

HEART DISEASE PREDICTION
USING MACHINE
LEARNING

ZHENG ZEYU

UNIVERSITI KEBANGSAAN MALAYSIA

HEART DISEASE PREDICTION USING MACHINE
LEARNING

ZHENG ZEYU

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

MERAMAL PENYAKIT JANTUNG MENGGUNAKAN
PEMBELAJARAN MESIN

ZHENG ZEYU

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT UNTUK MEMPEROLEH IJAZAH SARJANA
SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2025

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

06 January 2025

ZHENG ZEYU

P131656

LIBRARY FETSM

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to the Lord of all things, the Creator of heaven and earth, for all that has been bestowed upon us through learning His language. I am especially grateful for His gifts. For this research work, I would like to express my gratitude to Professor Suhaila Zainudin for her guidance and valuable technical support on my research project, which has enabled the results of my experiment to be more comprehensive, objective, and complete, and better realize the research value of the final academic achievements. In addition, I would like to thank the World Health Organization (WHO) for providing relevant explanations and concept introductions on heart disease, clearly demonstrating the importance of preventing heart disease for human health, thus supporting the necessity of conducting this research project.

Secondly, I would also like to express my gratitude to the Kaggle platform for providing relevant heart disease experimental dataset samples, as well as all the staff and reference materials provided by the National University of Malaysia Library. This has provided a certain evidence basis for the experiment and the reliability of the experimental conclusions. At the same time, I would like to thank the Malaysian government and people for their warm hospitality and all those who have contributed to the progress and prosperity of these facilities, which have enabled the experiment to have more extensive technical support and in-depth research conditions.

Finally, I would like to express my sincere gratitude to all the teachers and classmates from the Faculty of Information Science and Technology for their help and useful suggestions in further improving and simplifying the experimental process. I am also very grateful to my family who have silently supported me throughout my efforts, providing me with the confidence and strength to complete the experimental research.

ABSTRAK

Penyakit jantung merupakan isu kesihatan awam yang besar di dunia, yang menyebabkan banyak kematian setiap tahun. Kajian ini bertujuan menggunakan teknik pembelajaran mesin untuk meramalkan risiko penyakit jantung. Dengan menganalisis faktor berkaitan, model ramalan yang efektif dibangun untuk menyediakan dasar untuk pencegahan awal dan intervensi. Kajian itu menggunakan dataset penyakit jantung Framingham Kaggle, yang mengandungi 4,240 rekod dan 16 atribut. Kajian ini menggunakan pelbagai kaedah termasuk prapemprosesan data, pemilihan ciri, dan pembinaan model untuk menilai model. Mengguna bahasa pengaturcaraan Python dan pustaka berkaitan untuk pemprosesan dan analisis data, membina model ramalan menggunakan algoritma pembelajaran mesin seperti regresi logistik, mesin vektor sokongan, pokok keputusan, hutan rawak, Bayes naif, dan XGBoost, dan mengevaluasi prestasi model menggunakan pengesahihan salib 10 kali dan nisbah sekatan set data berbeza. Hasil menunjukkan bahawa model XGBoost berfungsi dengan baik dalam kebanyakan konfigurasi, dengan ketepatan yang tinggi, terutamanya apabila set latihan menyumbang sebahagian besar, mencapai maksimum 91.94%. Ia boleh lebih baik menangkap corak kompleks dalam data. Model hutan rawak juga mempunyai keupayaan generalisasi yang kuat, tetapi kerana pengaruh pembahagian set data, ia hanya melakukan yang terbaik dalam pengesahan silang sepuluh kali lipat, dengan ketepatan 78.81%. Model SVM berfungsi dengan baik dalam memproses data yang agak mudah, tetapi mempunyai kerumitan pengiraan yang tinggi, dengan ketepatan tertinggi 68.89% apabila set latihan menyumbang 90%. Prestasi model DT dan LR sederhana, sekitar 65%, manakala model NB tidak sangat sesuai kerana kadar panggilan yang rendah hanya sekitar 50%. Ujian ini menyediakan rujukan berharga untuk pembelajaran mesin dalam bidang ramalan penyakit jantung, yang boleh membantu mengesan risiko penyakit jantung awal dan meningkatkan prognosis pesakit. Penyelidikan masa depan boleh memperbaiki model, mengeksplorasi sumber data dan teknologi baru untuk meningkatkan kemampuan prediksi dan praktikal klinik.

ABSTRACT

Heart disease is a major global public health issue, leading to a large number of deaths every year. The aim of this study is to use machine learning techniques to predict the risk of heart disease. By analyzing relevant factors, an effective prediction model is constructed to provide a basis for early prevention and intervention. The study used the Framingham heart disease dataset from Kaggle, which contains 4,240 records and 16 attributes. The study used various methods, including data preprocessing, feature selection, and model building to evaluate the model. Using Python programming language and related libraries for data processing and analysis, constructing prediction models using machine learning algorithms such as LR, SVM, DT, RF, NB and XGBoost, and evaluating model performance using 10-fold cross validation and different dataset partitioning ratios. The results show that the XGBoost model performs well in most configurations, with high accuracy, especially when the training set accounts for a large proportion, reaching a maximum of 91.94%. It can better capture complex patterns in the data. The random forest model also has strong generalization ability, but due to the influence of dataset partitioning, it only performs the best in ten-fold cross validation, with an accuracy of 78.81%. The SVM model performs well in processing relatively simple data, but has a high computational complexity, with the highest accuracy of 68.89% when the training set accounts for 90%. The performance of DT and LR models is moderate, around 65%, while the NB model is not very suitable due to its low recall rate of only around 50%. This study provides valuable reference for machine learning in the field of heart disease prediction, which can help detect heart disease risk early and improve patient prognosis. Future research can further optimize the model, explore new data sources and technologies to improve predictive ability and clinical practicality.

TABLE OF CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENTS		iv
ABSTRAK		v
ABSTRACT		vi
TABLE OF CONTENTS		vii
LIST OF TABLES		ix
LIST OF ILLUSTRATIONS		x
LIST OF ABBREVIATIONS		xii
CHAPTER I	INTRODUCTION	
1.1	Research Background	1
1.2	Problem Statement	3
1.3	Research Objective	4
1.4	Research Significance	4
1.5	Thesis Structure	5
CHAPTER II	LITERATURE REVIEW	
2.1	Introduction	7
2.2	Machine Learning for Heart Disease Prediction	7
2.3	Introduction of Chi Square Test	21
2.4	Data Prediction	22
2.5	10-Fold-Cross Validation	22
2.6	Machine Learning Model	22
2.7	Performance Evaluation	26
2.8	Summary	29
CHAPTER III	RESEARCH METHODOLOGY	
3.1	Introduction	30
3.2	Experiment Environment	31
3.3	Research Dataset	32
	3.3.1 Basic Information of Dataset Feature Attributes	32
	3.3.2 Statistical Information about the Dataset	35

	3.3.3	Visualization of Characteristic Numerical Distribution	36
	3.3.4	Missing Value Information in the Dataset	38
3.4		Data Processing	40
	3.4.1	Missing Value	40
	3.4.2	Randomly Resampled Dataset	42
	3.4.3	Chi Square Test	43
	3.4.4	Standardization of Datasets	44
	3.4.5	Proportional Partitioning of Training and Testing Sets	44
	3.4.6	Implementation of 10-Fold Cross Validation	45
3.5		Implementation of Machine Learning	45
3.6		Summary	48
CHAPTER IV		RESULTS AND DISCUSSION	
4.1		Introduction	50
4.2		Research Results	50
	4.2.1	The Most Influential Features on the Final Result	50
	4.2.2	Parameter Selection for Machine Learning Models	55
	4.2.3	Model Capability Demonstration and Evaluation	57
4.3		Visualization and Analysis of Experimental Results	74
4.4		Overall Evaluation of Various Methods	78
4.5		Summary	79
CHAPTER V		CONCLUSION	
5.1		Introduction	80
5.2		Influence Factor	80
5.3		Model Parameters	81
5.4		Model Evaluation	82
5.5		Limitations of the Experiment	82
5.6		Future Work	83
REFERENCES			84

LIST OF TABLES

Table No.		Page
Table 2.1	Summary of Reference Views	15
Table 2.2	Confusion Matrix	27
Table 3.1	Introduction to Dataset Attributes	33
Table 3.2	Dataset attribute types	35
Table 3.3	The number of missing values for each attribute	38
Table 4.1	Evaluation metrics for LR	58
Table 4.2	Evaluation metrics for SVM	60
Table 4.3	Evaluation metrics for DT	62
Table 4.4	Evaluation metrics for RF	68
Table 4.5	Evaluation metrics for NB	70
Table 4.6	Evaluation metrics for XGBoost	72
Table 4.7	Comparison of the best performance of each model	78

LIST OF ILLUSTRATIONS

Figure No.		Page
Figure 3.1	Experimental procedure	31
Figure 3.2	Dataset Example	32
Figure 3.3	Dataset statistics	36
Figure 3.4	Dataset feature data distribution histogram	37
Figure 3.5	The proportion of missing features in each dataset	40
Figure 3.6	Changes in missing values in the dataset	41
Figure 3.7	Changes in sample size during resampling	43
Figure 3.8	Example after data standardization	44
Figure 4.1	The chi square values of different features in the dataset	50
Figure 4.2	SysBP distribution based on Heart Disease Risk	51
Figure 4.3	Glucose distribution based on Heart Disease Risk	52
Figure 4.4	TotChol distribution based on Heart Disease Risk	52
Figure 4.5	Age distribution based on Heart Disease Risk	53
Figure 4.6	CigsPerDay distribution based on Heart Disease Risk	54
Figure 4.7	DiaBP distribution based on Heart Disease Risk	54
Figure 4.8	PrevalentHyp distribution based on Heart Disease Risk	55
Figure 4.9	ROC curve of LR	59
Figure 4.10	ROC curve plot of LR (10-fold cross validation)	59
Figure 4.11	ROC curve of SVM	61
Figure 4.12	ROC curve of SVM (10-fold cross validation)	61
Figure 4.13	ROC curve of DT	63
Figure 4.14	ROC curve of DT (10-fold cross validation)	63
Figure 4.15	DT Model Diagram (90% Train 10% Test)	64
Figure 4.16	DT Model Diagram (80% Train 20% Test)	64

Figure 4.17	DT Model Diagram (70% Train 30% Test)	64
Figure 4.18	DT Model Diagram (60% Train 40% Test)	65
Figure 4.19	DT Model Diagram (50% Train 50% Test)	65
Figure 4.20	DT Model Diagram (40% Train 60% Test)	65
Figure 4.21	DT Model Diagram (30% Train 70% Test)	66
Figure 4.22	DT Model Diagram (20% Train 80% Test)	66
Figure 4.23	DT Model Diagram (10% Train 90% Test)	66
Figure 4.24	DT Model Diagram (10-fold cross validation)	67
Figure 4.25	ROC Curve of RF	69
Figure 4.26	ROC curve of RF (10-fold cross validation)	69
Figure 4.27	ROC Curve of NB	71
Figure 4.28	ROC Curve of NB (10-fold cross validation)	71
Figure 4.29	ROC curve of XGBoost	73
Figure 4.30	ROC curve of XGBoost (10-fold cross validation)	73
Figure 4.31	The time spent on model training and prediction	74
Figure 4.32	The time spent on model training and prediction (10-fold cross validation)	75
Figure 4.33	Comparison of Precision between Models	75
Figure 4.34	Comparison of F1 scores between models	76
Figure 4.35	Comparison of accuracy between models	76
Figure 4.36	Comparison of AUC values between models	77
Figure 4.37	Comparison of recall rates between models	77

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
AUC	Area Under the ROC Curve
BMI	Body Mass Index
CHD	Coronary Heart Disease
DBP	Diastolic Blood Pressure
DDS	Dual Discriminant Scoring
DT	decision tree
FM	fat mass
FN	False Negative
FP	False Positive
FPR	false positive rate
FFM	weight without fat
KNN	K-nearest neighbor algorithm
LR	logistic regression
LV	left ventricle
MLP	Multi layer Perceptron
NB	naive Bayes
QD	Quadratic Discriminant Analysis
RBC	red blood cells
RBF	Radial Basis Function
RF	random forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic

SBP	systolic blood pressure
SVM	support vector machine
TP	True Positive
TPR	true positive rate
TN	True Negative
WHO	World Health Organization
XGBoost	extreme Gradient Boosting

LIBRARY ETSM

CHAPTER I

INTRODUCTION

1.1 RESEARCH BACKGROUND

The heart plays a crucial role in the human body by pumping blood to all parts of the body. If the heart cannot function properly, it may lead to patient death. "Heart disease" includes various diseases, including thinning of the myocardium, heart failure, myocardial blood insufficiency, and narrowing or blockage of the heart's blood vessels that affect the pumping and circulation of blood throughout the entire organ, all of which are referred to as heart disease (Soni J et al. 2011). When the coronary artery becomes narrowed or blocked due to plaque buildup from cholesterol and other substances on the artery walls, leading to myocardial ischemia, hypoxia, or necrosis, it is known as coronary atherosclerotic heart disease, commonly referred to as coronary heart disease, a type of ischemic heart disease. (Burnett et al. 2020). The onset of heart disease can cause a decrease in blood flow in the human blood vessels, leading to a decrease in red blood cells (RBC) and ultimately resulting in hypoxia and loss of consciousness.

Cardiovascular diseases cause a significant number of deaths each year, making it a major public health issue and the leading cause of death globally. (Kale A et al. 2024) According to data from the WHO, as of 2023, cardiovascular disease causes approximately 17.9 million deaths annually, with a relatively high mortality rate. (World health organization 2021) In daily life, there are several factors that can contribute to the development of heart disease. For example, McGill (1990) had shown that smoking is strongly related to coronary artery disease and other atherosclerosis sequelae, including the vascular endothelial damage caused by smoking, the increase of harmful lipids in the blood, and the initiation of inflammatory reactions. Smoking can also cause vasospasm, increase blood viscosity, promote thrombosis, and further exacerbate the risk of arterial obstruction and heart disease. Additionally, according to Gao Z et al.

(2019), in terms of the gender-related impact of heart disease, women may experience a significant increase in the risk of cardiovascular disease after menopause, as the protective effect of estrogen on the cardiovascular system diminishes following menopause. After menopause, women's blood pressure increases and the risk of atherosclerosis increases. This factor makes the incidence rate of cardiovascular disease in women in old age gradually approach or even exceed that of men. But similarly, (Spence & Pilote 2015) testosterone deficiency in men also increases the risk of atherosclerosis events.

A high BMI value may also be linked to an increased risk of cardiovascular disease (Ortega et al. 2016). For example, individuals with higher fat mass (FM) often have greater lean body mass, which refers to the weight without fat (FFM). Obese individuals need to adapt to additional weight burdens. This adaptation can result in enhanced blood circulation, which in turn increases the stroke volume and cardiac output of the left ventricle (LV). Over time, this may elevate the risk of ventricular hypertrophy and heart failure. The longer the duration of obesity, the greater the risk of developing cardiovascular disease.

Moreover, there may be a connection between heart disease and diabetes (Olimjonovna, 2024), primarily due to the impact of high blood sugar on the cardiovascular system. In individuals with diabetes, elevated blood sugar levels can damage blood vessels and nerves that regulate heart function, thereby increasing the risk of heart disease. In addition, high blood sugar can lead to arteriosclerosis, which is the accumulation of plaques in the arteries, hindering blood flow to the heart and potentially causing serious cardiovascular events such as angina and heart attacks. The SBP and DBP of the heart are also closely related to heart disease (Fuchs & Whelton 2024). During each contraction of the heart, if the pressure (systolic pressure) in the arteries is high, the heart must use greater force to overcome this resistance and pump blood into the arteries. Similarly, during periods of cardiac relaxation, if the pressure inside the arteries (diastolic pressure) is high, the heart requires greater force to fill the blood. This sustained high load can lead to myocardial hypertrophy and impaired cardiac function.

For a long time, preventing heart disease has been a theme in the field of public health, and methods based on reducing the level of risk factors by changing the lifestyle of the population are receiving increasing attention in the prevention, detection,

evaluation, and treatment of heart disease. Early detection and classification are crucial for controlling heart disease (Kaminsky L A 2022). Machine learning technology has proven to be highly effective in improving the precision of medical diagnoses. It is capable of processing large volumes of data and identifying patterns and relationships that might not be evident through conventional statistical techniques.

With the development of information technology, machine learning and data mining are evolving and encounter an important role in serving doctors to make precise disease and prevent clinical errors. Data mining refers to the process of finding unobserved trends and patterns across the entire dataset and using knowledge extracted from such to make predictions. Data mining is a technique for the study of distribution of a large amount of source information in order to explore the existence of certain invisible trends, information and relations from which the traditional statistical analysis cannot find it. Therefore, data mining is utilized to extract valuable information from large databases (Jabbar M A et al., 2013). Clinical decision support systems for predicting heart disease have been developed using data mining and machine learning techniques. Data mining technology can enhance health policy development, reduce hospital errors, enable early disease detection, promote disease prevention, and reduce preventable hospital deaths.

1.2 PROBLEM STATEMENT

Most traditional medical diagnostic methods rely on doctors' clinical experience and expertise to make judgments (Patel 2002). Due to differences in personal experience, different doctors may make different heart disease risk assessments for the same patient's data. This may reduce the accuracy of predicting heart disease risk when facing different groups. The traditional method requires testing more health indicators, which will increase the medical expenses that the examinee needs to bear. The machine learning method can reduce the testing of additional health indicators, thereby reducing medical costs (Javaid M 2022). Moreover, using traditional methods for prediction comparison relies on structured and organized predictor variables, which underutilize the rich unstructured information in electronic medical records (such as clinical doctors' free text records). And this unstructured information may contain valuable details for predicting heart disease, such as specific descriptions of patient symptoms and other

special situations. Therefore, when faced with complex heart disease risk prediction problems, ignoring this information may lead to inaccurate predictions (Desai R J et al. 2020). Effective prevention and management are crucial for early and accurate prediction of heart disease risk, with the potential to save millions of lives each year.

1.3 RESEARCH OBJECTIVE

1. To harness the capabilities of machine learning in identifying the optimal parameters to develop heart disease risk prediction models.
 2. To identify the key factors that may influence the occurrence of heart disease, based on the characteristics of the dataset.
 3. To use methods to obtain evaluation indicators and training prediction time for different models in various situations to evaluate the efficiency of the model in the experiment, the most suitable prediction model for predicting the experimental dataset is selected.

1.4 RESEARCH SIGNIFICANCE

This study will do comparison of different models, and provide the corresponding insights for future heart disease prediction data preparation practice. The findings of this study are expected to offer valuable insights for the development of more effective clinical decision prediction models. By identifying the most significant risk factors, the study aims to provide targeted recommendations for future heart disease prevention, ultimately assisting healthcare providers in formulating more efficient prevention strategies. This study will also provide suggestions for improving data collection and processing in future research. By conducting this comprehensive study, the goal is to advance the field of heart disease prediction through the application of machine learning, ultimately enhancing the accuracy and reliability of heart disease predictions. This will better support clinical outcomes and contribute to the improvement of public health strategies.

1.5 THESIS STRUCTURE

This article is divided into five chapters, systematically addressing the problem of using machine learning methods to predict heart disease.

In the introduction section of Chapter 1, this article provides an overview of the research, including the background of heart disease, problem statement of traditional prediction methods, necessity of using machine learning, research objectives, significance, and overall structure of the paper. By emphasizing the importance of accurately predicting heart disease and the role that machine learning can play in achieving this goal, it lays the foundation for this study.

In the related work section of Chapter 2, existing literature on the use of machine learning for disease prediction was reviewed, with a focus on heart disease. The article discusses various machine learning types and their applications in predicting heart disease, including research comparing the efficiency of different methods. It also covers related concepts such as chi square test, data classification, 10-fold cross validation, and experimental evaluation indicators. This chapter is the foundation for understanding the current status and research methods in this field.

In Chapter 3, this article provides a detailed introduction to the experimental process of predicting heart disease, and introduces the research dataset, including its feature attributes, attribute classification, statistical information, and handling of missing values. This chapter also introduces techniques such as random resampling, chi square test in Python, dataset standardization, proportional partitioning of training and testing sets, 10x cross validation in Python, and implementation of machine learning in Python. This comprehensive approach ensures the reliability and effectiveness of the research results.

In Chapter 4, this article introduces the final research results of the experiment, including identifying the features that have the greatest impact on the final results through chi square test and visualization, discussing the parameter selection of different machine learning types, and introducing the model indicators used to evaluate machine learning and the model structure of some machine learning methods. By further analyzing the experimental results of each model and comparing the results of different types, a summary of the experimental results is provided.

In the final chapter 5, the experiment summarized this study, emphasizing the

key factors affecting heart disease risk identified through experiments and the results evaluation of different machine learning methods, highlighting the excellent efficiency and strong generalization ability of the XGBoost model. This chapter also acknowledges the limitations of current research and proposes future research directions, such as improving models, exploring new data sources, and enhancing the clinical practicality of predictions.

LIBRARY FTSM

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

Although there is a large amount of data in the healthcare system, there is a lack of corresponding analysis tools to explore the nonlinear relationship in the data. The answers to business questions, and prediction of various diseases can be answered and predicted through and data mining tools. It is important in heart disease prediction (Nikhar S & Karandikar A M 2016).

It is based on data, statistics, clinical record and hospital management. The health industry is a multi an enormous billion-dollar field (Kashyap A, 2018). It is highly important in medical data analysis and knowledge extraction. Researchers have been interested in incidence and mortality of heart disease. So they have conducted many studies to lower incidence rate (Patel, J et al. 2016).

2.2 MACHINE LEARNING FOR HEART DISEASE PREDICTION

Literature suggests that machine learning can be applied in healthcare field that changes day to day. Now its powerful tool in combating and predicting diseases and has become a focus.

Beunza et al. (2019) explore how several machine learning algorithms were capable of predicting clinical events in particular study of the risk of coronary heart disease. And they used a detailed comparison of decision trees was performed on the Framingham Heart Research Database. traditional logistic, RF, SVM and neural networks. For this, R-Studio and RapidMiner software are applied. They made some remarkable discoveries in data analysis. They found that neural networks and SVM in different data processing performed exceptionally well. The neural network in R-Studio

had an AUC of 0.71, In another example, when performing in RapidMiner, the support vector machine having an AUC of 0.75. In addition to improving diagnostic and predictive capabilities, this study not only unleashes the potential of machine learning algorithms, but also demonstrates their ability in traditional regression techniques for diagnosis and prediction.

Nikhil Bora et al. (2021) set out to explore the possibility. The ultimate goal is to use methods for predicting cardiovascular disease from early detection of the disease will help to reduce mortality. The first, came from the well-known UCI machine learning library, had 303 records and 14 attributes, while the second, a comprehensive dataset from Kaggle, contained 1290 records AND attributes rich in several medical parameters of patients, along with 12 attributes. Multiple machine learning methods were employed, including LR, SVM, RF, NB, KNN and XGBoost. They employed the Python programming language and Jupyter Notebook for data analysis and model training and a few of the libraries for data They turned up some interesting findings. The testing accuracy of 92% was reached with the SVM model. They found that the RF model achieved the highest testing accuracy, reaching 94.12%. They combined the datasets and tested their random forest model that remained the top performer. Of course, the authors also mentioned some of the key things that impact heart disease, including cholesterol, blood pressure, age and gender. In the conclusion, they emphasized that machine learning methods are effective in predicting heart disease. and selecting corresponding features helps improve the accuracy of the model.

In the Framingham heart disease dataset, Khouadji (2024) compares eight machine learning classification algorithms. In the research, the selection of probability cutoff points to transform the regression models into classifiers was explored, this was done to evaluate the generalization ability and sensitivity of bias of these algorithms for four types of training/testing scenarios. For example, algorithms already investigated for this thesis include XGB, SVM, RF, Logit, LD, QD and the newly proposed algorithms (DDS1 and DDS2). He found diverse results, XGB and SVM performed poorly in the case of imbalanced training dataset, while the DDS1 algorithm was performing well on all of the training/testing scenarios, especially when the training dataset ratio exceeds 40, keeping the stable TPR more than 75%. A novel method for extracting the optimal variable hierarchy of classification algorithms was also proposed and validated on Framingham data population, males and females, to give a robust

technical foundation for disease and epidemiological applications.

Hasan (2021) then compared and analyzed the applications of some machine learning algorithms in predicting potential heart disease. In this research, an efficient cardiovascular disease prediction model is built over the medical data and historical information to assist decision for lifestyle change and continuous medical supervision, in order to lower incidence rates and mortality of cardiovascular diseases. KNN, DT, NB, LR, and RF were used in the experiment, the advantages and disadvantages and applicability of these algorithms are compared to other algorithms that assist in predicting heart disease through research. Different data preprocessing and feature selection methods, such as Backward Elimination and RFE are used to optimize each algorithm. In particular, during its data preprocessing stage the study used substitution algorithms to process missing and outliers and standardized the features to avoid inaccuracy and instabilities in the model. During the phase of model evaluation, the study performed a detailed evaluation of the predictive efficiency of each model using various performance indicators. The RF algorithm was predicted to be the most correct, which is 98.17% accuracy, and the KNN is 87%. This study's research result proves that machine learning technology has a potential application on heart disease forecast. Optimizing models and data preprocessing methods further improves prediction accuracy and are strong supports of clinical decision making and public health strategies. Overall, this study provides a very useful reference for future heart disease prediction work and clearly demonstrates the broad potential applications of machine learning technology in medicine.

In the attempt to predict heart disease using Framingham dataset, Mahmoud W A et al. (2021) have used compared different machine learning and data mining techniques. This study aims to predict the heart disease with several machine learning classification methods like SVM, DT, KNN, LR, RF using 10-fold cross validation resampling methods. Then use various valuable experimental indicators to validate the methods used in the experiment. The accuracy of classification using KNN, SVM, DT, LR, RF methods were 83.95%, 84.50%, 84.82%, 84.89% and 85.05 percent, respectively. RF algorithm performed well for predicting Framingham dataset with an accuracy of 85.05%. In addition, the study shows that, for the predictive results of machine learning types, preprocessing steps such as data standardization and missing value processing can play a significant role. However, it was shown that smaller

datasets (4240 records) are more correctly predicted by the RF method rather than other algorithms. Thus, this paper demonstrates that such technology has great potential in heart disease prediction through comparison with other technology, and advantages using RF algorithm in the processing of high dimensional data and obtaining high forecast accuracy. The result of this research is serving as a reference for conducting new research into the prediction of heart disease as well as selecting and optimizing models for application in these fields.

According to Kumar et al. (2022), a model for predicting heart disease with the aim of reducing mortality was developed through the use of different machine learning methods in forecast. Multiple algorithms were used, like LR, SVM, KNN, DT, RF and XGBoost. They then preprocessed the data (cleaned, and selected features), trained the models on 80% of the data and tested on 20%. Results of these experiments showed that the RF method had accuracy is 95.16% and LR, NB, XGBoost had 85.25% and 74.5% accuracy respectively. The efficiency of KNN with an accuracy of 67.21% was not very good. From the experiment is able to demonstrate the potential of machine learning methods in heart disease forecast as well as how to properly feature select and preprocess data.

An efficient heart disease diagnostic system based on machine learning has been created (Ramma & Salman 2022). For heart disease prediction, the system employs classification techniques such as SVM, NB, and KNN. The study used data from the University of California, Irvine (UCI) database and data preprocessing and five-fold cross validation methods to optimize model efficiency in the data preprocessing before and during the model learning and testing stage such that can avoid selecting the same values in both these stages. Missing values are studied and processed in the dataset during the data preprocessing stage, as well as features are standardized. This method has boosted the precision and reliability of the model. Experiments have shown that the NB algorithm has the best efficiency of all machine learning algorithms with a prediction accuracy of 97%. In addition, SVM and KNN algorithm show prediction accuracy of 87.4% and 95.1%, respectively. This research results show that NB algorithm is a powerful solution for heart disease forecast, particularly in presence of noisy and incomplete data. Moreover, the study suggested that appropriate data preprocessing and cross validation technique can greatly improve the efficiency of machine learning methods in order to produce more accurate and fast suggestion to

clinical diagnosis. By showing the enormous potential of these machine learning based heart disease prediction systems they provide important reference for further development of such systems in practical medical applications.

In their review, Gupta and Seth (2022) compared different machine learning and deep learning algorithms on heart disease prediction. Framingham Heart Disease and UCI Heart Disease dataset were used to evaluate multiple methods such as DT, RF, KNN, SVM, MLP and optimized these algorithms through hyperparameter tuning. The Framingham dataset revealed that the RF method obtained the highest prediction accuracy (97.13%). the UCI dataset showed that the multi-layer perceptron and SVM resulted with the better accuracy rates (86.89%) on each of the datasets. In the experiment, the author pointed out the different factors that affect the occurrence of heart disease in two different datasets. This work concludes that RF and multi-layer perceptions work well in heart disease forecast. The analysis of feature importance will help to find the key risk factors, to improve the accuracy and reliability of early detection of the heart disease, and supply an important reference to the future heart disease forecast and prevention.

Using logistic regression models, Nishadi (2019) tries to predict the risk of heart disease and uncover the most predictive factors. The main purpose of this experiment is to use of techniques particularly LR for the detection and evaluation of earlier risk of heart disease, thus avoiding associated health complications and medical costs. The data accordingly will be from Framingham Heart Disease dataset in Kaggle, containing 4238 records and 15 features for data analysis and model validation by Python in Jupyter Lab. As research, the sequential of research steps are data acquisition, data preprocessing, picking out machine learning types and performing a data analysis, and at last, by regression analysis come to know which the most predictive factors are. The model results indicate that the risk of heart disease increases by about 7 percent for each year of age in males, but not in females. Furthermore, an increase in daily smoking and rise in systolic blood pressure also markedly increases the chance of heart disease. The model has an overall accuracy of 87% and AUC of 73.5%. The research result states that the LR model can predict and identify the main risk factors of the likelihood of heart disease. This study is thoroughly supportive to and provides ideas for future use of machine learning technology in heart disease forecast that will help in early detection and prevention of heart disease and consequently increase the survival rate and quality

of life of heart disease patients. The LR model is employed to assess the risk of heart disease and identify the key factors influencing the prediction of heart disease risk.

In 2021, Kwakye and Dadzie used different machine learning algorithms and identified the most relevant predictive variables to a coronary heart disease (CHD) prediction. Framingham Heart Disease dataset from Kaggle the world most famous heart disease database, containing over 4000 records and 16 attributes was used to conduct the experiment. They outline the process as research process including data acquisition, data preprocessing, feature engineering, and model developing. In the study the original dataset was analyzed by the LR, NB, RF, SVM, NB and KNN model, with LR and NB doing best on the original dataset. Once the dataset has been balanced using SMOTE, the RF model was the best with an AUC of 0.946337. The experiment also identified several factors that had the greatest impact on the experimental results. Overall, they find that with reasonable data preprocessing and feature selection, machine learning methods can be taught to predict the risk of coronary heart disease at highly competitive efficiency levels using either a logistic regression or a random forest model. This research proves the clinical value and the values that are of importance for applications, and the results offer effective clinical tools for clinical doctors to prevent and detect coronary heart disease earlier.

The logistic regression models used by Ambrish G et al. (2022) were used to predict the risk of cardiovascular disease (CVD) and determine the most important predictive factors. To reduce mortality, they study the development of an efficient method for predicting early diseased presence in heart diseases so that they can be detected and treated in early stages. The UCI Heart Disease dataset was studied with 13 features and 303 records, data acquired, data preprocessed, feature selected and model developed. Data preprocessing mainly means to clean data and to process the missing value, feature selection is based on has a high positive correlation with the output value. When the dataset is split in a 90:10 ratio, the LR model has the highest accuracy, at 87.10%. LR models have been proved to be able to predict the risk of disease occurrence by reasonable preprocessing and feature selection. As the proportion of training data increases, the model also gets more accurate and the research results give strong supporting and reference to the preliminary prediction of cardiovascular diseases using machine learning technology so as to enhance the early detection and prevention strategy, and improve the patient life and survival rate.

The purpose of the experiment is to use different methods for predicting heart disease and evaluating the effectiveness of those models in heart disease forecast. A heart disease dataset was obtained from Kaggle (Yahaya et al. 2020), comprising 70000 patient records and 12 features such as age, gender et al. For data preprocessing and scaling, the efficiency of the model was improved by applying the k-modes clustering. The machine learning methods considered in the study include RF, DT, MLP and XGB. The use of the RF method resulted in the highest accuracy (89.01%) for heart disease prediction among the classifiers. In the experiment, the author also provided the dataset features that could most affect the experimental results. Finally, they conclude the research conclusion — that using reasonable data preprocessing and feature selection, machine learning methods can markedly increase the accuracy with which heart disease is predicted. More importantly, the random forest model has exhibited very good results in several experiments, and this may indicate the use of the model in forecasting heart disease. This study is significant in that larger, more diverse datasets have been used to improve universality of prediction results, which provides useful information for early detection and prevention of heart disease. The research findings have also offered effective tools for clinical doctors to effectively conduct early diagnosis and prevention of heart disease through which the survival rate and quality of life will be improved for patients. The biggest contribution of this work is that it demonstrates the use of machine learning methods for predicting heart disease, as well as feature importance analysis to highlight high risk features in the heart disease forecast model.

In 2020, Chauhan explored using multiple machine learning methods to predict cardiovascular disease (CVD). The target of the study is to create a heart disease forecast system that helps to improve the patient's treatment outcomes and lower mortality. For this study, they used the provided dataset, Framingham that has 4238 patient record and 14 feature. Data cleansing (data cleaning), standardization, and dealing with missing values are done as parts of data preprocessing, and they use feature selection and feature engineering techniques to extract meaningful predictive variables from the data. Here for example, they are using LR, KNN, SVM, DT, and RF. Results from the experimental results indicated that LR model could predict heart disease, which had the accuracy of 89%; KNN and SVM could also predict heart disease, the accuracy is 88%. During the research process, the author also proposed some key

factors that can affect the experimental results. The conclusion of the research is that through simple preprocessing and feature selection the efficiency of machine learning methods in predicting heart disease can be significantly improved and especially, LR models perform well in multiple experiments, illustrating their potential in heart disease forecast. This study makes the main contribution to demonstrate the application of different methods to predict the heart disease and show the importance of key factors by feature importance analysis. The research findings provide the clinical doctor with effective means to enhance early detection and prevention strategies of heart disease to adopt improved survival and quality of life for patients. The research results can serve as the important reference and technical support in application of the machine learning technology for the future prediction of heart disease.

The key challenge of early prediction of heart disease, one of the leading causes of death worldwide, is addressed by Ramanathan G. and Jagadeesha S.N. (2023). This study aims to evaluate different machine learning methods to assess which machine learning method is best at predicting coronary artery disease (CAD). Researchers have used Kaggle's dataset and implemented algorithms such as LR, DT, RF, AdaBoost, gradient enhancement, extreme gradient enhancement, optical gradient enhancement and support vector machine. Use certain performance metrics to evaluate. Boosting algorithm robustly performed on both datasets, while decision trees performed very well on the Cleveland dataset. This study finds that accurate early prediction is important to prevent the risk treatment and assist with diagnosis. More research, particularly with the use of deep learning techniques, is needed to make the predictive models in the medical field more effective and reliable, the author suggests.

The objective of the study is to assess how accurately different methods can predict heart disease and identify the most effective one. The study uses a heart disease dataset that is taken from UCI machine learning library which has 303 records and 14 features (Abhisek Acharya 2017). The steps of data preprocessing are: data cleaning and standardization processing and feature selection. The machine learning types implemented include KNN, SVM, NB, XGBoost, RF and ANN. Experimental results exhibit that ANN and SVM respectively showed the best accuracy in predicting heart disease of 91 and 85 percent. Finally, based on reasonable amount of data preprocessing and feature selection, it concludes that machine learning methods can enhance significantly the prediction accuracy of heart disease, especially the superior

efficiency of ANN and SVM in multiple experiments, indicating that they can perform the prediction of heart disease well. The major contribution of this study is to make use of several machine learning types in predicting heart disease and identify the key risk factors using feature importance analysis. The research results enable clinical doctors to develop better early detection and prevention strategies for heart disease that can increase patient survival and quality of life.

Latifah et al. (2020) investigate the effectiveness of LR algorithms and RF algorithms for heart disease classification in order to discern the greatest advantage in preying heart disease. The dataset of the experiment is a publicly available Cleveland Heart disease dataset that had multiple heart disease related features. A feature selection and data preprocessing steps were used to optimize the model inputs. The author has used the RF method and the LR method for constructing a heart disease classification model based on research methods. Model construction and efficiency evaluation were done using the Python programming language and its associated data analysis libraries. They chose a publicly available dataset containing multiple heart disease related features where the model input was optimized through feature selection and data preprocessing steps. Random forest accuracy is found to be 84.4% and LR accuracy is 85.04%. Furthermore, the study also points out that a number of major factors contribute to a propensity towards heart disease. It turns out that LR method is more accurate and is capable of processing complex data that logistic classification algorithm was in the next test. This study's contribution in the sense that it offers more efficient diagnostic and preventive approach at early detection of heart disease and also kick start machine learning application in the medical field. The summary of the above literature is shown in Table 2.1.

Table 2.1 Summary of Reference Views

Citation	Dataset	Model	Result	Limitations
Beunza JJ et al. (2019)	Framingham Heart Study	Decision tree, random forest, support vector machine, neural network, logistic regression	Neural networks and support vector machine performed best across different software	Some obvious predictors of CHD, such as LDL-cholesterol blood levels, to be continued...

...continuation

			(R-Studio and RapidMiner), showing potential to enhance traditional regression for heart disease prediction.	were not included in the dataset, limiting model accuracy.
Nikhil Bora et al. (2021)	Cleveland Heart Disease Database	Naive Bayes, Decision Tree	Decision trees outperformed Naive Bayes and achieved higher classification accuracy. The study speculated that the improvement in accuracy may be due to the increase in the number of attributes used.	The overfitting or underfitting problems of the model have not been explored in depth, and the stability and generalization ability of the model need further study.
Kahouadji (2024)	Framingham CHD	Logistic regression, random forest, extreme gradient boosting, support vector machine	For logistic regression and random forest regression, good classification results can be achieved in different training/testing scenarios.	The research mainly focuses on the prediction performance and variable selection of the algorithm, and less on the parameter optimization of the algorithm.
Nishadi (2019)	Framingham Dataset	Logistic regression	The accuracy of the model is 0.87,	The study used only one to be continued...

...continuation

			but the sensitivity is low (the ability to predict diseased samples is weak). The area under the ROC curve is 0.735, indicating that the model has a certain level of classification accuracy, but there is still room for improvement.	algorithm and did not compare it to other machine learning algorithms, making it difficult to determine the strengths and weaknesses of the model in predicting heart disease risk.
Hasan (2021)	Framingham Cleveland	Logistic regression, decision trees, random forests, Gaussian naive Bayes	Gaussian Naive Bayes performs best with an accuracy of 91.2%.	The experiment did not perform hyperparameter optimization, which affected the performance of the model.
Mahmoud WA et al. (2021)	Framingham Dataset	Logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine	Taking all indicators into consideration, the true positive rate of the random forest algorithm is 84.8%, which outperforms other algorithms on this dataset.	The method of handling missing values in the dataset (mean imputation) may bias the results.
Kumar et al (2022)	UCI Dataset	Logistic regression, Naïve	The accuracy of random forest is	The dataset is relatively small
				to be continued...

...continuation

		Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, Extreme Gradient Boosting	the highest, reaching 95.16%. Overall, the random forest algorithm performs the best on this dataset.	and may not cover all features and situations related to heart disease, which limits the generalization ability of the model.
Rahma and Salman (2022)	Cleveland Dataset UCI Machine Learning Library	Naive Bayes, KNN, Support Vector Machines	The Naive Bayes algorithm performed best with an accuracy of 96.9%	There may be limitations in the data partitioning method and model selection in the experiment, and other better data partitioning ratios and more machine learning algorithms have not been fully explored.
Gupta and Seth (2022)	UCI Machine Learning Library Framingham Heart Dataset	Decision Trees, Random Forests, K Nearest Neighbors, Support Vector Machines	On the UCI dataset, SVM performed best with an accuracy of 86.89%; on the Framingham dataset, random forest performed best with an accuracy of 97.13%.	The algorithms used in the experiment are relatively limited and do not involve more emerging or complex algorithms.

to be continued...

...continuation

Kwakye and Dadzie (2021)	Framingham Heart Dataset	k-nearest neighbors, support vector machines, decision trees, logistic regression, naive Bayes, random forests	Random Forest performed best with an accuracy of 0.946. The original unbalanced data model generally had overfitting, while the balanced data model was mostly well fitted.	The study only compared a limited number of machine learning algorithms and did not cover all potentially effective algorithms.
Amrish G et al. (2022)	UCI Machine Learning Library	Logistic Regression	Used logistic regression on UCI data, found 90:10 split gave highest accuracy, showing preprocessing boosts CVD prediction accuracy.	The study only used one algorithm, logistic regression, for the experiment, making it difficult to determine the advantages and limitations of this algorithm.
Yahaya et al. (2020)	Cleveland Heart Disease Dataset Framingham Dataset Blue Mountain Eye Study Database	Naive Bayes, Decision Trees, Support Vector Machines, K Nearest Neighbors, Artificial Neural	NB performs well in predicting heart disease, with an accuracy rate of 86.6% - 94.80%.	There are deficiencies in the feature selection and model building process, which affect the

to be continued...

...continuation

		Networks, Logistic Regression, Random Forests	accuracy and reliability of the model.
Chauhan (2020)	Framingham Dataset	Logistic regression, KNN, SVM, decision tree, random forest	The logistic regression algorithm performed best among all algorithms, with an accuracy of 89%. There is still room for further improving the algorithm performance, such as tuning hyperparameters and improving feature selection methods.
Ramanathan G. and Jagadeesha S.N. (2023)	Cleveland Dataset Framingham Dataset	Logistic regression, decision trees, random forests, support vector machines, K nearest neighbors, extreme gradient boosting	The decision tree model has the highest accuracy of 0.98 in the Cleveland dataset, and extreme gradient boosting has the highest accuracy of 0.88 in the Framingham dataset. The performance of the algorithm may be affected by hyperparameter settings and data preprocessing, and different settings may lead to different results.
Abhisek Acharya (2017)	UCI Machine Learning Library	k-nearest neighbor, support vector machine, random forest, naive Bayes, Adaboost,	Compared algorithms from UCI data, artificial neural networks and SVM had best The relatively few ways to split the dataset result in an incomplete to be continued...

...continuation

		artificial neural network	accuracy, and identified key predictors.	evaluation of model performance.
				To be continued...
Latifah et al. (2020)	Framingham Heart Study	Logistic Regression random forest	Compared logistic regression and random forest on Framingham data, logistic regression was a better choice due to complex data handling and higher accuracy.	The dataset only comes from residents of Framingham and may not fully represent the situation of all heart disease patients, which has certain limitations.

2.3 INTRODUCTION OF CHI SQUARE TEST

The Chi-square test is a statistical approach used to evaluate the independence of categorical variables. In feature selection, the Chi-square test is employed to assess the relationship between each feature and the target variable (in this experiment, it tests the correlation between each feature and the risk of heart disease within 10 years in the sample). The calculation method of chi square statistic is as follows:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

Among them, O_i is the frequency of observation, E_i is the expected frequency, X^2 is the Chi Square value. The result of the chi square test includes a Chi-square value and a p-value. A larger Chi-square value indicates a stronger correlation between the feature and the target variable. The p-value is used to assess the significance of the correlation, with a significance level (such as 0.05) typically chosen to determine whether the correlation is statistically significant (Vikas P K 2021).

2.4 DATA PREDICTION

It's actually composed of two separate steps — learning step (building a model based on provided training data) and classification step (model predicts class labels given data). As a result, the classifier is very accurate only if the given training data is accurate. For this experiment, the study will split the dataset into several proportions based on the modeling settings, as this helps experiment to setup the model. Machine learning model training and evaluation (Birba D E 2020) is commonly used with splitting the dataset to use for training and testing using. The idea is to divide dataset into training set and testing set, train the model using the training set and since experiment can't train the model with part of the dataset, evaluates the model using the testing set. As the model don't see the data in the test set during training it can then effectively evaluate the model's capacity to generalize to unseen data. Experiments employ a test set to determine if the model is over fitting and take actions (e.g. regularization, decrease model complexity etc.) in order to improve the model's generalization ability (Myung, I. J., 2000). An opportunity is presented to explicitly segment a dataset, yielding an objective evaluation criterion to run experiments against and compare the effectiveness of various models and eventually select the best model to apply.

2.5 10-FOLD-CROSS VALIDATION

10-fold cross-validation is one of the most widely used types of cross-validations, which break the given dataset into ten equally small chunks and continue 10 rounds of train and testing. By doing this, each sample can act as both a portion of the training set and a portion of the testing set throughout the entire process, using data to the fullest extent, therefore the model can more accurately and objectively learn about its efficiency through many rounds of data results (Baumgart M 2024).

2.6 MACHINE LEARNING MODEL

For this experiment will use 6 machine learning methods to predict heart disease dataset like LR, RF, SVM, NB, DT and XGBoost. Below is a detailed introduction to the five

classification methods with mathematical formulae accompanying them.

It is well known that LR is one of the most popular statistical methods for binary response data modeling. This is particularly appropriate when the response variable is binary (with outcome 1 signifying success or the outcome of the event and outcome 0 indicating failure, or nonsuccess of the event). The LR model calculates the probability of an event occurring using a logistic function and analyzes the logarithmic probability of the response variable through a linear combination of predictive variables (Hilbe J M, 2011). The model formula can be expressed as:

$$\text{Ln} \left(\frac{1-P(Y=1)}{P(Y=1)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (2.2)$$

Accordingly, among them, $P(Y=1)$ is the probability of the event and standard of β_0 is the intercept term, $\beta_1, \beta_2, \dots, \beta_n$ is the regression coefficient for each predictor variable X_1, X_2, \dots, X_n and predict the predictor variable (LaValley 2008). The regression coefficient of LR is called the odds ratio, the unit change of an explanatory variable has an impact on the probability of an event occurrence. For example, a positive regression coefficient for a given feature means that if it grows, the likelihood to achieve an event increase (Ambrish G et al. 2022).

The efficiency of the model is also greatly dependent upon the model parameters. In the LR model of this experiment, the parameters that need to be adjusted include regularization strength, penalty, solver, and L1/L2 ratio (l1_ratio). The regularization strength parameter C is used to control the complexity of the model and balance the relationship between overfitting and underfitting. The smaller the value of C, the greater the regularization strength, and the model tends to be simplified. The larger the C value, the more the model will fit the training data (reducing the effect of regularization) The 'penalty' parameter defines the type of regularization, which affects the sparsity and stability of the model by selecting L1, L2, or Elastic Net regularization. Among them, L1 regularization (Lasso) sparsifies the weights to zero, allowing for feature selection. L2 regularization (Ridge) adjusts the complexity by penalizing the sum of squared weights to make the model tend towards smaller weight values. Elastic Net regularization combines the advantages of L1 and L2 regularization and is adjusted using the l1_ratio parameter. The 'solver' specifies an optimization algorithm for finding the optimal model parameters, where 'liblinear' is suitable for L1 and L2 regularization of smaller datasets, and 'saga' is suitable for efficient processing of large datasets and Elastic Net regularization optimization algorithms. l1_ratio (L1/L2 ratio) is only used

when using Elastic Net regularization to control the weight ratio of L1 and L2 regularization. When $l1_ratio=0$, it is equivalent to L2 regularization; When $l1_ratio=1$, it is equivalent to L1 regularization, and if it is an intermediate value, it is a mixture of the two.

SVM is a popular supervised learning algorithm used for classification and regression problems. Vapnik first discovered support vector machines in 1979. It was recommended again by Vapnik for regression and classification in 1995 (Veisi H 2023). SVM maximizes the spacing between data points of different categories by finding the optimal hyperplane, thereby achieving effective classification of data (Cervantes J 2020). For linearly separable data, SVM classifies by finding a hyperplane that maximizes the inter class interval. However, in practical applications, many data are not linearly separable, so SVM introduces kernel functions to transform the data to a higher dimensional space, allowing for linear separation in that space (Roman I et al. 2021). For linearly separable data, the objective of SVM is to identify a hyperplane that maximizes the margin. Given a training set (x_i, y_i) , where x_i is the feature vector, $y_i \in \{-1, +1\}$ is the label. SVM finds the optimal hyperplane by solving the following optimization problems:

$$\min_{w, b} \frac{1}{2} \|w\|^2 \text{ subject to } y_i (w \cdot x_i + b) \geq 1, \forall i \quad (2.3)$$

Among them,

w is the weight vector that specifies the direction of the hyperplane.

b is the bias term that determines the position of the hyperplane.

x_i is the feature vector.

y_i is the label, with a value of $\{-1, +1\}$.

For linearly indivisible data, introduce relaxation variables ξ_i and regularization parameter C :

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \text{ subject to } y_i (w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i \quad (2.4)$$

ξ_i is a relaxed variable that allows for misclassification of data points.

C is a regularization parameter used to balance the size of the interval and the penalty for misclassification.

SVM can transform data into a higher-dimensional space using kernel functions $K(x_i, x_j)$:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (2.5)$$

ϕ is a mapping function that maps the original data to a high-dimensional space.

Common kernel functions include:

Linear kernel: A linear kernel is a kernel function used in support vector machines (SVM) that directly calculates the inner product between feature vectors. It is suitable for linearly separable data scenarios, with high computational efficiency and strong model interpretability, and is commonly used in applications of high-dimensional data such as text classification. Due to the absence of nonlinear mapping, the decision boundary of linear kernels is a hyperplane, suitable for data that is already linearly separable in the original feature space (Huang H Y & Lin C J 2016).

$$K(x_i, x_j) = x_i \cdot x_j \quad (2.6)$$

Also: Gaussian Kernel. This kernel function is very useful because it can map data in high dimensional space, which can result data that are separable in low dimensional space but not linearly separable (Al-Mejibli I S 2020).

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (2.7)$$

It's based on simple but powerful probability classifier called Naive Bayes (NB) which is a byproduct of Bayesian theorem. Although there may be incorrect situations in reality, naive Bayesian classifiers are often good at classification in practice (Bafjaish S S 2020). In NB, it made decision about the classifying based on conditional probability of each feature for each class. With high computational efficiency and good efficiency in small or high-dimensional data processing, especially for the text classification and medical diagnosis, it is its main advantage. The naive Bayesian model is simple and easy to implement, and has strong interpretability. Naive Bayesian classifier is based on Bayesian theorem:

$$P(\text{Class } j|x) = \frac{P(\text{Class } j)P(x|\text{Class } j)}{P(x)} \quad (2.8)$$

Among them, $P(\text{Class } j|x)$ is what study wants to know, the Posterior probability of class j given a predictor x , $P(x|\text{Class } j)$ is the Likelihood, the probability of the predictor given a Class j . It's computed from the training-set. $P(\text{Class } j)$ is the prior probability of class j , what study knows about the class distribution before study considers x . $P(x)$ is the prior probability of the predictor, in practice, there's interest only in the numerator (denominator is effectively constant) (Bafjaish S S 2020).

DT is a classification and regression model based on a tree structure. By starting from the root node and gradually dividing the data into subsets based on different values

of features, until reaching the leaf node. Each leaf node is associated with a class label. The construction process of a decision tree model includes selecting the best feature as the splitting node and dividing the dataset into two or more subsets based on that feature. The commonly used splitting criteria include information gain, gain rate, and Gini index. Decision trees have advantages such as easy understanding and interpretation, handling non-linear relationships, and no need for feature standardization. Decision trees perform well in handling complex datasets and are an intuitive and powerful classification tool (Charbuty B & Abdulazeez A 2021).

RF is a powerful supervised learning algorithm composed of multiple decision trees. Each decision tree receives a randomly selected subset of features and samples during the training process and independently generates prediction results. Finally, the random forest determines the final prediction value by voting on the prediction results of all trees (classification task) or averaging (regression task). Random forest can handle datasets with a large number of features and make decisions by selecting the most relevant features. It has high robustness against noise and outliers in the data, as the average or voting mechanism of multiple trees can smooth out the instability of individual trees (Pal M & Parija S 2021).

XGBoost is a powerful machine learning method widely used in various classification and regression problems. XGBoost is an optimization implementation based on the gradient boosting framework, which improves prediction accuracy by constructing multiple weak classifiers (usually decision trees). Its core idea is to train each new model on the residuals of the previous model to reduce errors. XGBoost, on the other hand, adds a regularization term so the model will not become more complex than necessary, and will also prevent overfitting. This algorithm's main features are efficiency, flexibility, built in regularization, and fast processing of sparse data (Anbuselvan P 2020).

2.7 PERFORMANCE EVALUATION

Mainly, commonly used performance indication in data analysis and machine learning model evaluation are accuracy, precision, recall, F1 score, AUC, Receiver Operating Characteristic Curve, Confusion Matrix.

Table 2.2 shows that the confusion matrix is used to describe explicitly the

model's performance in classification tasks in TP, TN, FP, FN. Distribution of classification results is provided in detail, which allows us to understand how well the model predicts each category, or rather which category the model fails to predict. Experiment can comprehensively understand what the advantages and disadvantages are in the model by comprehensively evaluating these performance indicators, and chose and optimize the model.

Table 2.2 Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

In these, TN denotes the number of samples that are actually negative (0) and predicted to be negative (0), which means the model will classify patients who do not have a risk of heart disease as negative (0). The number of samples that are in fact negative (0), but predicted as positive (1), is also known as FP, meaning the model marked some of patients without risk of heart disease as at risk, resulting in excessive worries and additional medical test. As a fraction of the total number FN means the number of samples that are actually positive (1) but are predicted to be negative (0), i.e. samples that the model failed to identify as samples at risk of a heart disease. This error is especially dangerous as these patients may miss needed treatment and intervention. TP is the number of samples with true class 1 (actually positive, 1) and predicted with class 1 (also 1), which means that the model is able to distinguish the actual number of patients with high risk of heart disease and thus convey it on time, so that the patients can undergo timely examination and effective treatment.

The confusion matrix of the model can also be obtained by the model through the formula of the corresponding calculation formula, which brings very important reference value (Tharwat A 2021).

Precision index is one of the predictive indicators for measuring the accuracy of learning models, in evaluating machine learning methods, because it measures how well the model predicts Positive Classes. For example, incorrect diagnosis that a healthy

individual has a heart disease risk also means not only unnecessary psychological burden and economic burden, but also needless treatment. In this case, high precision models can reduce the negative impact of the false positives, by reducing the false positives themselves.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.9)$$

The fundamental measure to measure how accurate a model is, is using Accuracy which tells us the proportional value of correctly predicted samples compared with the total number of samples. Unfortunately, on imbalanced datasets, accuracy does not tell us much about how well the model does on minority classes. Accuracy gives us the accuracy of our model in predicting positive classes, i.e., the fraction of correctly predicted positive samples to the total predicted positive samples. In particular, for unpleasant effects of false positives — spam filtering and disease screening to name a few.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.10)$$

The recall rate defines the degree to which the model captures true positive samples, that is, the actual number of positive samples correctly predicted by the model for all actual positive samples:

$$\text{Recall} = \frac{TP}{FN+TP} \quad (2.11)$$

The F1 score is the harmonic average of precision and recall, and it combines the characteristic of both. On imbalanced datasets, we should opt for evaluation metric that balances precision and recall such as the F1 score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.12)$$

AUC is the area under the ROC curve, so it is another highlight of the model capacity to distinguish positive and negative samples. A higher AUC value indicates that the model is better at distinguishing between positive and negative samples. The false positive rate is represented on the horizontal axis, while the true positive rate is on the vertical axis. At a specific threshold, the ROC curve illustrates the relationship between the true positive rate and false positive rate of a classifier. The model's performance at various thresholds can be assessed by observing the shape of the ROC curve. Generally, the closer the curve is to the upper left corner, the better the model's performance.

2.8 SUMMARY

This chapter discusses in depth various machine learning algorithms and methods for predicting diseases. Through the study of feature selection techniques, different data partitioning methods and machine learning algorithms, a thorough experiment was conducted on the prediction of heart disease. In this experiment, the chi-squared test was introduced as a method for feature selection of the experimental data set, and the use of different data set segmentation ratios and 10-fold cross-validation methods for data set segmentation was explained. The machine learning models to be used in this experiment, including LR, RF, SVM, NB, DT and XGBoost, as well as the corresponding model evaluation indicators, were explained, providing more solid support for the experimental results.

CHAPTER III

RESEARCH METHODOLOGY

3.1 INTRODUCTION

In this chapter, the experiment conducted corresponding data processing on the Framingham heart disease dataset collected from Kaggle (<https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset#>).

After previewing the dataset structure, it was found that the corresponding values were missing in the dataset. Subsequently, mean imputation and resampling were used to fill the missing values and balance the corresponding dataset samples.

In order to make the data distribution of the experimental dataset appear clearer and more distinct, a data visualization method was used in the experiment to display the distribution of each attribute in the interval of the dataset in the data histogram. This way, the experiment can have a clearer understanding of the sample distribution in the dataset and make corresponding data processing to achieve sample balance in the dataset.

In addition, in order to more objectively evaluate the ability of various machine learning to predict on this data set, the experiment divided the data set by different methods, including the use of different proportions of test sets and training sets, and the use of 10 times cross validation method for segmentation. At the same time, the experiment also used random search to select the most suitable model parameters so as to improve the prediction ability of various models as much as possible. Chi square detection was also used to select the attributes that most affect the prediction results in the data set, reducing the interference attributes during model prediction, thereby enhancing the model's prediction efficiency.

The specific experimental process is shown in Figure 3.1:

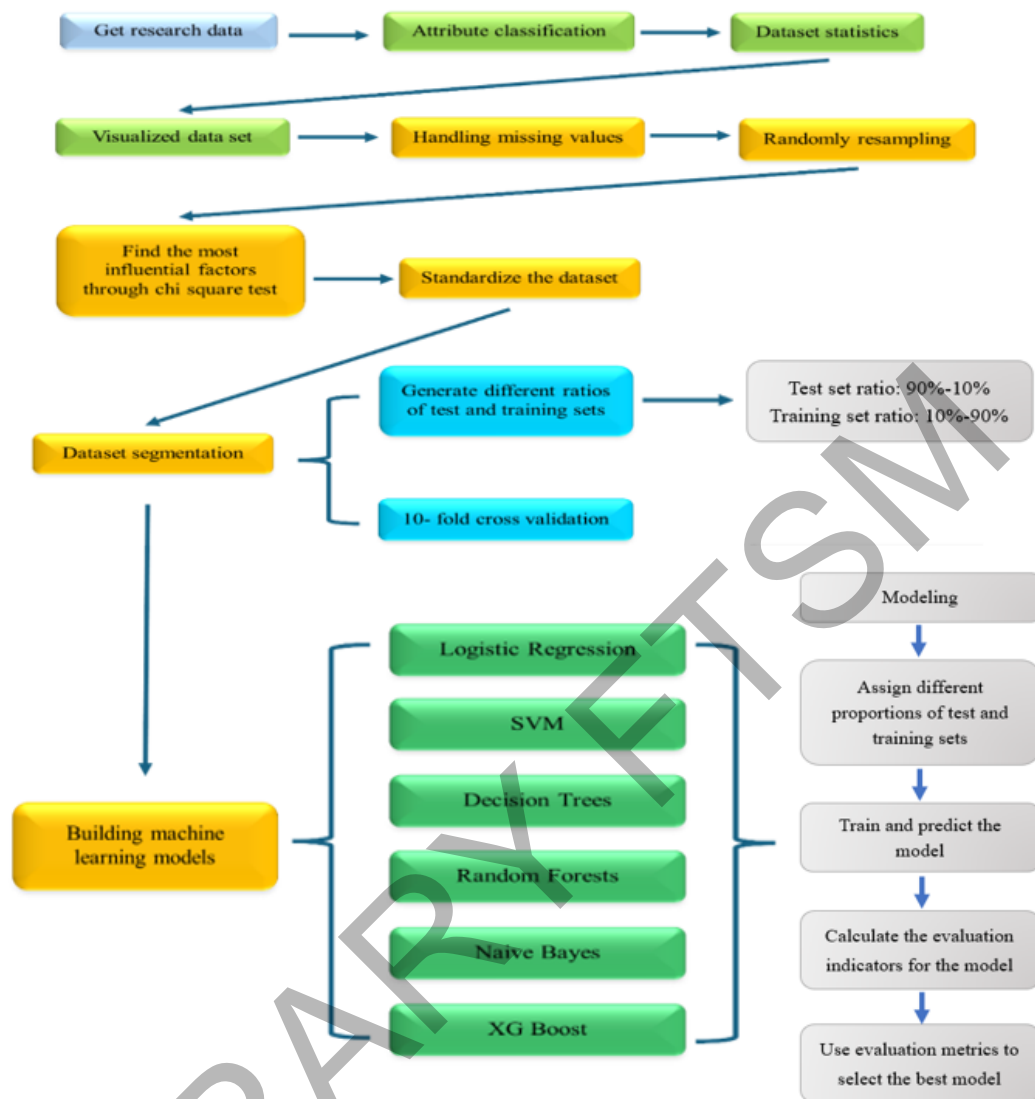


Figure 3.1 Experimental procedure

3.2 EXPERIMENT ENVIRONMENT

Python was selected as the development language for the project in this study because the Scikit learn library in Python has feature selection methods and machine learning frameworks necessary for experiments. In addition, all learning methods are evaluated via a complete set of functions. For instance, timing when calculating model prediction time you can use the time module function in Python. Hence, other module is not needed to be imported for experiments in experiment process during the experiment, making the experiment easier and quicker.

The experiment selected Google Collaboratory, a free, cloud-based Jupyter Notebook environment offered by Google for writing and running Python code, widely

used in the AI field. It allows researchers to conduct valuable experiments without the need to set up complex development environments locally — a browser is all that is required.

3.3 RESEARCH DATASET

3.3.1 BASIC INFORMATION OF DATASET FEATURE ATTRIBUTES

This experiment used the Framingham Heart Disease dataset downloaded from Kaggle, a great resource for researchers to get insight into risk factors for heart disease and other cardiovascular illness. On dataset, it has 4240 records and 15 feature attributes in which column 16 is the required feature attributes for the analysis. Some of these attributes include age, gender, daily smoking volume, heart rate and BMI and other key indicators that may increase risk of having heart disease. The specific contents are shown in Figure 3.2:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	EMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
...
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84.0	86.0	0
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86.0	NaN	0
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80.0	107.0	0
4238	1	40	3.0	0	0.0	0.0	0	1	0	185.0	141.0	98.0	25.60	67.0	72.0	0
4239	0	39	3.0	1	30.0	0.0	0	0	0	196.0	133.0	86.0	20.91	85.0	80.0	0

4240 rows x 16 columns

Figure 3.2 Dataset Example

The following is information about the features included in the dataset. The author of the dataset, Ashish Bhardwaj, provides a detailed description of the features in the dataset, as shown in Table 3.1:

Table 3.1 Introduction to Dataset Attributes

Attribute	Attribute Introduction
Male	A binary indicator (0 represents female, 1 represents male) that represents the gender of the participants.
Age	A number that represents the age of participants in years.
Education	Education level (1-4 education levels increase sequentially).
Current Smoker	A binary indicator (0 represents non-smokers, 1 represents smokers) that details the smoking status of participants.
CigsPerDay	A number that represents the number of cigarettes a participant smokes per day.
BPMds	A binary indicator (0 indicates no antihypertensive medication taken, 1 indicates antihypertensive medication taken) that reflects whether participants are taking antihypertensive medication.
Prevalent Stroke	Display whether there is a history of hypertension (1 indicates yes, 0 indicates no).
PrevalentHyp	Display whether there is a history of stroke (1 indicates yes, 0 indicates no).

to be continued...

...continuation

Diabetes	A binary indicator (0 means nonexistence, 1 means existence of diabetes), representing whether the participant has diabetes.
TotChol	Number, detailing the total cholesterol level (mg/dL).
SysBP	A number representing systolic blood pressure (in millimeters of mercury).
DiaBP	A number representing diastolic blood pressure (in millimeters of mercury).
BMI	A number representing the participant's body mass index.
Heart rate	A number that represents the participant's heart rate per minute.
Glucose	Number, detailing the glucose levels of participants.
TenYearCHD	Risk of coronary heart disease within ten years (0 indicates no risk, 1 indicates risk) is a target variable indicating the presence of heart disease risk.

In this study, to classify the dataset's features, two empty lists, `cate_val` and `cont_val`, were defined to store the column names of categorical and continuous variables, respectively. The experiment traversed all columns in the dataset and classified them based on the count of distinct values in each column using the `data[column].nunique()` method.

If the count of distinct values were less than or equal to 10, the column was considered categorical, and its name was added to the `cate_val` list. Otherwise, it was considered continuous, and its name was added to the `cont_val` list. This method effectively determined the attribute types for each feature in the dataset.

The categorization of the dataset attributes is presented in the following Table 3.2:

Table 3.2 Dataset attribute types

Attribute	Type
Male	Categorical
Age	Continuous
Education	Categorical
Current Smoker	Categorical
CigsPerDay	Continuous
BPMds	Categorical
prevalentStroke	Categorical
prevalentHyp	Categorical
Diabetes	Categorical
TotChol	Continuous
SysBP	Continuous
DiaBP	Continuous
BMI	Continuous
Heart rate	Continuous
Glucose	Continuous
TenYearCHD	Categorical

3.3.2 STATISTICAL INFORMATION ABOUT THE DATASET

To obtain a more comprehensive understanding of the dataset's characteristics, the report will use the `descript()` function to calculate the average and percentage of each field in the dataset. 'Describe()' is a function in the Pandas library used to generate descriptive statistical information for a data box or series. This function is very useful for quickly understanding the basic statistical characteristics of data, especially in data analysis and preprocessing. The detailed calculation results of the dataset are displayed in Figure 3.3:

	nale	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCVD
count	4240.000000	4240.000000	4135.000000	4240.000000	4211.000000	4187.000000	4240.000000	4240.000000	4240.000000	4190.000000	4240.000000	4240.000000	4221.000000	4239.000000	3852.000000	4240.000000
mean	0.429245	49.580189	1.979444	0.494104	9.005937	0.029615	0.005896	0.310613	0.025708	236.699523	132.354599	82.897759	25.800801	75.878981	81.963655	0.151887
std	0.495027	8.572942	1.019791	0.500024	11.922462	0.169544	0.076569	0.462799	0.158280	44.591284	22.033300	11.910394	4.079840	12.025348	23.954335	0.358953
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	107.000000	83.500000	48.000000	15.540000	44.000000	40.000000	0.000000
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	206.000000	117.000000	75.000000	23.070000	68.000000	71.000000	0.000000
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	234.000000	128.000000	82.000000	25.400000	75.000000	78.000000	0.000000
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.000000	1.000000	0.000000	263.000000	144.000000	90.000000	28.040000	83.000000	87.000000	0.000000
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.000000	1.000000	1.000000	696.000000	295.000000	142.500000	56.800000	143.000000	394.000000	1.000000

Figure 3.3 Dataset statistics

3.3.3 VISUALIZATION OF CHARACTERISTIC NUMERICAL DISTRIBUTION

To quickly grasp the data distribution of each feature attribute in the dataset, the experiment visualized the data distribution of the dataset. On the one hand, it is more convenient to observe the distribution pattern, central tendency, and dispersion of the data in the dataset. On the other hand, it can also visually check whether there are outliers in the dataset for replacement or deletion in subsequent experimental steps. Intuitive graphics can more easily explain data features and analysis results, and graphical representations are more persuasive and understandable than simple numerical descriptions.

In this experiment, graph objects will be created by calling the Python plotting library `matplotlib.pyplot`. The total width and height of the plotted graph objects will be defined using the `Figure()` method in the library, and the `gca()` method will be used to obtain subgraph objects for each feature in the dataset. Finally, the `hist()` method in the Pandas library will be used to draw a histogram of the data distribution, the results are presented in Figure 3.4:

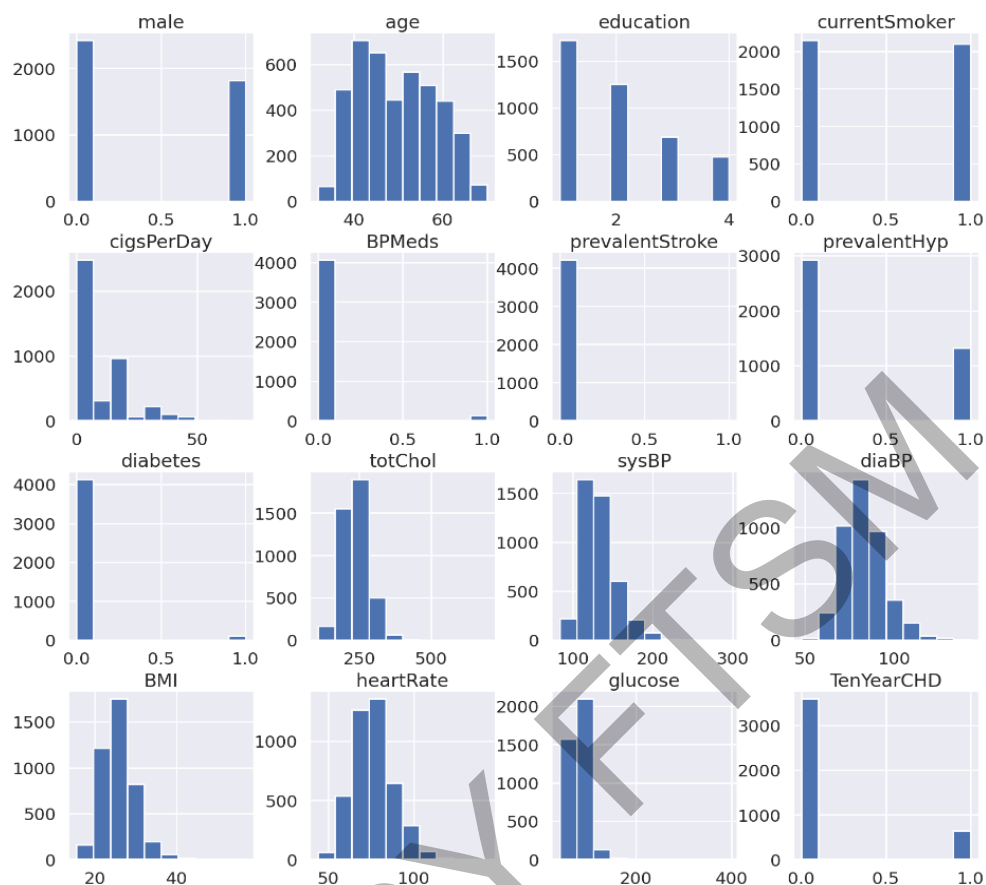


Figure 3.4 Dataset feature data distribution histogram

By observing the histogram of data distribution, it can be concluded that the male to female ratio is relatively even in this dataset, with slightly more males than females. The majority of participants are aged between 40 and 60, showing a normal distribution trend. The education level ranges from 1 to 4, and the majority of participants are concentrated in the two levels of 2 and 1, indicating that the majority of participants may not have a high level of education. The number of smokers and non-smokers in the experimental sample is relatively balanced, with slightly more non-smokers. Most smokers smoke between 0 and 10 cigarettes per day, and a very small number smoke more than 30 cigarettes per day. Most participants have no history of hypertension and have not used antihypertensive drugs, but there may also be a portion of people with a history of hypertension who have not taken antihypertensive drugs. In the data set, the vast majority of subjects have no history of stroke, and there are fewer people with diabetes. Regarding the characteristics of total cholesterol level, systolic blood pressure, diastolic blood pressure, BMI, heart rate, etc., the sample distribution of experimental data shows a normal distribution, that is, within a certain interval, the

number of sample distributions will sharply increase. In terms of blood glucose levels, the majority of participants had blood glucose levels concentrated between 70 and 100, with a few having higher levels. In this experimental dataset, the majority of subjects collected had no risk of developing heart disease within 10 years, with only a few having a risk of developing heart disease within the same period.

These histograms provide important insights into the health status and risk factors of study participants. They revealed the typical characteristics of participants' blood pressure, cholesterol and heart rate, as well as other possible influencing factors related to smoking, diabetes and potential elevated blood glucose levels. The distribution of these attributes helps the experiment to understand the potential relationship between the lifestyle factors of the subjects in the data set and the final results. This visual analysis is crucial for further statistical testing and modeling work aimed at predicting cardiovascular disease risk.

3.3.4 MISSING VALUE INFORMATION IN THE DATASET

In this study, missing values in the dataset were first processed and a Boolean data box of the same size as the original dataset was generated using the `data.isnull()` method, where missing values were marked as True and non-missing values were marked as False. Next, use the `data.sum()` function to separately count the number of missing values for each feature in the dataset, and calculate the data missing rate of the dataset.

Missing data. `data.sum()` represents the total number of missing data in the dataset, while `data.shape[0]` represents the total number of records for all information in the dataset. By running `Missing data. sum()` function can determine the count of missing values present in each attribute, the results are presented in Table 3.3:

Table 3.3 The number of missing values for each attribute

Attribute	Missing value
male	0
age	0
education	105
currentSmoker	0
cigsPerDay	29
BPMeds	53
prevalentStroke	0
prevalentHyp	0
Diabetes	0

to be continued...

...continuation

totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

The proportion of missing values in the dataset can be calculated using the following method:

$$\text{total_percentage} = \frac{\text{missing_data.sum}()}{\text{data.shape}[0]} * 100 \quad (3.1)$$

The ratio calculated between the two can be used to determine the proportion of missing values in the experimental dataset. Preliminary evaluation shows that the entire dataset has varying degrees of completeness, with some attributes such as education, CigsPerDay, BPMeds, TotChol, BMI, heart rate, and glucose showing missing values. After calculation, the overall missing value rate is 15.21%. Missing data may introduce bias in the analysis, leading to underestimation or overestimation of relevant risk factors.

In order to more intuitively display the proportion of missing values for each feature in the dataset, the experiment visualized the ratio of missing values for each attribute relative to the total number of missing values. By calculating the percentage between the number of missing values for each feature and the total number of rows in the dataset, the proportion of missing values for each feature to the entire dataset can be determined.

$$\text{missing_data['Percentage']} = \frac{\text{missing_data['Total']}}{\text{len(data)}} * 100 \quad (3.2)$$

Missing data [' Percentage '] is a new column used to store the proportion of missing values for each feature. Missing data [' Total '] represents accessing the 'Total' column in the Data Frame, which contains the number of missing values for each feature. Len (data) will return the total number of rows in the dataset, representing the overall number of samples. By calculating the proportion of missing values for each feature relative to the total number of rows, and storing the results in a new column called 'Percentage'.

Subsequently, the experiment visualized the data in the missing data box using the sns. barplot function, which is a function in the Seaborn library used to create a bar graph where the name of each feature is displayed on the x-axis and the proportion of missing values for each feature is displayed on the y-axis. Through this visualization

method, the experiment can intuitively understand the proportion of missing values for each feature. The result is shown in Figure 3.5:

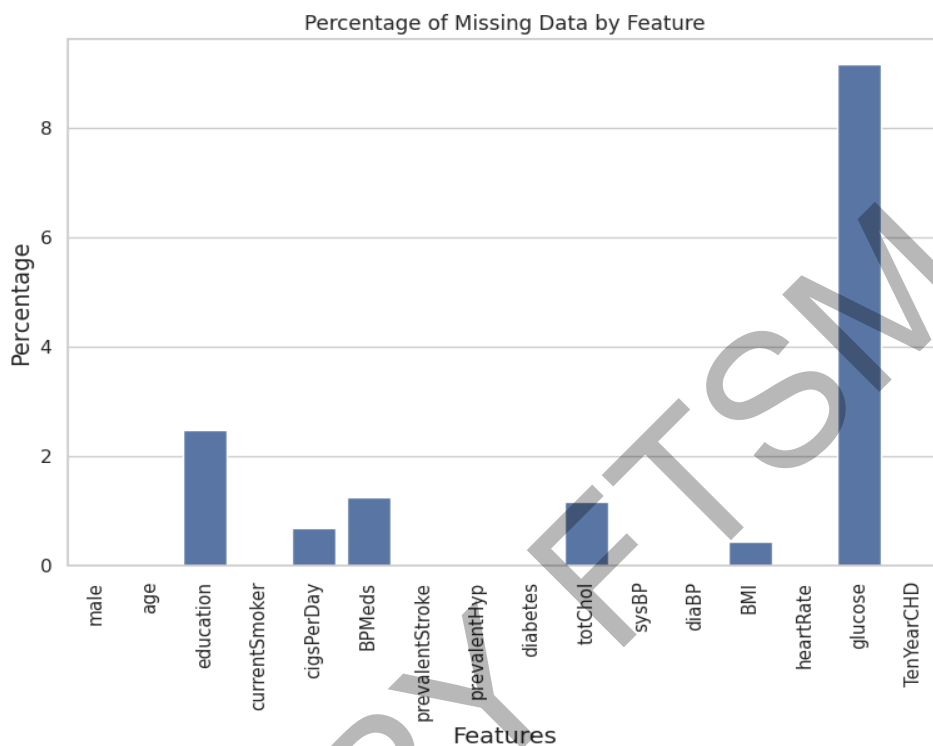


Figure 3.5 The proportion of missing features in each dataset

3.4 DATA PROCESSING

In the following sections of this experiment, the study will conduct in-depth analysis of each attribute. The experiment will provide appropriate methods for handling missing data values of attributes that may affect the overall quality of the dataset, as well as recommendations for improving modeling accuracy and efficiency in the future, in order to prepare for effective analysis and modeling of the dataset.

3.4.1 MISSING VALUE

The structure and properties of the dataset indicate that statistical modeling and machine learning can provide a deeper understanding of cardiovascular health risks. However, handling missing values in the experimental dataset and unifying the range of feature attribute values in the dataset are crucial to ensure the reliability of any analysis or

model developed using this dataset.

It is crucial to address missing data in order to improve the quality and reliability of the Framingham Heart Disease dataset used for cardiovascular disease prediction. In data preprocessing methods, replacing missing values with the average value can preserve more data, maintain the mean and distribution characteristics of the dataset, avoid data loss caused by directly deleting samples containing missing values, and enable the model to use more information for training, improve model efficiency, and reduce bias caused by missing values. Secondly, the average replacement method is easy to understand and explain, and can clearly explain the process and logic of data processing, enhancing the transparency and acceptability of the method.

In this project, the SimpleImputer function from the sk-learn library was used to replace missing values with mean values. Setting the strategy parameter of the SimpleImputer function to 'mean' indicates that the missing values in the dataset will be replaced with the average value. Then, the data is fitted and transformed using the fit_transform method, replacing missing values with the mean of the corresponding features. Finally, check the missing values after replacement using the isnull (). sum() method to ensure that the missing values have been successfully replaced. After the above processing, the number of missing values of each feature in the statistical data set has become 0. The specific changes can be seen in Figure 3.6.

Missing data:		Missing data:	
male	0	male	0
age	0	age	0
education	105	education	0
currentSmoker	0	currentSmoker	0
cigsPerDay	29	cigsPerDay	0
BPMeds	53	BPMeds	0
prevalentStroke	0	prevalentStroke	0
prevalentHyp	0	prevalentHyp	0
diabetes	0	diabetes	0
totChol	50	totChol	0
sysBP	0	sysBP	0
diaBP	0	diaBP	0
BMI	19	BMI	0
heartRate	1	heartRate	0
glucose	388	glucose	0
TenYearCHD	0	TenYearCHD	0
dtype: int64		dtype: int64	
The total percentage of missing data is 15.21%		The total percentage of missing data is 0.0%	

Figure 3.6 Changes in missing values in the dataset

3.4.2 RANDOMLY RESAMPLED DATASET

During the process of visualizing the dataset, it was observed that the proportion of individuals at risk of developing heart disease was lower than that of individuals at no risk within the last 10 years. This means that the number of minority class samples is less than that of the majority class samples. This may lead to the model tending to predict the majority class in the subsequent machine learning modeling process, as it can achieve high accuracy overall, but may ignore the correct prediction of minority class samples, resulting in very low recall and F1 scores for minority class samples and poor ability in identifying minority class samples. In practical applications, this may lead to serious consequences, such as people who were originally at risk of heart disease being misdiagnosed by the model as healthy individuals, which may result in medical accidents.

So, in the experiment, the method of random oversampling will be adopted to modify the data in the dataset. Data resampling techniques such as random oversampling are used in the case of class imbalance problems. Increase the number of replicas of minority class samples so that the number of samples for both minority and majority classes is equal. Random oversampling is a simple, effective method to greatly improve classifier efficiency on imbalanced datasets. To randomly resample some data in this experiment, defined two arrays, target 1, and target 0, to store samples with TenYearCHD values of 1 and 0 respectively. Then, the `resample()` method was used to perform random resampling on the target 1 array, so that the number of samples in the target 1 array was equal to the number of samples in the target 0 array. Finally, the `concat()` method in the Panda library was used to merge the processed target 1 and the original target 0 into a new dataset. In the new dataset, the experiment ensured that the number of risky samples was consistent with the number of risk-free samples, thereby improving the accuracy of the model for disease prediction. The results are shown in Figure 3.7.

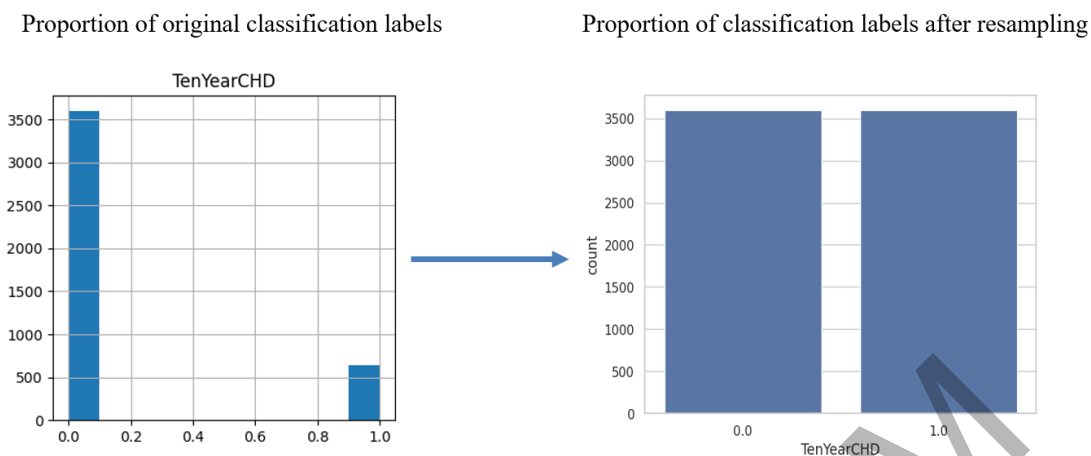


Figure 3.7 Changes in sample size during resampling

3.4.3 CHI SQUARE TEST

In order to effectively improve the modeling efficiency during the experimental process, the study will use the chi square test method to determine the most influential features for predicting heart disease from this experimental dataset, thereby achieving the goal of effectively reducing the amount of data used to train the model. The experiment simplified the experimental dataset by reducing the number of sample features, thus achieving the goal of maximizing the running speed of the training model and model prediction while ensuring model accuracy, thereby improving the efficiency of the model.

In this experiment, to implement chi square detection using Python, it is necessary to import the `chi2()` method from the `sklearn.feature_selection` module to calculate the chi square value and p-value. In order to observe the experimental results more intuitively, the calculated chi square values will be sorted in descending order, and the `plot.bar()` method will be used to visualize the bar chart of the chi square values of each feature and the risk of heart disease in the sample within ten years, so as to clearly and intuitively identify the strength of the correlation between the feature and the target variable.

3.4.4 STANDARDIZATION OF DATASETS

Many machine learning methods perform better when features have the same scale. Therefore, the data needs to be converted to standard deviation based on its mean, so that the distribution of the data conforms to the standard normal distribution, which helps to accelerate the convergence speed of the model and improve its efficiency. If not standardized, features with a larger numerical range may dominate the algorithm (not because they are more important, but simply because they have a larger numerical range).

In this regard, the experiment will use the `fit_transform()` method of the `StandardScaler` function, a standardization tool in the Scikit learn library, to standardize the data after data cleaning. This operation can adjust the mean of the data to 0 and the standard deviation to 1, so that the standardized data has the same scale in statistics, thereby eliminating dimensional differences between different features. The standardized results are shown in Figure 3.8:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1.0	-1.448340	4.0	0.0	-0.780758	0.0	0.0	0.0	0.0	-0.994440	-1.239876	-1.126196	0.201665	0.306649	-0.245463	0.0
1	0.0	-0.634463	2.0	0.0	-0.780758	0.0	0.0	0.0	0.0	0.187865	-0.645424	-0.285687	0.615472	1.549247	-0.278069	0.0
2	1.0	-0.401927	1.0	1.0	0.848098	0.0	0.0	0.0	0.0	0.080383	-0.387829	-0.362097	-0.181577	-0.107550	-0.473700	0.0
4	0.0	-0.634463	3.0	1.0	1.092357	0.0	0.0	0.0	0.0	0.940241	-0.288754	-0.056457	-0.708241	0.720848	0.015379	0.0
5	0.0	-0.983267	2.0	0.0	-0.780758	0.0	0.0	1.0	0.0	-0.285057	1.692751	1.930200	0.984607	0.058130	0.471853	0.0

Figure 3.8 Example after data standardization

As mentioned above, the report has standardized certain element values in the dataset to a standard range, making the data follow a standard normal distribution. However, properties such as Male and Current Smoker, which use 0 or 1 as variables, have not been standardized yet. This is appropriate because the values of these variables are already 0 and 1, and represent categories and should not be standardized.

3.4.5 PROPORTIONAL PARTITIONING OF TRAINING AND TESTING SETS

In this study, the experimental objective is to understand the efficiency of each type of machine learning under different training and testing set ratios, in order to objectively evaluate the model. In order to make the experimental data as accurate and diverse as

possible in terms of results, the experiment selected multiple scenarios for segmentation ratio: starting from 90% of the training set, each experimental scenario gradually decreased in a 10% gradient until reaching 10%, while the training set ratio started from 10% and gradually increased by 10% for each experimental scenario until reaching 90%. This study aims to comprehensively evaluate the ability of the model by manufacturing different model settings. In the experimental project a function `train_test_split ()` of Scikit learning library will be used to randomly divide the dataset to the corresponding proportions. An important activity of this partitioning method is to evaluate the ability of the model on unseen data in order to avoid overfitting and thus raising the model's generalization capacity.

3.4.6 IMPLEMENTATION OF 10-FOLD CROSS VALIDATION

Moreover, to enforce more total effective evaluation bias from data segmentation, the experiment also described the model ability under 10-fold cross validation conditions.

Here the study performed 10-fold cross validation on model using the `cross_validate ()` function from the scikit learn library. The reason of using this function was to find time required for the model to train and predict, the prediction labels and probabilities for each sample by using `cross_val _predict ()`, to later get different model evaluation metrics for the model for the future.

3.5 IMPLEMENTATION OF MACHINE LEARNING

In order to prevent overfitting or poor ability of some models in predicting diseases during the experiment, the key parameters of machine learning types such as LR, DT, and RF were adjusted using random search to optimize model efficiency and prevent overfitting. Random search is a hyperparameter optimization technique that trains and evaluates models by randomly selecting parameter combinations. The experiment defined a parameter list for the above machine learning methods, combined them into machine learning models with different parameters through random search, and then determined the most suitable parameters for the model without fitting by predicting the AUC values generated after training different models. The random search conducted in this experiment used 10-fold cross validation and searched for the optimal parameter

combination in 50-100 iterations. This method effectively improved the predictive ability of the model while maintaining the rational utilization of computing resources.

For LR, the C parameter is set in the parameter list to control the regularization strength and balance the complexity and fit of the model. A smaller C value corresponds to stronger regularization to prevent overfitting. The penalty parameter is set to specify the type of regularization used, including l1 (Lasso) and l2 (Ridge), Elasticnet combines the advantages of both, adjusting through linear combination and setting the solver parameter to determine the choice of model optimization algorithm. Different optimizers are suitable for different types of regularization and data sizes. For example, liblinear is suitable for small datasets, while Saga is suitable for processing large-scale data and Elasticnet regularization. In addition, the l1_ratio parameter set in the list is only used when the model chooses Elasticnet regularization, used to adjust the mixing ratio of l1 and l2 regularization, thus flexibly controlling the sparsity and complexity of the model.

For DT, the max_depth parameter is set in the parameter list to control the maximum depth of the tree and avoid overfitting by limiting the depth of the tree. The min_Samples_split and min_Samples_leaf parameters define the minimum number of samples required for node splitting and leaf nodes, respectively, to control the branching structure of the tree and reduce model complexity. This max_features parameter limits the max number of features that will be looked at during each split, lowering the variance and helping generalizing better. The criteria parameter determines a splitting quality value based on Gini coefficient or information gain and thus determines a splitting strategy of the model. Finally, in another step, the experiment decided on optimal parameters for pruning the decision tree model via post pruning with different ccp-alpha parameters; with different values to AUC values generated with a changing value of the ccp-alpha. By adjusting this parameter, the study remove unnecessary branches and make the model simpler and more generalization.

In the parameter list of the model, the n_estimators parameter was set to the number of decision trees in this experiment to provide information for random forests. The study usually found that the more trees in the model, the more stable and accurate the prediction is. To keep the model or their predictions from over fitting, max_depth was set to control the maximum depth of each tree. Min_stamples_split and min_stamples_leaf for the list specify minimum number of samples for the internal

nodes and leaf nodes respectively and are set to reduce model complexity and avoid overfitting. By setting the `max_features` parameter, the study performs feature selection by limiting the number of features considered for each split further improving generalization efficiency by increasing model diversity. The `criteria` parameter in the end of the list was set to select the evaluation criteria for node splitting; in this case, Gini index and information gain influenced the decision-making strategy of the model. To make random forest robust and effective, study adjusted these parameters.

In the experiment, the study calls several different machine learning model instances in the scikit learn for training and evaluating each model. When the experiment is run in Python, the study will be using the `fit()` function from the scikit learn library for fitting models, the core of machine learning model training function itself. The `fit()` function is either an important part of fitting the model to the training data if the model is a supervised learning model (such as classification and regression), or an unsupervised learning model (such as clustering and dimensionality reduction). In this case, the study have dataset features in matrix `X` and dataset result labels in array `y` that can be used to train different machine learning methods by providing these parameter pairs.

To run the experiment of comparing the training and prediction time for each model, the study used the `time()` function from the Python standard library's `time()` module. To start timing before training a model, `time()` was used. Subsequently, `time()` will be used to stop the time when the subsequent model ends its prediction. The time spent by the model during this period can be obtained by subtracting the two times. For the use of models for prediction, the experiment will use the `predict()` function in the Python library as the instruction for the model to predict the test set results. The `predict()` function is a core method used for machine learning models to make predictions. After training the model, the `predict()` function will accept input data and output prediction results. The experiment will calculate the relevant evaluation indicators for the model based on this result.

In this experimental project, the research first calculated the AUC value. First, the `predict_proba` function in the sk-learn library was called to calculate and extract the probability that each sample in the test set was predicted to have heart disease (category 1). Then, the `roc_auc_score()` function was called to calculate the AUC value with the actual test set results and the previously measured probability as parameters. AUC is

calculated based on the area under the ROC curve, which is plotted with FPR as the horizontal axis and TPR as the vertical axis.

Among them, TPR is the proportion of correctly predicted positive samples among all positive samples.

$$TPR = \frac{TP}{(TP+FN)} \quad (3.3)$$

FPR, which is the proportion of negative samples that are incorrectly predicted as positive.

$$FPR = \frac{FP}{(FP+TN)} \quad (3.4)$$

As the threshold changes from high to low, TPR and FPR will change, and the ROC curve will be drawn. Specifically, the trapezoidal rule will be used to calculate the area under the ROC curve. The specific formula is:

$$AUC = \sum_{i=1}^{n-1} (x_{i+1} - x_i) \times \frac{y_{i+1} + y_i}{2} \quad (3.5)$$

Among such points, one of them is x_{i+1} and another is x_i that are two adjacent points of the FPR. Moreover, y_{i+1} and y_i are two adjacent points of corresponding TPR.

Then, the experiment can use the ROC curve obtained by each model for different conditions as parameters of the `plt.plot()` method, taking the FPR and TPR as parameters.

Ability evaluation of machine learning methods is a key step for which the experiment obviously needs to build a confusion matrix. It can improve experiments being able to better understand how well the model does in classification tasks specifically the efficiency of correctly and incorrectly classifying. The project in the experiment has used parameters as `y_test` classification label of the actual test set and `y_pred` classification label generated by model prediction. The confusion matrix of the model was built with the help of `confusion_matrix()` function. Using the provided data from the confusion matrix, the experiment calculated the accuracy, precision, F1 score, recall, and other key metrics from corresponding formulas according to the conditions of each model.

3.6 SUMMARY

The Framingham Heart Disease dataset from Kaggle containing 4240 records and 15 feature attributes is the dataset which is selected and processed for this project. Through these data processing steps, the experiment had created a solid and reliable data

foundation that will be built and evaluated using machine learning methods.

The data processing stage covers multiple important steps, and introduces the implementation method of the experimental process in Python, including sorting out the basic information of the dataset, attribute classification, statistical information analysis, visualization processing, missing value processing, random resampling, chi square test, feature selection, standardization of the dataset, different proportion division of the training and testing sets, and 10-fold cross validation. When dealing with missing values, the mean substitution method was used to reduce the impact of data loss. Random search was used to select the optimal machine learning model parameters, chi square test was applied to determine the key features that can affect the occurrence of heart disease in the dataset, standardization was used to make the data conform to a standard normal distribution, and different proportions of dataset partitioning and 10-fold cross validation were adopted to ensure the comprehensiveness and accuracy of the model evaluation.

In this section, the dataset features are screened and selected using Python and predictive analysis is performed on various machine learning types and on differing partitions of the dataset to generate ability evaluation indicators for each model and find whether the samples will likely have heart disease in 10 years.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 INTRODUCTION

In the project of this section, Python is utilized to screen and select the features of the experimental dataset and perform predictive analysis on various machine learning methods under various dataset partitions to obtain ability evaluation indicators for each model in order to decide if the samples would have heart disease in the next 10 years. This is done to determine what variables are the biggest influencers of heart disease risk in the given experimental dataset and screen the machine learning methods with the best efficiency in the heart disease dataset.

4.2 RESEARCH RESULTS

4.2.1 THE MOST INFLUENTIAL FEATURES ON THE FINAL RESULT

After passing the chi square test and visualizing the final result, the result is shown in Figure 4.1:

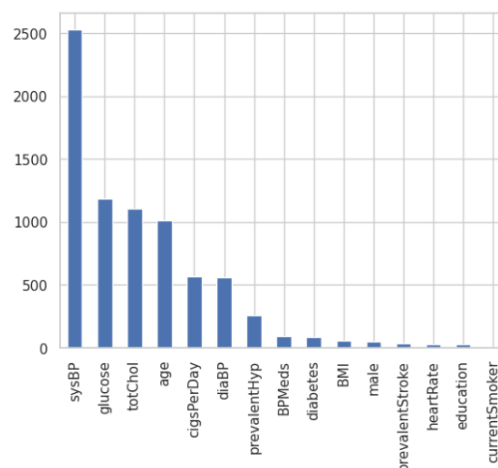


Figure 4.1 The chi square values of different features in the dataset

Through the chi square test, it is possible to visually observe the relationship between various factors in the dataset and the risk of developing heart disease within 10 years. From the results, it can be seen that the features of the dataset such as "sysBP", "glucose", "age", "cigsPerDay", "totChol", "diaBP", and "PrevalentHyp" are closely related to the disease risk of the final sample within 10 years.

In order to more intuitively demonstrate how various factors affect the final risk of disease, the experiment visualized the distribution histograms of the above features in the sample with and without disease risk.

Firstly, as shown in Figure 4.2, from the distribution map of systolic blood pressure (sysBP), individuals with heart disease (orange line) have a higher frequency in the range of higher systolic blood pressure (approximately above 140), while individuals without heart disease (blue line) are concentrated in the lower range of systolic blood pressure. As systolic blood pressure increases, the risk of heart disease in the sample also increases.

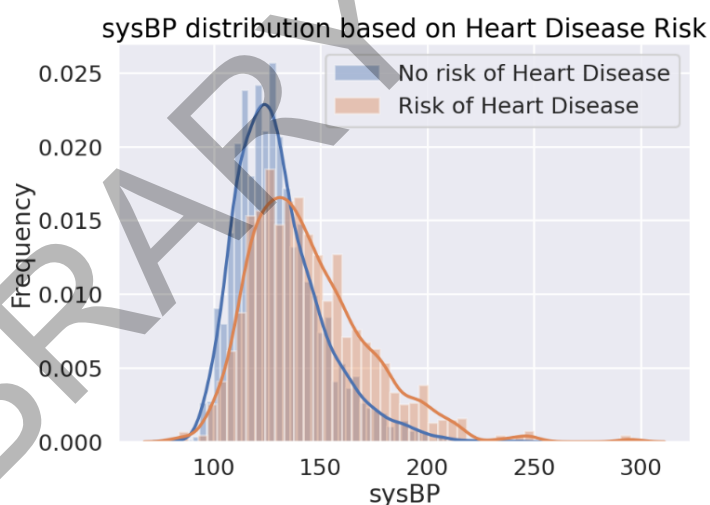


Figure 4.2 SysBP distribution based on Heart Disease Risk

Similarly, in the glucose distribution map shown in Figure 4.3, individuals at risk of developing heart disease have a slightly higher frequency within the low blood sugar range, which may indicate that in this dataset, individuals with low blood sugar may be more susceptible to heart disease.

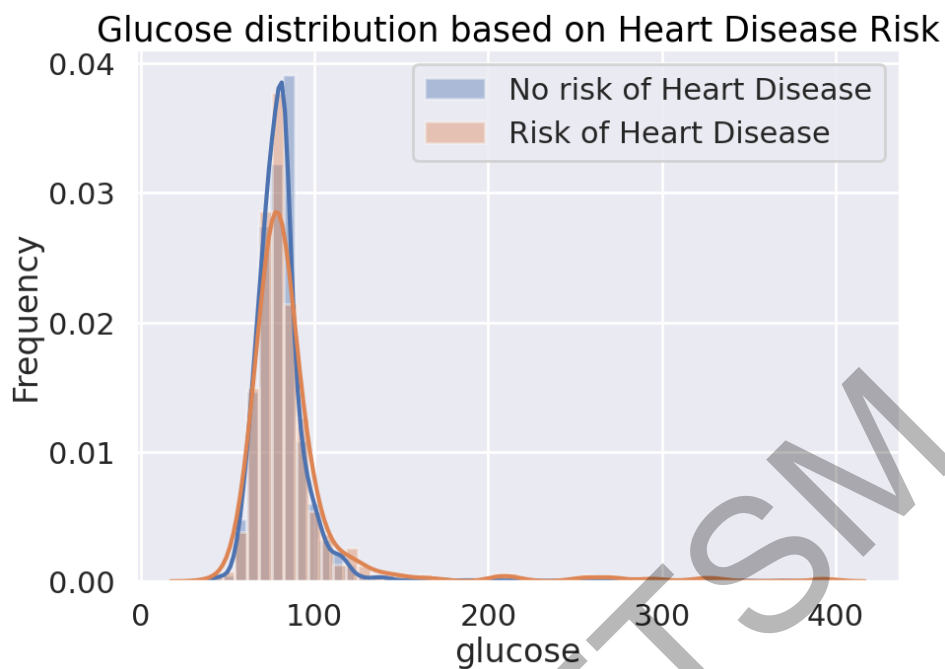


Figure 4.3 Glucose distribution based on Heart Disease Risk

As shown in Figure 4.4, the distribution map of total cholesterol (totChol) shows that individuals with heart disease have a higher frequency in the range of higher total cholesterol, especially during the period of 260 and above, while individuals without heart disease are concentrated in the lower cholesterol range, and this pattern is particularly prominent at 200 and below. Therefore, it can be seen that the risk of heart disease in the sample increases with the increase of total cholesterol (totChol) within 10 years.

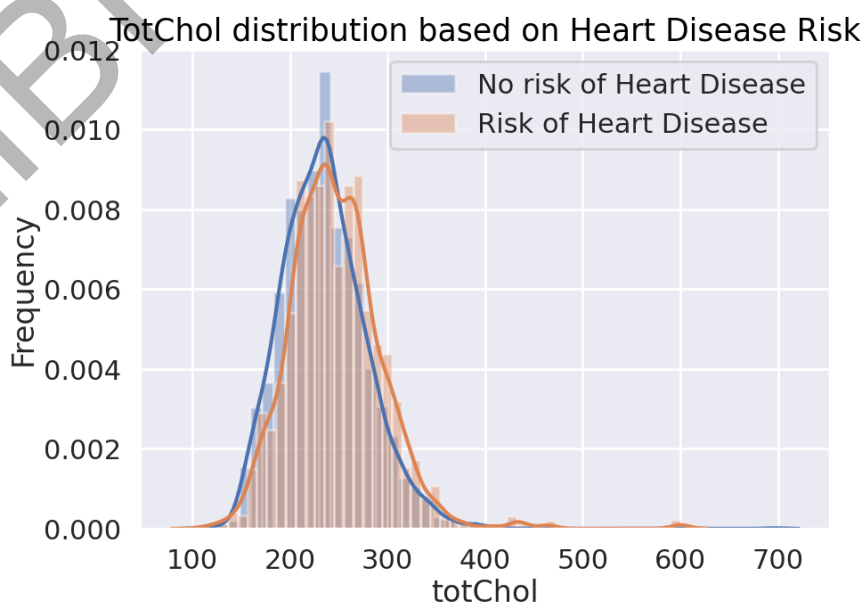


Figure 4.4 TotChol distribution based on Heart Disease Risk

As shown in Figure 4.5, by analyzing the distribution map of age characteristics, experiments can gain a more detailed understanding of the impact of age on heart disease. From the graph, it can be seen that individuals without heart disease are mainly concentrated between the ages of 30 and 50, and the frequency of no heart disease risk reaches its peak between the ages of 35 and 50. In contrast, individuals with heart disease are mainly concentrated between the ages of 50 and 70, with a peak frequency around 55-60 years old. From this, it can be seen that the risk of heart disease in the sample significantly increases after the age of 50.

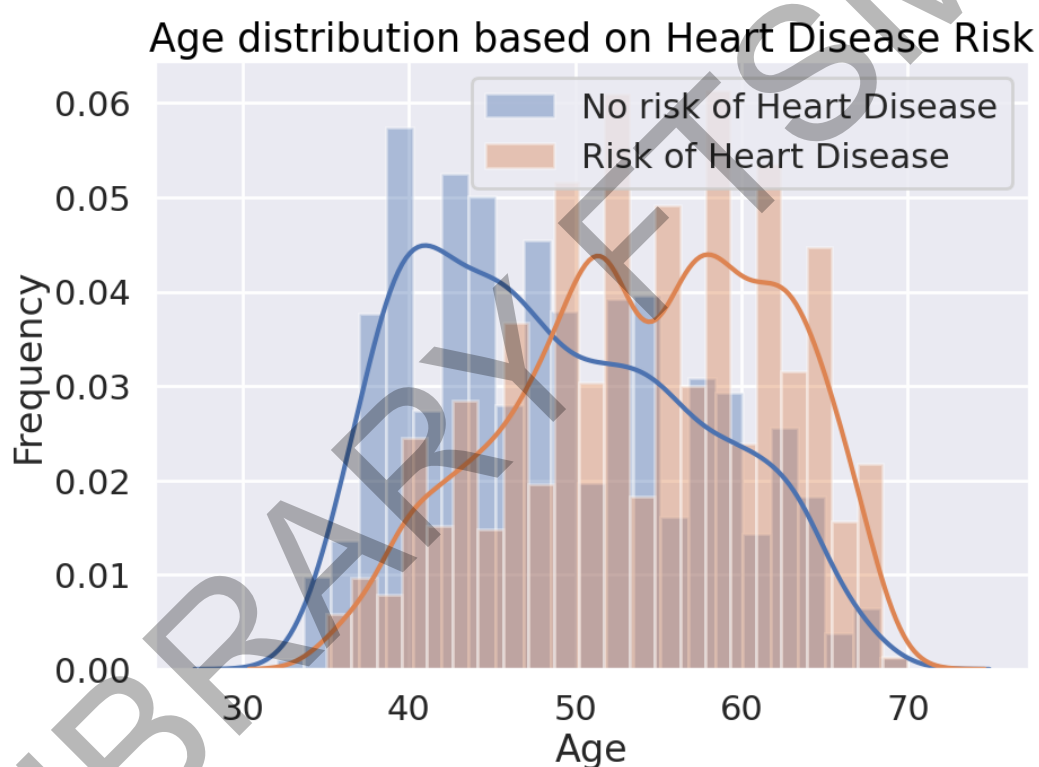


Figure 4.5 Age distribution based on Heart Disease Risk

The distribution of daily smoking volume (CigsPerDay) between individuals at risk of heart disease and those without heart disease shown in Figure 4.6 indicates that the proportion of individuals at risk of heart disease is higher among those who smoke more per day. After the number exceeds 20, the proportion of samples with heart disease risk exceeds that of samples without heart disease. Therefore, the above results indicate a correlation between smoking volume and the risk of heart disease. People who smoke heavily are more likely to suffer from heart disease, and controlling or reducing smoking is of great significance for preventing heart disease.

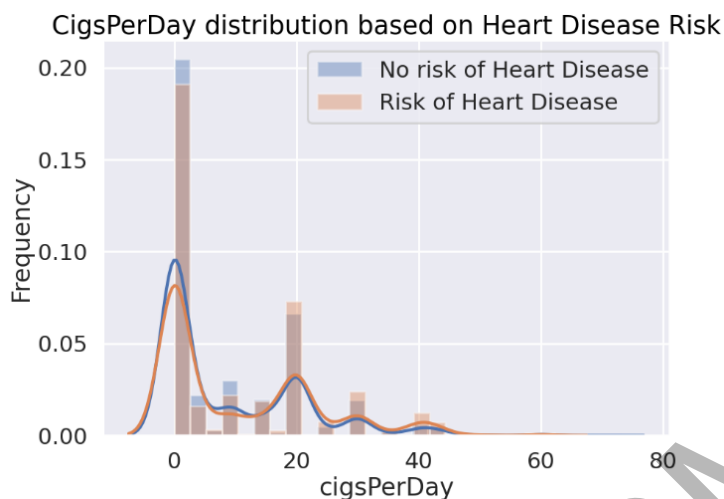


Figure 4.6 CigsPerDay distribution based on Heart Disease Risk

As shown in Figure 4.7, the distribution of diastolic blood pressure (DiaBP) between individuals at risk of heart disease and individuals without heart disease indicates that individuals with higher diastolic blood pressure have a higher proportion of individuals at risk of heart disease. Especially when the diastolic blood pressure starts around 90, as the diastolic blood pressure continues to rise, the proportion of people at risk of heart disease is significantly higher than that of people without heart disease risk. When the diastolic blood pressure is below 80, the proportion of people without heart disease risk is much higher than that of people at risk. Therefore, it can be concluded that controlling diastolic blood pressure within the normal range is of great significance for preventing heart disease. The diastolic blood pressure starts at around 90, and as the diastolic blood pressure increases, the risk of disease in the sample also increases.

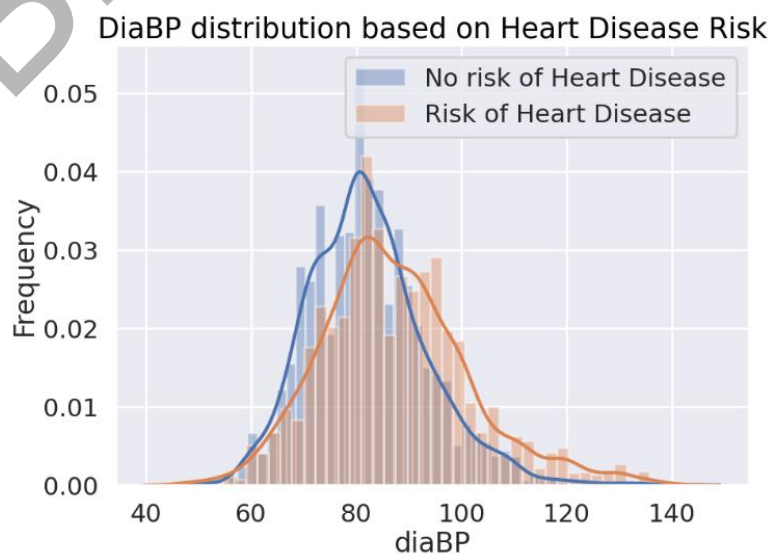


Figure 4.7 DiaBP distribution based on Heart Disease Risk

The visualized bar chart shown in Figure 4.8 displays the distribution of whether the sample has a history of hypertension in the population at risk of heart disease and the population without heart disease. From the chart, it can be seen that people with a history of hypertension have a higher risk of developing heart disease, while those without a history of hypertension have a lower risk of developing heart disease. This indicates that people affected by hypertension have a higher risk of developing heart disease, suggesting that hypertension is one of the important risk factors for heart disease.

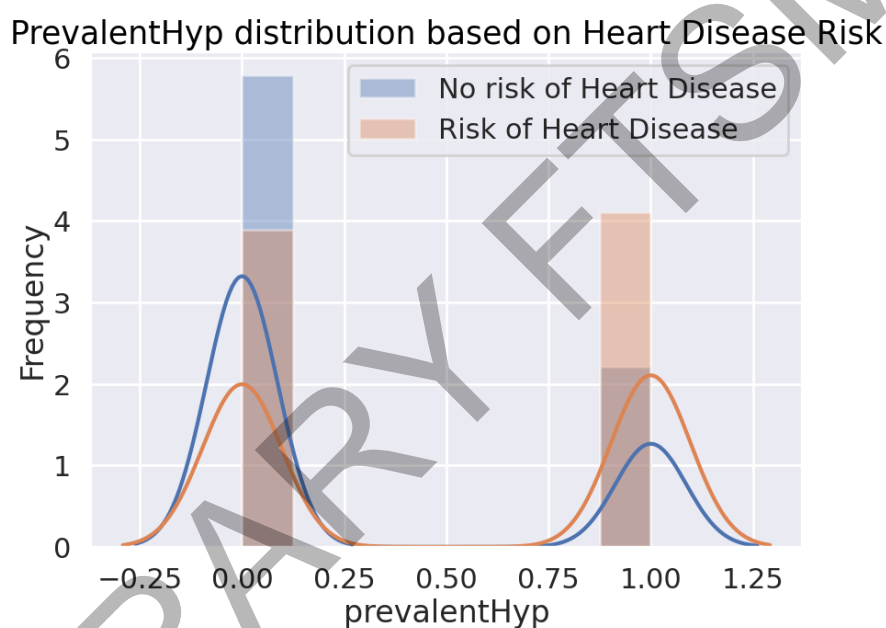


Figure 4.8 PrevalentHyp distribution based on Heart Disease Risk

4.2.2 PARAMETER SELECTION FOR MACHINE LEARNING MODELS

In the experiment, each machine learning type selected the optimal model parameters through random search under different dataset partitioning ratios to ensure that the model can achieve optimal efficiency in specific situations.

In the testing of the logistic regression model, the results of random search showed that the optimal parameters of the model changed under various configurations such as training set proportion from 90% to 10%, testing set proportion from 10% to 90%, and 10-fold cross validation. In terms of solver parameters, the majority of configurations use the 'saga' solver for the model. However, when the training set accounts for 90% and the testing set accounts for 10%, and the training set accounts for