values obtained.

In addition, this study also utilized Matplotlib in Python to visualize correlation coefficients.

```
time=importances_df.transpose()
time['average']=average_importances.transpose()
time.sort_values('average', ascending=False, inplace=True)
time
```

**Figure 3. 11 Code of Combining the average value with the feature importance value obtained three times**

```
time_xg['Factors'].str.split('_').str[0]
# Split the 'Factors' column by '_'
time_xg[['Factors', 'Subcategory']] = time_xg['Factors'].str.split('_', n=1, expand=True)

# Group by the 'Factors' column and calculate the mean for 'First', 'Second', 'Third', and 'Average'
time_xg = time_xg.groupby('Factors').mean().reset_index()

# Reorder the columns
xg=time_xg[['Factors', 'First', 'Second', 'Third', 'Average']]
xg
```

**Figure 3. 12 Code of Merge Feature Factors**

Figures 3.11 and Figure 3.12 show the code used to calculate the important features between coronary heart disease and other factors and to combine and merge this important special data. This code provides accurate data on important features between coronary heart disease and other factors. Here are further operations on the data.

```
# Create the bar plot
plt.figure(figsize=(10, 6))
plt.bar(df_sorted[df_sorted['Average']>=0.01]['Factors'], df_sorted[df_sorted['Average']>=0.01]['Average'], color='red')
plt.xticks(rotation=45)
plt.xlabel('Factors')
plt.ylabel('Average')
plt.title('Top Factors (Features Importance>0.01)')
plt.tight_layout()
# Show the plot
plt.show()
```

**Figure 3. 13 Code of Visualization of top factors that associated with CHD**

```
df=pd.read_csv('CHD.csv')
selected_columns = ['CHD', 'DiffWalking', 'Stroke', 'GenHealth', 'AgeCategory', 'Sex', 'Diabetic', 'Smoking', 'KidneyDisease']
newdataset=df[selected_columns]
newdataset.to_csv('CHDNew.csv', index=False)
```

**Figure 3. 14 Code of Generate a new dataset based on identified factors**

Figures 3.13 and 3.14 show the specific process of identifying lifestyle and health factors closely related to the development of coronary heart disease and generating a new dataset based on the identified lifestyle and health factors.

And then repeat the 3.2.3.1 step.

#### **3.2.3.4 Partition Of Test Set and Training Set**

In the realm of machine learning, data sets are commonly bifurcated into two distinct subsets: the training set and the test set. The primary function of the training set involves instructing the machine learning model, whereas the test set serves the purpose of gauging the model's effectiveness and its capacity to apply acquired knowledge to new data. This partition entails allocating 20% of the authentic data to the test set, while the remaining 80% constitutes the training set, maintaining the integrity of the initial dataset.

```
# import train_test_split from sklearn.model_selection
from sklearn.model_selection import train_test_split
# split the dataset into train and test set with 80% and 20% respectively
train_data, test_data = train_test_split(df, train_size=0.80)
train_data.shape, test_data.shape

((255836, 51), (63959, 51))
```

**Figure 3. 15 Split the dataset into train set and test set**

#### **3.2.3.5 Handling Imbalanced Data**

A Balanced Dataset is characterized by a near-even distribution of classes within the target column, while an Imbalanced Dataset exhibits a significant imbalance in class distribution within the target column. The challenge posed by imbalanced datasets lies in the potential bias of the model towards the majority class, resulting in suboptimal performance in predicting the minority class. Effectively addressing this requires focused strategies and techniques, including resampling approaches like oversampling or undersampling, the adoption of suitable evaluation metrics such as precision, recall, or F1 score, and the utilization of advanced algorithms explicitly tailored for imbalanced data,

such as SMOTE, ADASYN, or algorithms designed for handling the costs associated with imbalanced distributions.

It is important to recognize the nature of the dataset (balanced or imbalanced) and apply appropriate strategies to address the specific challenges associated with each type during the machine learning process.

```
1 df['CHD'].value_counts()
No      229787
Yes      23893
Name: CHD, dtype: int64
```

**Figure 3. 16 Target values Rate in Dataset**

Referring to Figure 3.16, it's clear that the dataset displays an uneven distribution among its classes. 'No' instances are prominent at 292,787, whereas 'Yes' instances are notably limited at 23,893. This considerable disparity points to an imbalanced dataset, necessitating specialized techniques and attention to confront the predicaments linked with imbalanced data in machine learning. One approach involves under-sampling, wherein the objective is to rectify this class imbalance by replicating instances from the minority class to achieve a balanced sample size comparable to the majority class. To illustrate, imagine an imbalanced training dataset initially containing 1,000 records.

Before under-sampling:

- Target class 'Yes' has 900 records.
- Target class 'No' has 100 records.

After under-sampling:

- Target class 'Yes' has 900 records.
- Target class 'No' has 900 records.

Following the under-sampling technique, both classes now have an equal sample size.

Using 'SMOTE' for over\_sampling and 'NearMiss' for under\_sampling.

```
1 print('Original: {}'.format(Counter(y_train)))
2 print('SMOTE: {}'.format(Counter(y_train_smote)))

Original: Counter({0: 233908, 1: 21928})
SMOTE: Counter({0: 233908, 1: 233908})
```

**Figure 3. 17 Balance the dataset using Smote**

### 3.2.3.6 K-Fold Cross Validation

K-Fold Cross-Validation involves partitioning a given dataset into K folds, where each fold serves as the testing set during various stages of evaluation. For instance, consider a scenario with 5-Fold Cross-Validation ( $K=5$ ), where the dataset is evenly split into five folds. Throughout evaluation:

In the first round, the initial fold acts as the testing set, and the other four folds form the training set.

Following this, the second fold becomes the testing set in the second round, and the remaining folds constitute the training set.

This sequence continues, assigning a distinct fold as the testing set in each subsequent iteration, while the rest serve as training data.

This cycle persists until all five folds have been used as the testing set once.

In this study, a 10-Fold cross-validation approach will be employed ( $K=10$ ).

```
# kfold cross validation
from sklearn.model_selection import KFold

# make a 10 fold cross validation
cv = KFold(n_splits=10, random_state=None, shuffle=False)
```

**Figure 3. 18 10-Fold Cross-Validation**

### 3.3.4 Modeling

The central objective of this research revolves around harnessing diverse machine learning methodologies to anticipate the likelihood of coronary heart disease. Drawing from the analysis of existing literature, AdaBoost, Random Forest, Decision Tree, KNN, and Naive Bayes emerge as the foremost machine learning techniques for prognosticating coronary heart disease. Consequently, these five algorithms will be selected in this investigation to construct a comprehensive machine learning framework, thereby determining the optimal algorithm for forecasting CHD risk. Presented below are the operational principles and elucidations of each of these algorithms:

#### **AdaBoost (Adaptive Boosting):**

AdaBoost constitutes an ensemble learning technique that amalgamates numerous feeble classifiers to formulate a potent classifier. The methodology involves sequential training of feeble classifiers on distinct subsets of data, assigning greater significance to misclassified instances. The ultimate forecast stems from the accumulation of predictions from these

feeble classifiers, with escalated emphasis on those demonstrating heightened accuracy (Freund Y. et al., 2021). For prognosticating coronary heart disease (CHD) susceptibility, AdaBoost can be instructed employing lifestyle and health determinants encompassing variables like smoking habits, body mass index (BMI), physical exertion, cholesterol levels, among others. Each frail classifier might specialize in distinct attributes or trends, such as the repercussions of smoking or the influence of BMI on CHD risk assessment. By fusing the forecasts of these feeble classifiers, AdaBoost can present a holistic risk evaluation for an individual grounded in their lifestyle and health constituents.

### **Random Forest:**

At its core, the fundamental principle of Random Forest involves an ensemble learning technique. This algorithm creates multiple decision trees, amalgamating their predictions into a final conclusive prediction. With every decision tree being trained on a distinct random subset of data and contemplating a different set of features for each division, the ultimate prediction is reached through either majority voting or an average of these tree predictions, as elucidated by Breiman in 2020. For instance, in the context of assessing the risk of Coronary Heart Disease (CHD), the Random Forest method proves valuable. By instructing decision trees using lifestyle and health attributes, each tree can factor in varying combinations of traits like age, blood pressure, diabetes presence, physical activity level, and more. Through the confluence of individual decision tree forecasts, Random Forest furnishes a more resilient and precise CHD risk evaluation.

### **Decision Tree:**

Decision Tree operates on a foundational principle within predictive modeling. It constructs a tree-shaped model to facilitate decision-making, relying on the attributes of features. This algorithm recursively segments data through inquiries grounded in feature parameters, ultimately assigning class identifiers to end nodes. The decision process hinges on a series of conditional rules, gleaned from the dataset (Quinlan et al., 2021). To illustrate, envision the utilization of a Decision Tree in gauging the likelihood of coronary heart disease (CHD) risk. By crafting a tree that bifurcates into realms of lifestyle and health attributes, inquiries like "Does the individual smoke?" or "Does their BMI surpass a certain threshold?" might arise. By tracing the branches guided by responses to these queries, the Decision Tree effectively categorizes individuals' risk levels, predicated upon their distinctive lifestyle and health elements.

### **K-Nearest Neighbors (KNN):**

K-Nearest Neighbors (KNN) technique operates on the premise of similarity within the feature space. It's a non-parametric approach that categorizes data points by gauging their resemblance to nearby instances. This involves assessing a new data point's class label through the prevailing classification of its K closest neighbors. Here, K represents a user-defined parameter (Cover et al., 2020). For instance, when forecasting the risk of Coronary Heart Disease (CHD), KNN finds utility by evaluating how much a new individual's lifestyle and health attributes align with those of its K nearest counterparts within the training dataset. Noteworthy attributes for consideration encompass factors like smoking habits, Body Mass Index (BMI), physical activity levels, and cholesterol readings, among others. The technique capitalizes on identifying the K most akin individuals in the training dataset and, consequently, approximates the CHD risk for the new individual based on the prevalent class observed among these immediate neighbors.

#### **Naive Bayes:**

Naive Bayes functions as a probabilistic classifier grounded in Bayes' theorem, presuming self-reliance among attributes. It computes the probabilities of classification tags in relation to attribute values and designates the most likely classification tag to a fresh occurrence (Rish, I., 2019). As an illustration, Naive Bayes has the capability to anticipate the likelihood of coronary heart disease (CHD) by evaluating probabilities linked to distinct risk tiers derived from lifestyle and health variables. It takes into account the conditional probabilities of varied attributes like smoking, BMI, level of physical activity, cholesterol levels, and more. Through approximating the probabilities of diverse risk tiers considering these attributes, Naive Bayes assigns the most probable risk tier to an individual grounded in their lifestyle and health components.

By employing these species of machine learning algorithms, this study aims to leverage their respective strengths and capabilities in predicting the risk of coronary heart disease. Each algorithm brings its unique approach to the task, and by comparing their performances, insights can be gained into the most effective algorithms for this particular predictive task.

### 3.3.5 Developing Web Application

Developing a Web application based on lifestyle and health factors to predict coronary heart disease risk is very important for our life. The main steps of developing this web application are as follows:

Setting up the development environment includes installing the tools and libraries required for web application development. The application development will use Python and Dash (the framework used to build web applications) to write code.

Developing web application code: This development will use Python and Dash frameworks to write code for the web application. This code will include necessary user interface elements, data processing logic, and machine learning model integration.

Integrated machine learning model: Integrate the machine learning model that evaluates the best coronary heart disease risk prediction obtained in objective 2. The model will be integrated into our Web application code, allowing users to input their lifestyle and health factors and obtain risk predictions. As shown in Figure 3.19.

```
user_data = pd.DataFrame({
    'HighBP': [1 if highbp=='Yes' else 0],
    'BMI': [int(weight/(height*height))],
    'Stroke': [1 if Stroke=='Yes' else 0],
    'Age': [age],
    'Sex': [1 if gender=='MEN' else 0],
    'HighChol': [1 if highchol=='Yes' else 0],
    'DiffWalk': [1 if diffwalk=='Yes' else 0],
    'Smoker': [1 if smoking=='Yes' else 0],
    'MentHlth': [mental],
    'Income': [income_category],
    'GenHlth_Excellent': [1 if genhealth == 'Excellent' else 0],
    'GenHlth_Fair': [1 if genhealth == 'Fair' else 0],
    'GenHlth_Good': [1 if genhealth == 'Good' else 0],
    'GenHlth_Poor': [1 if genhealth == 'Poor' else 0],
    'GenHlth_Very good': [1 if genhealth == 'Very good' else 0],
    'Diabetes_No': [1 if diabetic == 'No' else 0],
    'Diabetes_No, borderline diabetes': [1 if diabetic == 'No, borderline diabetes' else 0],
    'Diabetes_Yes': [1 if diabetic == 'Yes' else 0]
})

# Make the prediction using the trained model
prediction = ada_model.predict(user_data)[0]
if prediction == 1:
    result =html.P('Have Risk.', style={"color": "red"})
else:
    result =html.P('No Have Risk.', style={"color": "green"})
return html.Div(result)
```

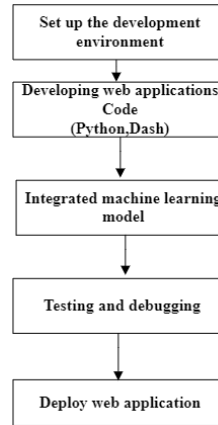
**Figure 3. 19 Integrated machine learning model**

Testing and debugging: It is important to thoroughly test the web application to ensure it runs as expected. Moreover, various scenarios will be simulated, and different values will be input to verify the accuracy and reliability of risk prediction. Any errors or issues encountered during the testing process will be debugged and resolved.

Deploying web applications to Python Anywhere: Python Anywhere is a platform that allows all users to host and deploy Python web applications. We will deploy the web



application we have developed on Python Anywhere, allowing users to access it through the Internet. This step includes configuring necessary settings, such as setting up server and database connections, to ensure the smooth operation of the application. As shown in Figure 3.20.



**Figure 3. 20 Develop Web Application**

This application can help individuals assess risks and take necessary preventive measures to maintain a healthy lifestyle.

### **3.3.6 Testing And Evaluating Web Application**

After developing the web application for predicting the risk of coronary heart disease based on lifestyle and health factors, the next crucial step is testing and evaluating its functionality and performance. This ensures that the application is reliable, user-friendly, and capable of delivering accurate predictions. The following testing and evaluation steps can be conducted:

**Unit testing:** Unit testing is performed to verify the correctness of individual components and functions within the web application. It involves writing test cases that cover various scenarios and inputs and then executing these tests to ensure the expected behavior of each unit. In the context of this application, unit testing would involve testing functions responsible for data processing, model prediction, and user interface interactions. By conducting unit testing with a small group of participants (around 5), we can identify and address any issues or discrepancies in the application's behavior (Noor S. et al., 2019).

**Data validation:** Validating user input data is critical for ensuring the accuracy and reliability of the predictions. The web application should implement robust data validation mechanisms to check for input errors, missing values, or incorrect formats. This step involves verifying that the input data adheres to the expected criteria and ranges. By

performing thorough data validation, we can minimize the potential for erroneous predictions due to incorrect or unreliable input.

Cross-browser and cross-device testing: It is important to test the web application across different browsers (such as Chrome, Firefox, Safari, and Edge) and devices (desktop, tablets, and mobile phones). This ensures that the application functions correctly and displays appropriately across various platforms, operating systems, and screen sizes. By conducting cross-browser and cross-device testing, we can identify and address any compatibility issues or inconsistencies in the application's appearance and functionality.

Error handling and logging: Implementing robust error handling mechanisms is crucial for providing a smooth user experience and identifying issues within the application. The web application should handle errors gracefully by displaying informative error messages to users when unexpected situations occur. Additionally, logging errors and exceptions that occur during the application's runtime allows for efficient troubleshooting and debugging. Proper error handling and logging mechanisms enable us to identify and resolve issues promptly, ensuring the application's stability and reliability (Meissner, J., 2018).

By thoroughly testing and evaluating the web application through unit testing, data validation, cross-browser and cross-device testing, and implementing effective error handling and logging mechanisms, we can enhance its quality, accuracy, and user experience. These steps help in identifying and resolving potential issues before deploying the application to production, ensuring that users receive reliable risk predictions for coronary heart disease based on their lifestyle and health factors.

## CHAPTER 4:RESULT

### 4.1 Identify The Lifestyle And Health Factors

#### 4.1.1 Feature Importance Statistics And Analysis

Feature Importance refers to the utilization of metrics such as Gain, Cover, and Weight to calculate scores for each input feature in a given model. These scores indicate the relative "importance" of each feature. A higher score signifies that the corresponding feature has a greater impact on the predictive capabilities of the model, specifically in relation to the target variable. By considering the average values of Cover, Gain, and Weight, a comprehensive assessment of feature importance can be obtained. Below is a specific explanation of these three values.

Table 5.1, a model constructed using the XGBoost machine learning algorithm is proposed to study the important feature relationships between coronary heart disease and other lifestyle and health factors. In order to reduce errors, this model was used to obtain the percentage relationship of Gain, Cover, and weight values between 'CHD' and other elements, and their corresponding average values were calculated based on these three values. Thus, an overall CHD and other factors analysis will be conducted based on the final average score obtained.

**Table 4. 1 XGBoost model feature importance statistics.**

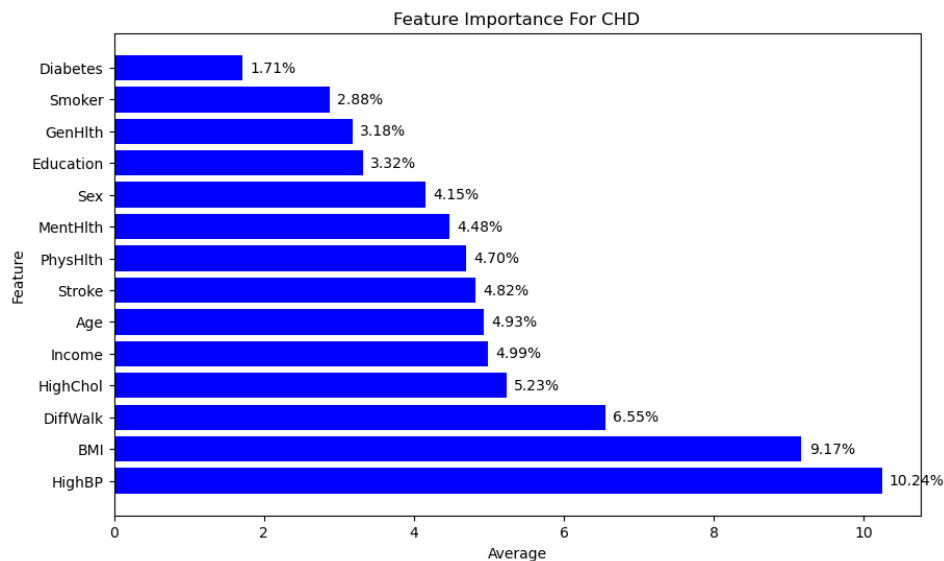
ID	Feature	Gain	Weight	Cover	Average
1	HighBP	22.40%	2.46%	5.86%	10.24%
2	BMI	4.86%	18.04%	4.60%	9.17%
3	DiffWalk	12.85%	2.83%	3.97%	6.55%
4	HighChol	8.09%	2.57%	5.04%	5.23%
5	Income	3.92%	8.84%	2.21%	4.99%

6	Age	2.59%	10.59%	1.61%	4.93%
7	Stroke	7.83%	2.79%	3.84%	4.82%
8	PhysHlth	2.91%	9.09%	2.09%	4.70%
9	MentHlth	2.45%	8.11%	2.88%	4.48%
10	Sex	5.64%	3.48%	3.35%	4.15%
11	Education	1.54%	6.19%	2.22%	3.32%
12	GenHlth	2.89%	1.29%	5.37%	3.18%
13	Smoker	3.02%	3.11%	2.51%	2.88%
14	CholCheck	0.60%	0.92%	6.60%	2.71%
15	HvyAlcoholConsump	0.48%	0.85%	6.23%	2.52%
16	Diabetes	1.22%	1.40%	2.53%	1.71%
17	AnyHealthcare	0.61%	1.24%	3.11%	1.65%
18	NoDocbcCost	0.47%	1.77%	2.31%	1.52%
19	Fruits	0.74%	2.38%	1.43%	1.52%
20	PhysActivity	0.46%	2.01%	1.58%	1.35%
21	Veggies	0.45%	2.05%	1.50%	1.34%

Based on the analysis of Table 4.1, we observe that several lifestyle and health factors exhibit significant importance in relation to coronary heart disease (CHD) prediction. Notably, HighBP, BMI, DiffWalk, HighChol, Income, Age, Stroke, PhysHlth, MentHlth, and Sex emerge as the top influential features associated with CHD. These factors demonstrate substantial gain and weight values, indicating their strong impact on the model's predictive capabilities.

However, it is worth noting that there are certain features for which the gain or weight value falls below the 1% threshold. This implies that these particular features may have a relatively lower contribution to CHD prediction compared to the aforementioned influential factors. It is essential to conduct further research and analysis to better comprehend the relationship between these features and CHD. Additional investigations can help ascertain whether these features possess hidden connections or if their impact on CHD prediction requires more nuanced exploration.

Below, further analysis will be conducted using visualization and further narrowing of gain and weight values. As shown in Figure 4.1.



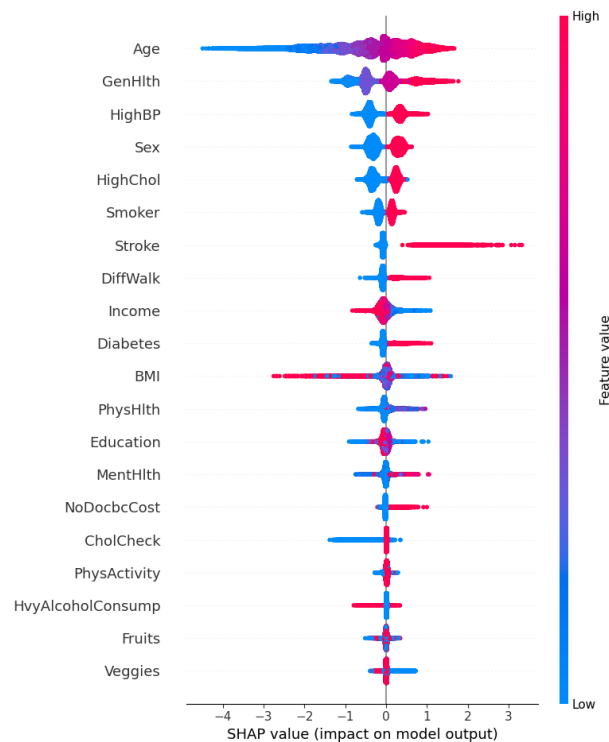
**Figure 4. 1 Bar chart of the degree of correlation between lifestyle and health factors related to coronary heart disease(Weight>=1%,Gain>=1%and Cover>=1%)**

From Figure 4.1, it can be seen HighBP, BMI, DiffWalk, HighChol, Income, Age, Stroke, PhysHlth, MentHlth, Sex, Education, GenHlth, Smoker, Diabetes is very

important factors for CHD. Therefore, through the analysis of feature importance, it is found that these 15 factors are relatively important for CHD.

#### 4.1.2 Shape Figure Analysis

In SHAP explanations, elevated covariate values are depicted in red, while lower values are shown in blue. The X-axis illustrates the shift in log-odds for coronary heart disease (CHD).



**Figure 4. 2 Overall SHAP explanations**

Based on Figure 4.2, this graph intuitively displays the importance of lifestyle and health factors with the strongest correlation with coronary heart disease. From the graph, it can be seen that Age, GenHlth(Health), highBP(hypertension), Sex, Income, HighChol(High Cholesterol), Smoker, Stroke, DiffWalk(Difficult Walking), Diabetes, BMI, MentHlth(Mental Health), NoDocbcCost(Needed to see a doctor but could not because of cost), HvyAlcoholConsump(Heavy Alcohol Consumption) have the highest importance for the characteristics of coronary heart disease, indicating that these factors are closely related to coronary heart disease.

### 4.1.3 Summary

Based on the examination of Table 5.1 and Figures 5.1 and 5.2, a comprehensive analysis reveals several important factors that contribute to the risk of coronary heart disease (CHD). These factors include HighBP, BMI, DiffWalk, HighChol, Income, Age, Stroke, PhysHlth(Physical Health), MentHlth(Mental Health), Sex(Gender), GenHlth(General Health), Smoker, Diabetes, Education, NoDocbcCost (Needed to see a doctor but could not be because of cost), and HvyAlcoholConsump (Heavy Alcohol Consumption).

However, to obtain a more definitive understanding of their significance, it is necessary to perform a detailed analysis that takes into account both feature importance and the shape figure. By finding the intersection of these two components, the final set of crucial factors for CHD can be determined.

Based on this comprehensive analysis, it can be concluded that HighBP(hypertension), BMI, DiffWalk(Difficult Walking), HighChol(High Cholesterol), Income, Age, Stroke, MentHlth(Mental Health), Sex, GenHlth(Health), Smoker, and Diabetes emerge as the most important factors in relation to CHD. These factors, which exhibit significant feature importance and align with the patterns observed in the shape figure analysis, play a substantial role in influencing the risk of developing CHD.

By considering both the feature importance analysis and the shape figure, researchers can obtain a more reliable and comprehensive understanding of the factors that have the greatest impact on CHD risk. This knowledge is valuable for informing preventive measures, interventions, and further research in the field of cardiovascular health.

## 4.2 Comparison Of Model Performance

The Confusion Matrix serves as a tool to gauge the effectiveness of machine learning models in classification tasks. Its purpose is to assess how well the model's predictions align with the actual labels in the dataset. This matrix, presented as a 2x2 table, encapsulates diverse scenarios of predicted and actual labels. The breakdown of this matrix includes:

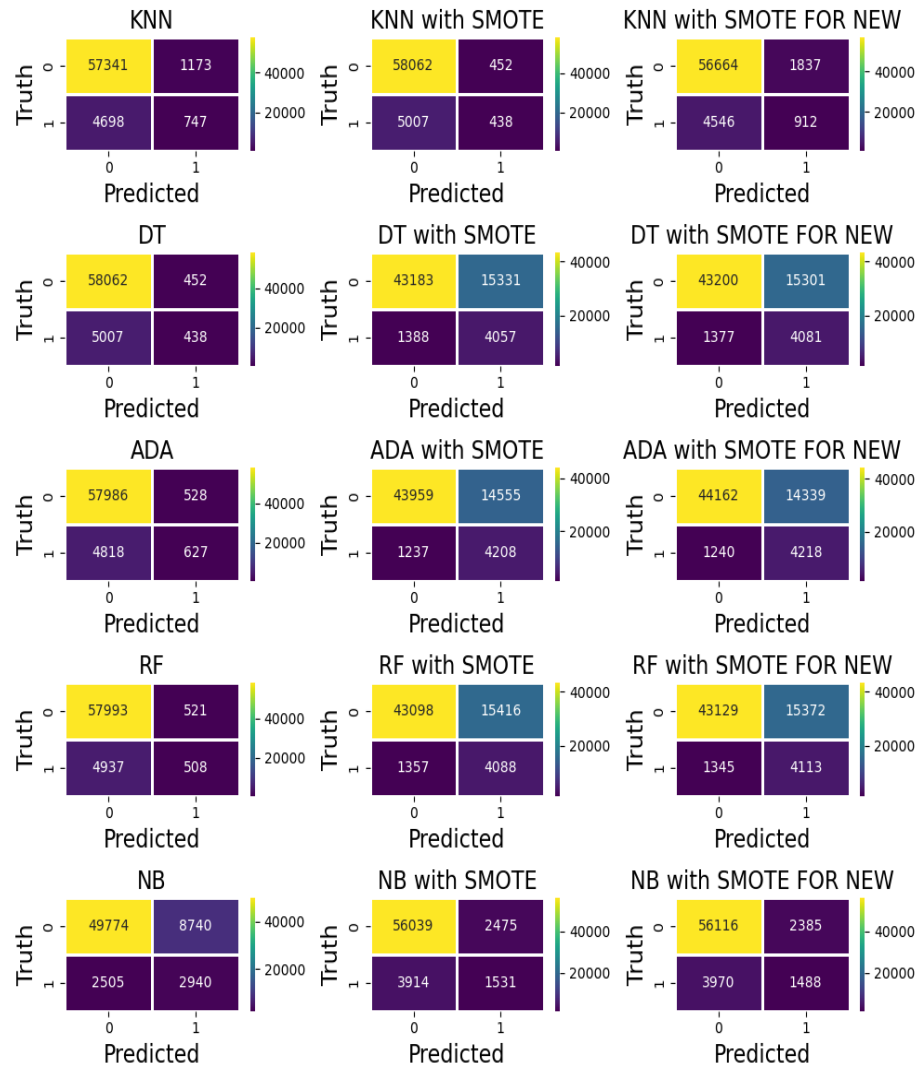
True Positive (TP): Instances accurately predicted as positive (CHD=Yes).

True Negative (TN): Cases correctly predicted as negative (CHD=No).

False Positive (FP): Instances wrongly predicted as positive (CHD=Yes) when they're negative.

False Negative (FN): Instances erroneously predicted as negative (CHD=No) when they're positive.

Illustrated in Figure 4.3 is a visual representation of the Confusion Matrix used in this study.



**Figure 4.3 Confusion Matrix**

The Confusion matrix allows the calculation of various performance indicators, such as accuracy, precision, recall rate (also known as sensitivity or true positive rate), and F1 score. These metrics provide a deeper understanding of model performance and its ability to classify instances correctly.

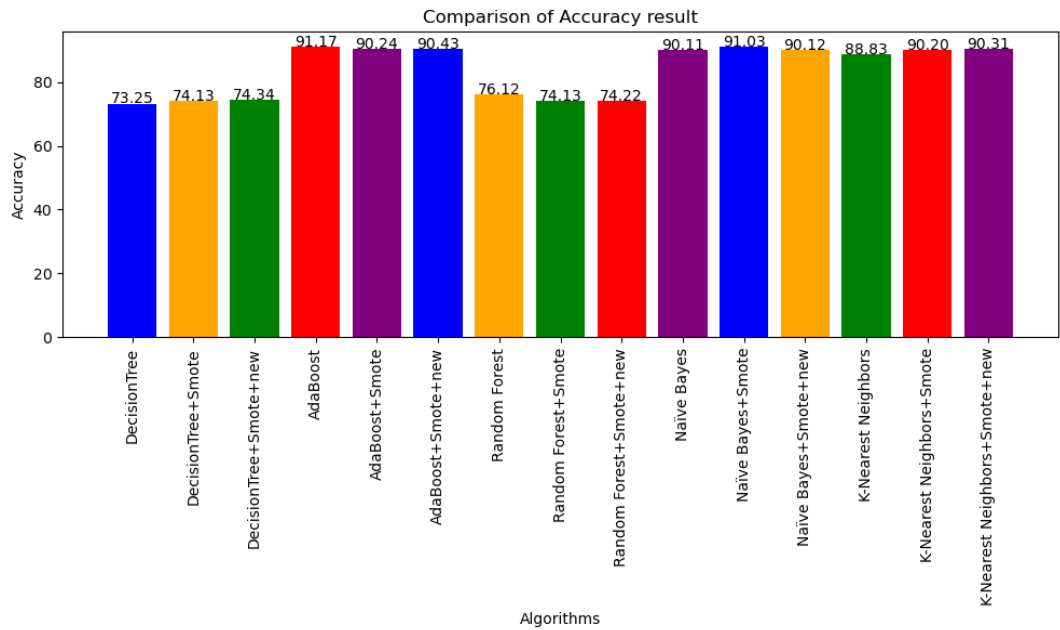
#### 4.2.1 Accuracy

Accuracy is a common performance metric used to evaluate classification models. It measures the overall correctness of the model's predictions. The formula for accuracy using the Confusion Matrix is:



$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

The top part (numerator) of the equation, comprising the sum of true positives (TP) and true negatives (TN), embodies the overall correct predictions achieved by the model. Meanwhile, the bottom part (denominator), consisting of true positives, true negatives, false positives (FP), and false negatives (FN), encompasses the complete instances within the dataset. The accuracy metric spans from 0% to 100%, with 100% denoting flawless precision (every prediction is accurate) and 0 signifying an absence of accurate predictions (James, G. et al., 2021).



**Figure 4. 4 Accuracy For Different Model's Performance**

Based on the accuracy results of different models shown in Figure 4.4, we can analyze the performance of the models in terms of prediction accuracy.

Firstly, examine the model without applying any additional techniques. The decision tree algorithm achieved an accuracy of 73.25%, indicating a moderate level of performance. However, when applying SMOTE technology to handle class imbalance (DecisionTree+SMOTE), the accuracy slightly improved to 74.13%. In addition, the accuracy was improved by combining with SMOTE for the second prediction (Decision\_Tree+SMOTE+new) to 74.34%. These improvements show that solving the class imbalance problem can have a positive impact on the performance of the decision Tree model.

Turning to the AdaBoost algorithm, we observed a significantly higher accuracy of 91.17%. The result indicates that AdaBooster is effective in accurately predicting the risk

of coronary heart disease. Similar to DecisionTree, the application of SMOTE (AdaBoost+SMOTE) and the second run (AdaBoost+SMOTE+new) slightly improved the accuracy, reaching 90.24% and 90.43%, respectively. These results indicate that compared to DecisionTree, AdaBoost did not achieve significant performance improvements from SMOTE and other modifications.

For Random forest, the accuracy rate without any additional technology (76.12%) is relatively high compared with the decision tree. However, when SMOTE (Random forest+smoke) is applied with other modifications (Random forest+smoke+new), the accuracy rate decreases to 74.13% and 74.22%, respectively. The result indicates that SMOTE and other modifications may benefit Random forests less than decision trees.

Regarding naive Bayes, the initial accuracy is 90.11%, which is relatively high. When applying SMOTE (Naive Bayes+SMOTE) with further modifications (Naive Bayes+SMOTE+new), the accuracy remained high, at 91.03% and 90.12%, respectively. These results indicate that naive Bayes performs well in predicting the risk of coronary heart disease, and additional techniques do not significantly affect its accuracy.

Finally, the accuracy of K-nearest neighbors reached 88.83%. Applying SMOTE (K-nearest neighbor+SMOTE) and additional modifications (K-nearest neighbor+SMOTE+new) improved the accuracy by 90.20% and 90.31%, respectively. These improvements demonstrate the effectiveness of SMOTE and other modifications in enhancing the performance of the K-nearest neighbor model.

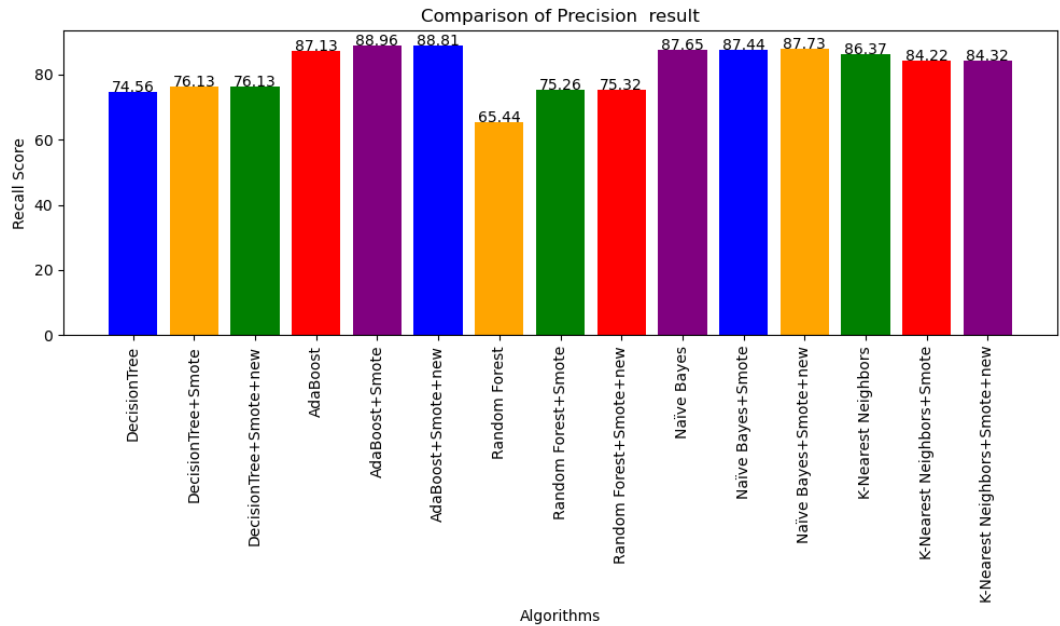
The analysis shows that AdaBoost consistently achieves the highest accuracy in the evaluated model. Naive Bayes also demonstrated strong performance, while DecisionTree and K-Nearest Neighbors benefited from applying SMOTE and additional modifications. However, other performance indicators such as accuracy, recall, and F1 score must be considered to gain a more comprehensive understanding of the performance of each model.

#### **4.2.2 Precision**

Precision is a performance metric used in classification problems that measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It is calculated using the Confusion Matrix. The formula for precision is:

$$\textit{Precision} = TP / (TP + FP)$$

Precision focuses on the accuracy of positive predictions and provides insights into the model's ability to avoid false positive errors. A higher precision value indicates fewer false positive errors, meaning the model is more reliable in identifying positive instances (James, G. et al., 2021).



**Figure 4.5 Precision For Different Model's Performance**

According to Figure 4.5, analyzing the model's performance for Precision can lead to the following.

The DecisionTree algorithm achieved a precision of 74.56%, indicating a moderate performance level in correctly predicting positive instances. When applying the SMOTE technique to handle class imbalance (DecisionTree+Smote) and further modifications (DecisionTree+Smote+new), the precision remains consistent at 76.13%. These results suggest that the additional techniques do not significantly impact the DecisionTree model's precision.

The AdaBoost algorithm achieved a precision of 87.13%, indicating high accuracy in correctly predicting positive instances. When applying SMOTE (AdaBoost+Smote) and further modifications (AdaBoost+Smote+new), the precision increases to 88.96% and 88.81%, respectively. These improvements indicate that AdaBoost benefits from applying SMOTE and additional modifications, resulting in a more precise model.

For Random Forest, the precision without additional techniques (65.44%) is relatively lower than other models. When SMOTE is applied (Random Forest+Smote) and combined with additional modifications (Random Forest+Smote+new), the precision improves to

75.26% and 75.32%, respectively. These results indicate that SMOTE and additional modifications positively impact improving the precision of the Random Forest model.

Regarding Naïve Bayes, the initial precision is relatively high at 87.65%. When SMOTE is applied (Naïve Bayes+Smote) and combined with further modifications (Naïve Bayes+Smote+new), the precision remains consistently high at 87.44% and 87.73%, respectively. These results suggest that Naïve Bayes performs well in correctly predicting positive instances, and the additional techniques do not significantly impact its precision.

Lastly, K-Nearest Neighbors achieved a precision of 86.37%. Applying SMOTE (K-Nearest Neighbors+Smote) and additional modifications (K-Nearest Neighbors+Smote+new) results in slightly lower precision values of 84.22% and 84.32%, respectively. These findings indicate that SMOTE and additional modifications might slightly negatively impact the precision of the K-Nearest Neighbors model.

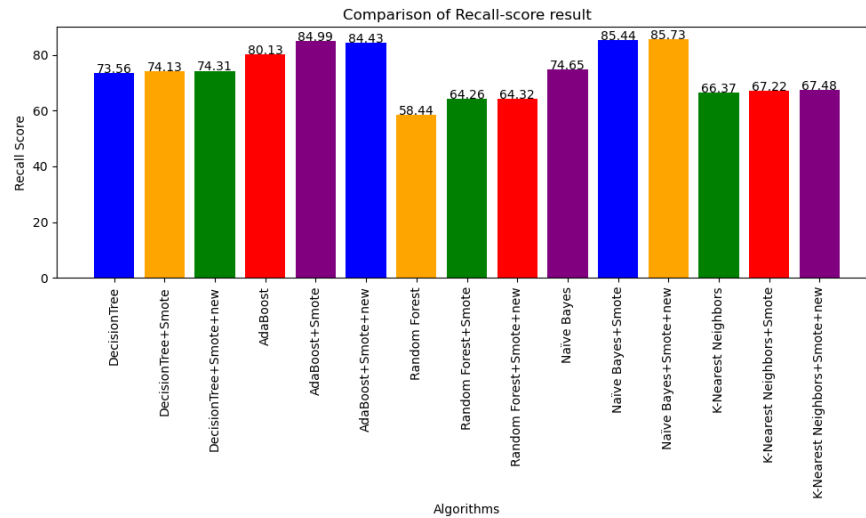
AdaBoost demonstrates a consistently high precision, while Random Forest, Naïve Bayes, and DecisionTree show moderate precision levels. K-Nearest Neighbors exhibits a slight decrease in precision by applying additional techniques. It is important to consider other performance metrics and the problem's specific context to understand each model's performance comprehensively.

#### 4.2.3 Recall

In classification tasks, recall, alternatively termed sensitivity or the true positive rate, serves as an evaluative measure. It gauges the ratio of accurately predicted positive cases to the overall true positive instances, within the Confusion Matrix framework. The calculation of recall is determined by the formula:

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

Retrieval centers on the model's capacity to accurately recognize affirmative cases, proving especially valuable in situations where the consequences of missing true positives (erroneously categorizing affirmative cases as negative) carry substantial repercussions. A higher recall value indicates a higher ability to capture positive instances and avoid false negative errors (James, G. et al., 2021).



**Figure 4. 6 Recall-Score For Different Model's Performance**

Based on Figure 4.6, analyzing the Recall results of different models, evaluate the model's performance based on its ability to identify positive instances correctly. The recall rate of the decision tree algorithm is 73.56%, indicating a moderate level of performance in correctly identifying positive instances. The application of SMOTE technology and further modifications only resulted in a slight increase in the recall rate of DecisionTree. On the other hand, the AdaBoost algorithm exhibits a relatively high recall rate of 80.13%, indicating its strong ability to identify positive instances correctly. The application of SMOTE and additional modifications have further improved the recall rate of AdaBoost, thus obtaining a higher value.

In the case of Random Forest, the recall rate without any additional technology is relatively low, at 58.44%. However, the application of SMOTE and additional modifications has shown positive effects in improving the model's ability to capture positive instances, as evidenced by the increase in recall values.

Similarly, Naive Bayes exhibits moderate ability in correctly identifying positive instances, with an initial recall rate of 74.65%. The application of SMOTE and additional modifications have further improved the recall rate of naive Bayes, highlighting the advantages of these technologies in improving model performance. Finally, the K-nearest neighbor showed a moderate recall rate of 66.37%, with a slight increase in recall rate with the application of SMOTE and additional modifications. The Result indicates that these technologies have little impact on improving the model's ability to capture positive instances.

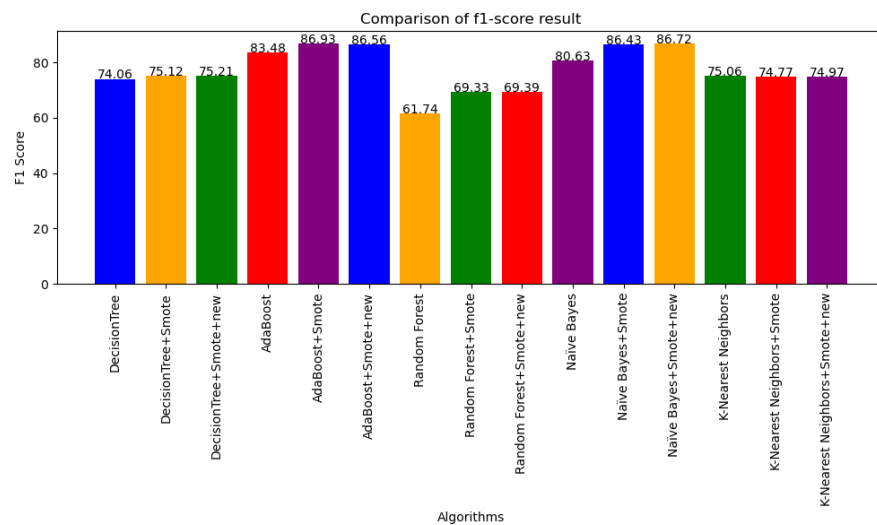
To sum up, AdaBoost and naive Bayes have relatively high Recall Score, while Random forest and decision tree have medium Recall Score, and KNN has the worst performance.

#### 4.2.4 F1-Score

The F1 score, a frequently employed performance measurement, merges precision and recall into a unified metric, offering an equilibrium between a model's precision in correct identifications and its ability to accurately predict positives. The F1 score calculation utilizes the Confusion Matrix as follows:

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

The precision metric gauges the correctness of positive forecasts, while recall assesses the model's aptitude in correctly recognizing positive occurrences. By amalgamating these two measures through their harmonic mean, the F1 score emerges. This score spans from 0 to 1, with 1 denoting the optimal F1 score (exemplifying impeccable precision and recall), and 0 indicating the least favorable outcome. Particularly valuable in scenarios involving imbalanced datasets—where distinct classes hold notably differing instance counts—the F1 score furnishes an equitable evaluation of the model's efficacy, encompassing both erroneous positive classifications and erroneous negative classifications.



**Figure 4. 7 F1-Score For Different Model's Performance**

Based on the F1 score results, we analyzed the models' performance in terms of overall accuracy and the balance between precision and recall.

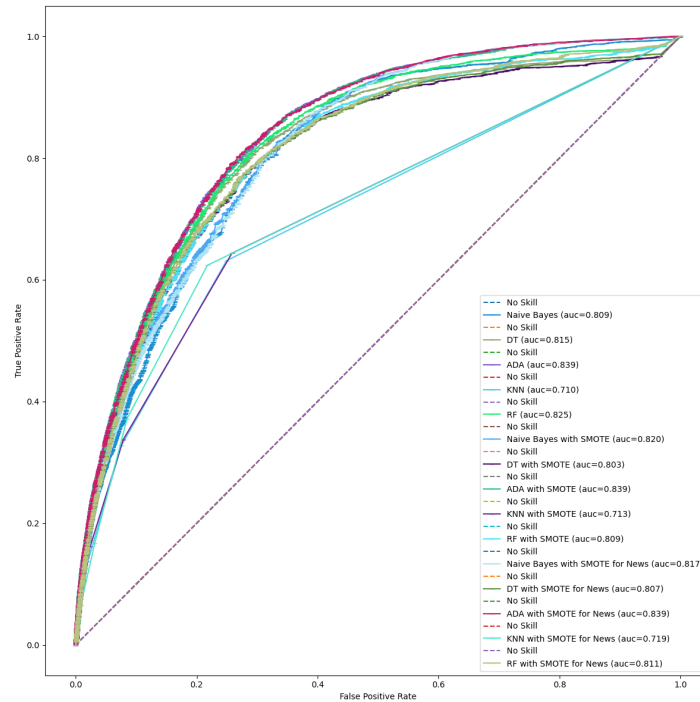
The DecisionTree algorithm had a moderate F1 score of 74.06%. Applying SMOTE and further modifications slightly improved the score to 75.12% and 75.21%, respectively, indicating a small positive impact. AdaBoost achieved a high F1 score of 83.48%,

indicating accurate prediction of positive instances and a good balance between precision and recall. With SMOTE and additional modifications, the score improved to 86.93% and 86.56%. For Random Forest, the initial F1 score without modifications was relatively lower at 61.74%. However, applying SMOTE and additional modifications increased the score to 69.33% and 69.39%, respectively, improving the model's overall performance. Naïve Bayes showed a high initial F1 score of 80.63%, demonstrating good accuracy and balance. Applying SMOTE and additional modifications maintained consistently high scores of 86.43% and 86.72%. K-Nearest Neighbors achieved an F1 score of 75.06%. With SMOTE and additional modifications, the scores decreased slightly to 74.77% and 74.97%, indicating a minor negative impact on performance.

In summary, AdaBoost(+Smote) consistently had the highest F1 score, showing strong overall performance. Naïve Bayes also demonstrated a good balance between precision and recall. DecisionTree and K-Nearest Neighbors benefited from modifications, while Random Forest had a lower overall performance. Considering other metrics and the specific problem context is essential for a comprehensive understanding of model performance.

#### **4.2.5 ROC Curve**

The Receiver Operating Characteristic (ROC) curve is a graphical representation illustrating the efficacy of a binary classification model. It contrasts the true positive rate (TPR) with the false positive rate (FPR) across varying thresholds. This curve serves to evaluate the model's proficiency in distinguishing between positive and negative instances at different thresholds. The construction of the ROC curve involves ranking the model's predictions based on probability scores or decision thresholds. Modifying the threshold leads to the classification of instances as positive or negative, enabling computation of TPR and FPR by comparing accurately classified positive cases and inaccurately classified negative cases. The ROC curve materializes by plotting TPR against FPR at distinct thresholds, offering insights into the model's performance. Positioned closer to the upper-left corner signifies superior performance, while a diagonal alignment represents a random classifier. The ROC curve facilitates performance assessment and threshold selection. An often-used metric, the Area Under the Curve (AUC), quantifies discriminative ability, with higher values indicating superior performance (Fawcett, T., 2006).



**Figure 4. 8 The Receiver Operating Characteristic (ROC) Curve For Different Model's Performance**

Based on Figure 4.8, AdaBoost with SMOTE and additional modifications achieved the highest AUC of 0.839. The result indicates that the AdaBoost model, when combined with SMOTE and the additional modifications, can differentiate between positive and negative instances.

On the other hand, Naive Bayes achieved an AUC of 0.820, while Naive Bayes with SMOTE achieved an AUC of 0.817. These values indicate that Naive Bayes models possess a reasonable discriminative ability, although slightly lower than AdaBoost. The remaining models have AUC values below the AUC of Naive Bayes and Naive Bayes with SMOTE. These results suggest that those models have comparatively lower discriminative abilities in distinguishing between positive and negative instances.

#### 4.2.6 Summary

Comprehensive analysis can be seen based on evaluation indicators, including accuracy, F1 score, recall rate, precision, and AUC in ROC.

AdaBoost with SMOTE and other modifications is the best-performing algorithm among the evaluated algorithms. It reached the highest AUC of 0.839, indicating excellent discrimination ability in distinguishing positive and negative instances. AdaBoost also



showed high accuracy, F1 score, recall rate, and accuracy values of 91.17%, 83.48%, 80.13%, and 87.13%, respectively. In contrast, Naïve Bayes with and without SMOTE also performed relatively well. Their AUC values are 0.817 and 0.820, respectively. Although their accuracy, F1 score, recall, and accuracy scores are slightly lower than AdaBoost, they still demonstrate competitiveness. Other models, such as Decision Tree, Random Forest, and K-Nearest Neighbors, have shown relatively low performance in accuracy, F1 score, recall, accuracy, and AUC values.

Therefore, based on the evaluation of multiple indicators, AdaBoost, SMOTE, and other modifications have become the best algorithms for predicting the risk of coronary heart disease in this situation.

### 4.3 Evaluating Web Application

#### 4.3.1 Unit Testing

**Table 4. 2 Unit Testing**

i d	use r	Hei ght (m)	Wei ght (Kg )	Ag e	Gend er	Stro ke	Smoki ng	Hig h Ch oles tero l (m g/d l)	Hy pert ension ( sys toli c pre ss ure )	Diff icult Wal king	He alt h De gre e	Ment al Healt h	Inc ome	Dia bete s	Pred ic Butt on	Result
1	use r1	0	0	23	Men	No	No	200	130	No	No	30	50 00 0	No	Clic ked	Please fill height or weight.
2	Us er2	1.5 6	60	0	Wom en	No	No	178	120	No	No	30	20 00 0	No	Clic ked	Please fill Age.
3	Us er3	1.8 2	80	22	Men	No	No	184	117	No	No	0	0	No	Clic ked	Please change Mental Health or Income if these two factors are not 0.
4	Us er4	1.8 4	72	24	Men	No	No	176	120	No	No	30	10 00 0	No	Clic ked	No Risk
5	Us er5	1.6 5	65	25	Wom en	No	No	186	125	No	No	30	70 00 0	No	Clic ked	No Risk

Based on the Table 4.2 unit testing table for the CHDs (Coronary Heart Disease) Risk Prediction System, we can draw the following conclusions:

User 1: The user did not provide values for height and weight, resulting in a validation error. The system prompted the user to fill in either height or weight. No other risk factors were identified.

User 2: The user-provided values for height, weight, and gender but did not fill in the age field. As a result, the system prompted the user to fill in the age. No other risk factors were identified.

User 3: The user-provided values for height, weight, age, and gender, but the mental health and income values were set to zero. The system prompted the user to change the mental health or income values if they were not zero. No other risk factors were identified.

User4: The user-provided values for all the required fields, and no risk factors were identified. The system indicated that there was no risk.

User 5: The user-provided values for all the required fields, and no risk factors were identified. The system indicated that there was no risk.

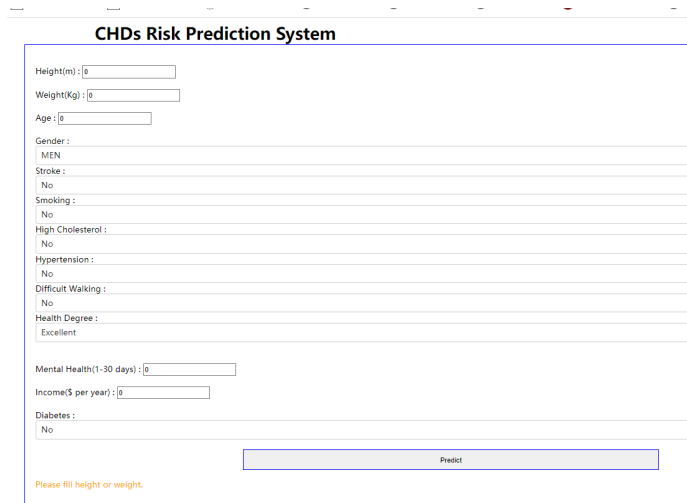
Based on the unit testing results, the CHDs Risk Prediction System can handle validation errors and prompt users to provide missing information when necessary. The system correctly identifies cases where there are no risk factors present and indicates that there is no risk.

#### **4.3.2 Data Validation**

Data validation is the process of ensuring that data is accurate, consistent, and meets certain predefined criteria or rules. It is an important step in data quality management to maintain the integrity and reliability of data. In the context of the CHDs Risk Prediction System, data validation is used to check the input provided by users and ensure that it is valid and complete.

In Table 4.2 unit testing table, these examples of data validation checks performed by the system:

Height and Weight Validation: User1 did not provide values for height and weight, which violated the validation rule. The system prompted the user to fill in either height or weight before proceeding.



**CHDs Risk Prediction System**

Height(m) :

Weight(Kg) :

Age :

Gender :

MEN :

Stroke :

No :

Smoking :

No :

High Cholesterol :

No :

Hypertension :

No :

Difficult Walking :

No :

Health Degree :

Excellent :

Mental Health(1-30 days) :

Income(\$ per year) :

Diabetes :

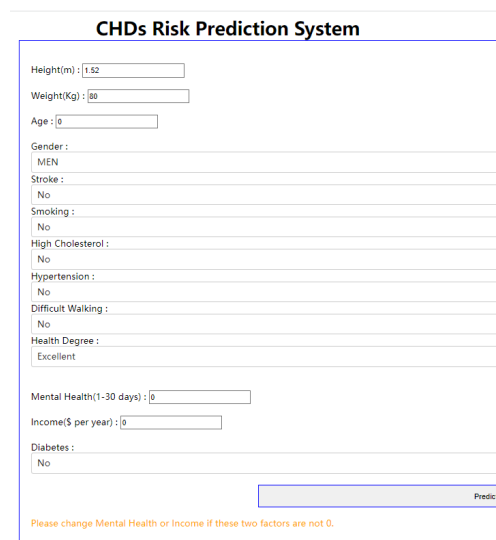
No :

Please fill height or weight.

**Figure 4. 9 Height and Weight Validation**

Age Validation: User2 did not provide an age value, resulting in a validation error. The system prompted the user to fill in the age field before proceeding.

Mental Health and Income Validation: User3 provided values of zero for both mental health and income. The system detected this and prompted the user to change either the mental health or income values if they were not both zero.



**CHDs Risk Prediction System**

Height(m) :

Weight(Kg) :

Age :

Gender :

MEN :

Stroke :

No :

Smoking :

No :

High Cholesterol :

No :

Hypertension :

No :

Difficult Walking :

No :

Health Degree :

Excellent :

Mental Health(1-30 days) :

Income(\$ per year) :

Diabetes :

No :

Please change Mental Health or Income if these two factors are not 0.

**Figure 4. 10 Mental Health and Income Validation**

Other Factors Validation: In order to ensure the data validity of other feature values, specific default values are selected as No values, while Health Degree defaults to selecting Excellent values.

These examples demonstrate how the CHDs Risk Prediction System performs data validation by checking for missing or invalid values and providing appropriate feedback or

prompts to the users. Data validation helps ensure that the system receives accurate and complete input, which is crucial for generating reliable risk predictions.

**4.3.3 Cross-browser And Cross-device Testing**

Cross-browser and cross-device testing encompasses the examination of a website or application across diverse web browsers and devices to ensure consistent user experience and functionality. This testing is crucial due to the potential disparities in how various browsers and devices interpret and showcase web content, leading to possible compatibility complications. Throughout cross-browser testing, the site or app is assessed on a range of web browsers like Google Chrome, Mozilla Firefox, and Microsoft Edge, spotlighting browser-specific issues encompassing rendering discrepancies, CSS harmony, JavaScript responses, and HTML5 compatibility. Concurrently, cross-device testing involves evaluating the site or application on a spectrum of devices, including desktop PCs, laptops, tablets, and smartphones. These devices exhibit distinct screen dimensions, resolutions, operating systems, and capabilities, all of which can influence the layout, responsiveness, and operation of the site or application. Further information is available in Table 4.3.

**Table 4. 3 Cross-browser And Cross-device Testing Table**

Device	Name	Layout	Responsiveness	Functionality
Computer Web Browsers	Google Chrome	Normal Operation	Normal Operation	Normal Operation
	Mozilla Firefox	Normal Operation	Normal Operation	Normal Operation
	Microsoft Edge	Normal Operation	Normal Operation	Normal Operation
Mobile Phone Browsers	Huawei Browser	Normal Operation	Normal Operation	Normal Operation
	Iphone Broswer	Normal Operation	Normal Operation	Normal Operation

By conducting cross-browser and cross-device testing, ensuring that this web projects work seamlessly across different platforms, browsers, and devices, providing a consistent user experience for all users.

#### 4.3.4 Error Handling And Logging

Error handling and logging are key aspects of software development used to identify, track, and handle errors or exceptions during program execution. Error handling involves implementing mechanisms for elegantly capturing and managing errors, while logging refers to recording information about errors and events for debugging and analysis purposes.

Log files in Python Anywhere handle errors and log information in web applications such as CHDs risk prediction systems.

**Access log:** The access log file named "xianlongdi. python anywhere. com. Access. Log" records information about each HTTP request made to the web application. It includes detailed information such as the requester's IP address, request URL, HTTP response status code, and other related information. This log can be used to monitor and analyze application traffic.

**Error Log:** An error log file named 'xianlongdi. pythonanywhere. com. Error. Log' captures information about errors that occurred during the execution of a web application. It includes backtracking information, error messages, and relevant detailed information that helps diagnose and fix problems. Monitoring this log can provide insight into runtime errors and exceptions encountered by the application.

**Server log:** The Server log file named "xianlongdi. python anywhere. com. Server. Log" contains general server-level information and messages related to the hosting environment. It may include startup and shutdown messages, configuration changes, and other server-related events.

In addition, Python Anywhere periodically rotates log files to prevent them from becoming too large. Older log files can be found in the directory '/var/log.' This directory stores archived log files that are no longer actively written but may contain valuable historical data for troubleshooting or analysis.

## Log files:

The first place to look if something goes wrong.

Access log: [xianlongdi.pythonanywhere.com.access.log](https://xianlongdi.pythonanywhere.com/access.log)

Error log: [xianlongdi.pythonanywhere.com.error.log](https://xianlongdi.pythonanywhere.com/error.log)

Server log: [xianlongdi.pythonanywhere.com.server.log](https://xianlongdi.pythonanywhere.com/server.log)

Log files are periodically rotated. You can find old logs here: [/var/log](#)

**Figure 4. 11 Log files**

```
2023-07-11 10:03:40,787: File /usr/local/lib/python3.10/site-packages/joblib/numpy_pickle.py, line 579, in load
2023-07-11 10:03:40,787: with open(filename, 'rb') as f:
2023-07-11 10:03:40,787: *****
2023-07-11 10:03:40,787: If you're seeing an import error and don't know why,
2023-07-11 10:03:40,787: we have a dedicated help page to help you debug:
2023-07-11 10:03:40,787: https://help.pythonanywhere.com/pages/DebuggingImportError/
2023-07-11 10:03:40,787: *****
2023-07-11 10:03:44,694: Error running WSGI application
2023-07-11 10:03:44,695: FileNotFoundError: [Errno 2] No such file or directory: '/home/xianlongdi/chd/ADA.pkl'
2023-07-11 10:03:44,695: File "/var/www/xianlongdi.pythonanywhere.com/wsgi.py", line 111, in <module>
2023-07-11 10:03:44,695: from chd import app
2023-07-11 10:03:44,695:
2023-07-11 10:03:44,695: File "/home/xianlongdi/chd/chd.py", line 14, in <module>
2023-07-11 10:03:44,695: ada_model = joblib.load("/home/xianlongdi/chd/ADA.pkl")
2023-07-11 10:03:44,696:
2023-07-11 10:03:44,696: File /usr/local/lib/python3.10/site-packages/joblib/numpy_pickle.py, line 579, in load
2023-07-11 10:03:44,696: with open(filename, 'rb') as f:
2023-07-11 10:03:44,696: *****
2023-07-11 10:03:44,696: If you're seeing an import error and don't know why,
2023-07-11 10:03:44,696: we have a dedicated help page to help you debug:
2023-07-11 10:03:44,696: https://help.pythonanywhere.com/pages/DebuggingImportError/
2023-07-11 10:03:44,697: *****
2023-07-12 04:32:29,996: OSError: write error
```

**Figure 4. 12 Error log**

In order to effectively handle errors and log information in web applications on Python Anywhere, using these log files can effectively track and debug issues, gain a deeper understanding of application usage, and monitor the overall health and performance of the CHDs risk prediction system.

## **CHAPTER 5: Conclusion**

The study aimed to identify lifestyle and health factors associated with the development of coronary heart disease (CHD), construct a predictive model for CHD risk, and develop a user-friendly web application. This study utilized publicly available datasets on lifestyle and health factors, such as smoking, alcohol consumption, physical activity, high blood pressure, and cholesterol levels.

Using the XGBoost algorithm, the study identified hypertension, BMI, Difficult Walking, High Cholesterol, Income, Age, Stroke, MentHlth, Sex, GenHlth, Smoker, and Diabetes as the most important lifestyle and health factors for predicting CHD risk. Five machine learning algorithms (AdaBoost, Random Forest, Decision Tree, KNN, and Naive Bayes) were employed, and a Smote balanced dataset was used to construct the models. The accuracy of the machine learning algorithms was evaluated using various metrics such as Recall, F1 score, Precision, Accuracy, and ROC curve values.

The best prediction model, AdaBoost+Smote, achieved an accuracy of over 90% in predicting CHD risk. This model was integrated with Python and Dash to develop a convenient web application. The web application underwent testing and evaluation, including Unit Testing and Data Validation, to ensure its functionality and accuracy. The final test and evaluation results indicated that the web application met the expected operational goals.

In conclusion, this study successfully identified important lifestyle and health factors associated with CHD risk, developed an accurate predictive model using machine learning techniques, and created a user-friendly web application for predicting the risk of developing coronary heart disease. The findings provide valuable insights into understanding and assessing CHD risk based on lifestyle and health factors. The developed web application can serve as a useful tool for both users and healthcare practitioners in managing and preventing CHD.



## References

- Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R. I., Tamam, N., Sulieman, A., & Pathan, R. K. (2022). The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. *Healthcare*, 10(6), 1137. <https://doi.org/10.3390/healthcare10061137>
- Fadnavis, R., Dhore, K., Gupta, D., Waghmare, J., & Kosankar, D. (2021). Heart disease prediction using data mining. *Journal of Physics: Conference Series*, 1913(1), 012099. <https://doi.org/10.1088/1742-6596/1913/1/012099>
- Hassan, Ch. A. ul, Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., Alturki, N., & Ullah, S. S. (2022). Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. *Sensors*, 22(19), 7227. <https://doi.org/10.3390/s22197227>
- K. AL-Taie, R. R., Saleh, B. J., Falih Saedi, A. Y., & Salman, L. A. (2021). Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(6), 5229. <https://doi.org/10.11591/ijece.v11i6.pp5229-5239>
- Luo, M., Hu, Y., Bai, R., & Xu, Z. (2022). Data Mining-Based Analysis of Modern Chinese Medicine for the Treatment of Stable Angina Pectoris in Coronary Heart Disease. *Journal of Healthcare Engineering*, 2022, 1–6. <https://doi.org/10.1155/2022/3511974>
- Meda, S., & Bhogapathi, R. (2018). Identification of heart disease using fuzzy neural genetic algorithm with data mining techniques. *Advances in Modelling and Analysis B*, 61(2), 99–105. [https://doi.org/10.18280/ama\\_b.610208](https://doi.org/10.18280/ama_b.610208)
- Shen, T., Liu, D., Lin, Z., Ren, C., Zhao, W., & Gao, W. (2022). A Machine Learning Model to Predict Cardiovascular Events during Exercise Evaluation in Patients with Coronary Heart Disease. *Journal of Clinical Medicine*, 11(20), 6061. <https://doi.org/10.3390/jcm11206061>
- Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*, Volume 13, 121–124. <https://doi.org/10.2147/IJN.S124998>
- Wan Zunaidi, W. H. A., Saedudin, R. R., Ali Shah, Z., Kasim, S., Sen Seah, C., & Abdurrohman, M. (2018). Performances Analysis of Heart Disease Dataset

using Different Data Mining Classifications. *International Journal on Advanced Science, Engineering and Information Technology*, 8(6), 2677. <https://doi.org/10.18517/ijaseit.8.6.5042>

Wang, G., Luo, L., & Zhao, X. (2021). Diagnosis of Arrhythmia for Patients with Occult Coronary Heart Disease Guided by Intracavitary Electrocardiogram under Data Mining Algorithm. *Journal of Healthcare Engineering*, 2021, 1–8. <https://doi.org/10.1155/2021/1640870>

Methaila, A. , et al. "Early Heart Disease Prediction Using Data Mining Techniques." *Fourth International Conference on Computational Science, Engineering and Information Technology* 2014.

Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. 2012 Japan-Egypt conference on electronics, communications and computers (pp. 173-177). IEEE. DOI: 10.1109/JEC-ECC.2012.6186978

Azzam, F. (2022). Data Mining Techniques for Heart Disease Prediction—A Review. <https://doi.org/10.13140/RG.2.2.31949.31204>

Behera, M. P. (2019). Analysis on Data Mining Application. 09(03), 3.

Fadnavis, R., Dhore, K., Gupta, D., Waghmare, J., & Kosankar, D. (2021). Heart disease prediction using data mining. *Journal of Physics: Conference Series*, 1913(1), 012099. <https://doi.org/10.1088/1742-6596/1913/1/012099>

Mostofi, S., Kordrostami, S., Refahi, A. H., Faridi Masooleh, M., & Shokri, S. (2022). Data mining and diagnosis of heart diseases: A hybrid approach to the b-mine algorithm and association rules. *International Journal of Research in Industrial Engineering*, 11(1). <https://doi.org/10.22105/riej.2022.302672.1243>

Salve, I., & Borikar, D. A. (2022). Heart Disease Prediction using Data Mining Techniques. *JOURNAL OF ALGEBRAIC STATISTICS*, 13(3), 10.

Alaa, A. M. , et al. "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants." *PLoS ONE* 14.5(2019):e0213653-.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). DOI: 10.1145/2939672.2939785

XGBoost Documentation. (n.d.). XGBoost Python Package. Retrieved from <https://xgboost.readthedocs.io/>