Problem Chosen
A

2024
MCM/ICM
Summary Sheet

Team Control Number
JS06012024193

Prediction of Cardiovascular Disease Based on Machine Learning

Cardiovascular disease is a general term for heart and vascular disease, having become the first attribute to all risk factors. The rise in the number of individuals affected by cardiovascular diseases has made the diagnosis and prediction of these conditions a significant concern within the medical field. Machine learning method has a wide range of applications in the field of medical big data, so in this paper **machine learning method** is adopted to predict cardiovascular disease.

In Task 1: To preprocess and conduct exploratory analysis on the data, several challenges and characteristics specific to medical diagnostic data are identified, including **redundancy** in text, **unreliable data** due to record errors, **incomplete data** from imbalanced proportions and unclear problem descriptions. The data is obtained from Kaggle platform and focuses on patients with CVD. The data preprocessing involves steps which are **data cleaning, transformation and reduction**. Exploratory data analysis reveals correlations between variables, which indicates **ap_hi, ap_ho, age and cholesterol** have **stronger correlations** with cardiovascular disease.

In Task 2: To predict whether a patient suffers from cardiovascular disease from the patient's indicators, we establish six classification prediction models based on machine learning method, which includes three **single models(Logistic, BP Neural Network, SVM)** and three **integrated models(Random Forest, XGBoost, LightGBM)**. The model parameters are set in advance and the optimal parameters are determined after using the training set to obtain the prediction model. Also, we visualize all of the training results.

In Task 3: To compare the prediction performance of each model, this paper calculates the accuracy, $F1.score$ and AUC of each model. By comparing the various indicators of models, it is found that the prediction performance of **integrated models** is **better** than that of **single models**, in which **Logistic**, as a linear model, has the **worst** prediction accuracy(66.6%) and the **LightGBM** model has the **best** accuracy(80.2%) in various indicators.

At the end of the paper, a conclusion is made that the use of machine learning techniques has gained significant attention in CVD prediction but still exists drawbacks. This paper successfully processes and analyzes the data, identifying key factors that play a major role in cardiovascular diseases. The nonlinear model, specifically the LightGBM model with optimized parameters, demonstrates the best predictive performance, while it also has limitations, including the need for extensive training data and large data to support.

Key Words: Cardiovascular; Machine Learning; Correlation Analysis; Single and Integrated Models

Contents

1 Introduction

1.1 Background and Problem Statement

Cardiovascular disease is a general term for heart and vascular diseases, which includes hypertension, coronary heart disease, cerebrovascular disease, peripheral vascular disease, heart failure, rheumatic heart disease, congenital heart disease and cardiomyopathy. According to the World Health Organization, approximately 17.5 million people died from cardiovascular disease in 2012, accounting for 31% of global deaths. Among these deaths, an estimated 7.4 million died from coronary heart disease and 6.7 million died from stroke. Due to the increasing number of patients with cardiovascular diseases, the diagnosis and treatment of cardiovascular diseases have become a major issue in the medical industry.

So we urgently need to address the challenges of prevention and treatment of cardiovascular diseases. Machine learning and data mining algorithms have been effectively applied in numerous diseases. With medical big data and algorithms such as data mining at its core, experts in the field state that building data models for cardiovascular disease prediction can solve the challenges of cardiovascular diseases and hold significant significance and value in clinical practice (Guo, Li & Luo, 2013)[1]. This technology not only helps patients with early prevention but also provides doctors with effective treatment plans, helping patients seek medical attention in a timely manner, reducing patients' risk of illness and mortality.

Here, we need to accomplish the following objectives:

- Taking advantage of the given data, we should preprocess and conduct exploratory analysis on it.
- We should use the classification method in machine learning to predict whether the patient has cardiovascular disease from the patient's physiological indicators, medical detection indicators and subjective information.
- We also need compare the prediction performance under different classifiers and draw a conclusion at last.

1.2 Related Work

The diagnosis of cardiovascular diseases is often inferred through various physiological indicators of the body. In the early 21st century, medical experts in the United States gradually conducted detailed research on the risk factors of cardiovascular diseases and developed predictive models for cardiovascular diseases[2]. The influential factors in the models include basic patient information such as age, gender, height, weight, as well as laboratory and examination indicators. In 1998, researchers transformed continuous variables into categorical variables, established relationships, and developed risk prediction models[3]. Through experiments, it was ultimately determined that the main risk factors for coronary heart disease in middle-aged and elderly people are cholesterol, creatinine, blood pressure and blood glucose levels. In 1967, research institutions in the United States used logistic regression to establish predictive models for the incidence of coronary heart disease and subsequently introduced various new disease prediction models[4].

In the article "Research on the Prediction Model of Cardiovascular Health Status Level Based on Data Mining," Lu Yang identified risk factors for cardiovascular diseases, including systolic blood pressure, diastolic blood pressure, and fasting blood glucose as eight vital signs[5]. Geng Zhongze, in the paper "Research on the Auxiliary Diagnosis Function of the Inspection Information System Based on Decision Tree," used the ID3 algorithm to generate a decision tree model for disease prediction in cardiovascular disease. The paper mentioned the combination of fuzzy clustering and stepwise recognition analysis[6]. In 2014, Chen Qingyun and colleagues used logistic regression analysis to analyze the relationship between cardiovascular diseases and factors such as blood pressure, cholesterol and blood glucose levels[7]. Pang Xiantao from Jilin University applied BP neural networks to the prediction of heart disease in his master's thesis and achieved good predictive results[8].

1.3 Our Work

The work we have done in this problem is mainly shown in the following Figure 1.

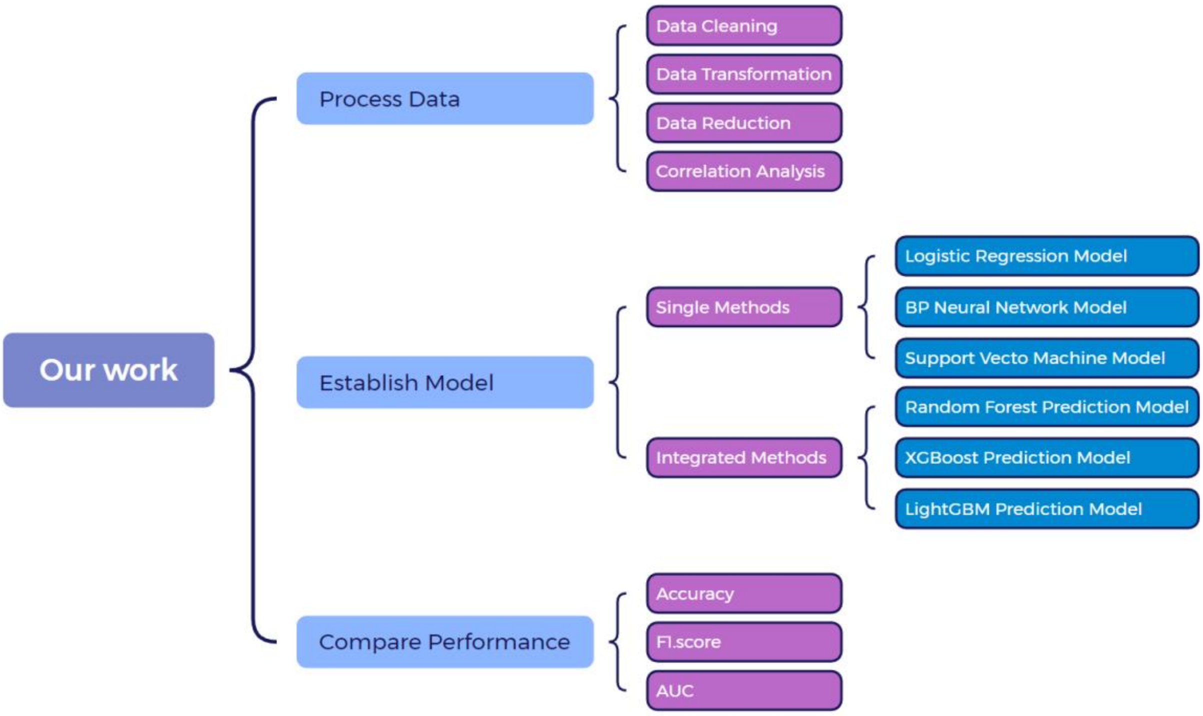

Figure 1: Our Work

2 Preparation of the Models

2.1 Assumptions

We made the following assumptions to help us with our modeling. These assumptions are the promise of our subsequent analysis.

- Assuming that the data is obtained through surveys, which means our analysis accords with the truth.
- Assuming that the influencing factors of cardiovascular disease are only eleven elements given in the dataset to make our predictions more accurate.
- Assuming that errors may occur in the given data but there exists no lag, which means that the data provided by the patient is real-time.

2.2 Notations

The primary notations used in this paper are listed in Table 1. Variables not mentioned in Table 1 will be explained in detail when they are used in the following section.

Table 1: Parameter Settings

Parameter	Description
Q	The Degree of Similarity Between Two Variable
n	Sample Size
a	The Parameter of Machine Learning
λ	The Parameter of Machine Learning
ρ_{xy}	Correlation Coefficient Between Two Variables
E_0	Effect of Different Units of Measurement
x_i	A Variable With the Number of Indicators i
a_i	The Coefficient Before the i-th Variable
p	Sigmoid Function
C	Inverse of Regularization Coefficient
lr	Learning Rate
d	Distance Between Hyperplanes
K	Kernel Function
l	Loss Function
Ω	Regular Term
y_i	True Value Corresponding to the Outcome
$\widehat{y}_i$	Predicted Value of the Model
k	Number of Trees in a Random Forest
f_i	Nonlinear Function

3 Task 1: Preprocess and Conduct Exploratory Analysis on the Data

3.1 Data Quality Analysis

Because of inconsistent data sources and different acquisition methods, there are differences between the experimental results and the expected results. The running process also encounters various problems, leading to inaccurate results. In this paper, medical diagnostic data samples are used, which have unique characteristics that pose various challenges during the experimental process[9]. These challenges include redundancy in the text, unreliable data due to record error, incomplete data due to patients' oral accounts leading to missing information, imbalanced proportions, and unclear problem descriptions.

- Duplication: Medical data differs from other scientific research data in that a large amount of data is generated every day. Doctors create medical records and examination information based on patient identities, but since the patients may go to different hosptals for check and treatment, repetive information may be included.

- Incompleteness: Medical data originates from different sources, including the doctor's workstation and patients' self-reports. This can result in data biases, incompleteness, and inconsistencies due to the urgency of treatment procedures.

- Temporality: Typically, there is a time gap between a patient's onset of illness and body check. Patients seek medical attention after falling ill, and doctors require some time to aquire physiological data.

- Imbalance: Due to variations in the affected population and different standard to determine severity of the disease, data collection can lead to imbalanced data in terms of demographic characteristics and disease severity.

In this paper, the feature parameters of the training and testing samples are obtained by analyzing the medical records of cardiovascular disease patients through research. The training samples are used to build a cardiovascular disease prediction model. The classification algorithm is continuously modified by adjusting the weights and thresholds of the training function to optimize the structure. Finally, the performance of the training results is analyzed and determined, the best model is selected, the training parameters are saved, and the testing samples are used to evaluate and analyze the predictive performance of the model.

3.2 Data Sources

Data of this article is obtained from cardio_train project on kaggle. The original dataset contains 70000 records. Firstly, the focus of the study is primarily on patients with cardiovascular diseases, which can be analyzed by their vital sign information. Secondly, we aim to extract the medical examination information of patients, including basic personal information such as the patient's id, gender, age, weight, height, subjective information such as exercise situation, smoke and alcohol consumption, as well as examination results such as blood pressure test, cholesterol test, and blood glucose test. These data contain a large amount of sign values that are not related to cardiovascular diseases. From the medical examination data, the selected data sources mainly include age, gender, smoking history, drinking history, exercise situation, BMI, diastolic pressure, systolic pressure, cholesterol, glucose, totaling 10 variables.

3.3 Data Preprocessing

Data processing plays a crucial role in the entire process of data mining. Based on the characteristics of big data, this study utilizes the Python platform and Pandas library for data interaction and querying. And the final data is output in a document with the suffix xlsx. The focus of this paper's data processing is primarily on the prevention and prediction of cardiovascular diseases. With the guidance of medical experts and through mining medical information, accurate and reliable information is discovered while discarding irrelevant fields. This reduces the workload of data processing in the experimental process and consequently improves the efficiency of data mining, which ensures that the experimental results are more persuasive and reliable. The data processing process is illustrated in Figure 2.

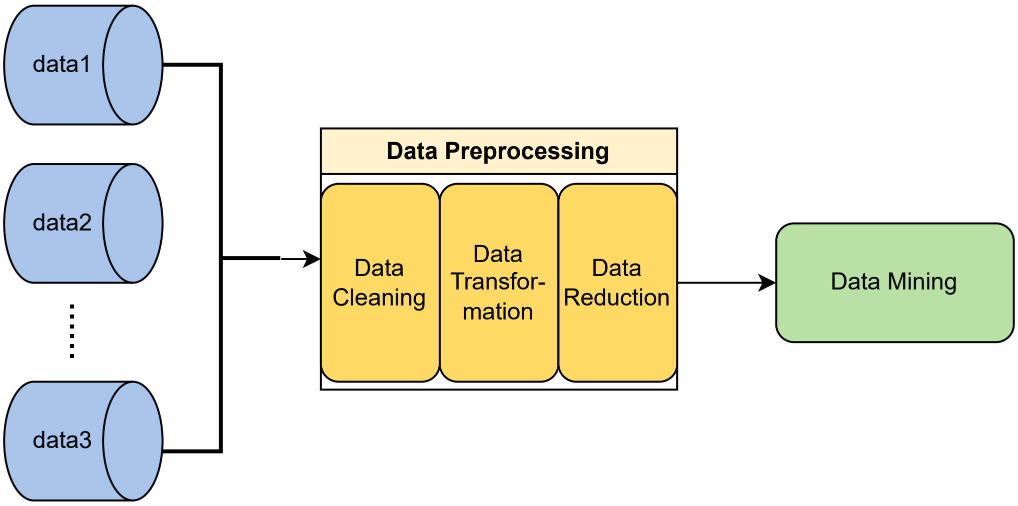

Figure 2: The Flow Diagram of Data Processing

3.3.1 Data Cleaning

Data cleaning is the first step in data processing, which is aimed at filtering out duplicate records and ensuring the validity of the remaining records. It involves filling in missing values, correcting erroneous information and determining the validity of the data for the experiment. The principles of data cleaning are illustrated in Figure 3.

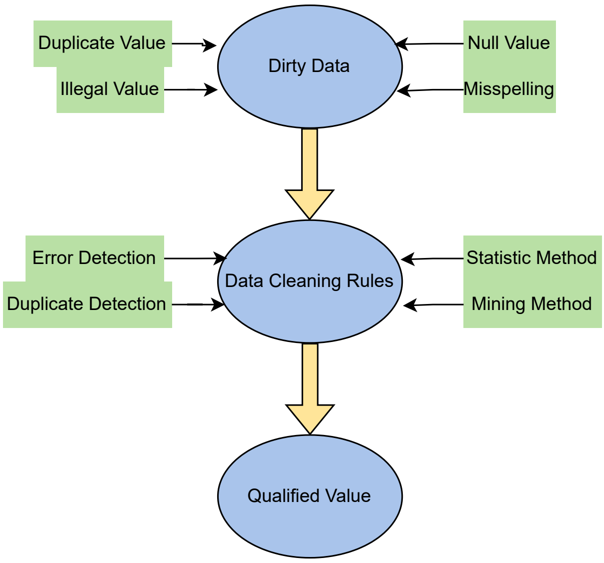

Figure 3: The Principles of Data Processing

We use tuple deletion and regression methods for data cleaning, which means deleting objects with missing attribute values to obtain a complete dataset and utilizing regression models to handle relatively complete data.

In this dataset, by using python command as follows, we can find that there are no missing values.

```
for colum in df_train.columns:
    if colunm in imp_list:
       current_label0_mean = df train[df_train[’sign’]==0][column].mean()
       current_label1_mean = df_train[df_train[’sign’]==1][colunn].mean()
for i in range(len(df_train)):
    if pd.isna(df_train.loc[i,colum]):
       if df_train.loc[i,’sign’]==0:
          df_train.loc[i,column]=current_label0_mean
       else:
          df_train.loc[i,column]=current_label1_mean
```

In this paper, the hospital measurement standards and threshold ranges are followed to facilitate data statistics and analysis. First, each data item is assigned a unique identifier code for monitoring purposes. This can be implemented using a Pandas library. The cleaning and storage of data are performed in the Jupyter Notebook, and the following method is used for handling various exceptional cases.

Due to the nature of disease data, outliers and writing errors often occur. It is important to define appropriate ranges for each attribute field and perform threshold-based queries. Data outside of these ranges can be considered as outliers. There are two main types of outliers: outliers with identifiable causes and outliers with unknown causes. By defining appropriate ranges, performing data comparisons, and identifying inconsistencies, outliers and errors can be detected and corrected in disease data. For example, if the systolic blood pressure is recorded as 280, this data should be deleted since reading of higher than 180 is considered "hypertensive crisis".

3.3.2 Data Transformation

Data transformation refers to converting data into a format suitable for data mining. It involves the following aspects:

- Smoothing: Removing noise from the data to make it smoother, which can be achieved by applying techniques such as filtering, averaging, or interpolation to eliminate outliers or irregularities in the data.
- Aggregation: Grouping or combining data to create aggregated representations. This can involve summarizing data at higher levels of granularity[10].
- Generalization: Transforming data by replacing detailed or raw values with higher-level concepts or hierarchies, which can involve generalizing specific attributes or values to more abstract or broader categories.
- Normalization: Standardizing data to a common scale or range. This ensures that different attributes with varying measurement units or scales are transformed to a comparable and consistent format.

In this paper, generalization is mainly used. According to British Heart Foundation, being Overweight is a significant cause of the CVD. Overweight analysis is conducted by generalize height and weight to body mass index.

$$BMI = weight(kg)/(height(m) * height(m)) \tag{1}$$

3.3.3 Data Reduction

When dealing with experimental data samples, which often contain a large amount of data, medical data in particular is inherently complex and has numerous features. Therefore, data reduction techniques are commonly applied to medical data. However, it is crucial to maintain the intrinsic characteristics of the data while ensuring that the final representation is intuitive and easy to comprehend.

We use the methods of data cube aggregation and numerical reduction to dispose of the problem, which stand for allowing multidimensional modeling and visualization of the data.

3.4 Exploratory Data Analysis

3.4.1 Definition of Correlation Coefficient

We use the Pearson correlation coefficient, which represents the degree of linear correlation between the study variables. This correlation coefficient is calculated by the product differentiation method based on the difference between these two variables and their mean values[11]. In this paper, we study the correlation between two variables. Let Q_0 be two vectors, and the data be taken from the overall sample. Then, the correlation between the vectors is measured by calculating the minimum value of the sum of squared errors, which is shown as the equation(2).

$$Q(a,\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - \lambda x_i)^2 \tag{2}$$

The Q_0 used to describe the degree of similarity between two variables. Derive the a and λ and make it equal to 0. As shown in equation(3).

$$\begin{cases} \frac{\partial Q}{\partial a} = -\frac{2}{n}\sum\limits_{i=1}^{n}(y_i - a - \lambda x_i) = 0 \\ \frac{\partial Q}{\partial a} = -\frac{2}{n}\sum\limits_{i=1}^{n}[(y_i - a)x_i - \lambda x_i^2] = 0 \end{cases} \tag{3}$$

Then, we can get the solution for equation(3).

$$\begin{cases} \lambda = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2} \\ a = \overline{y} - \lambda\overline{x} \end{cases} \tag{4}$$

We substitute equation(4) into equation(2). Then, we can get the equation(5).

$$\begin{cases} Q_0 = \frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2(1-\rho_{xy}^2) \\ \rho_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}} \end{cases} \tag{5}$$

The minimum relative error sum of squares can also be obtained as follows.

$$E_0 = \frac{Q_0}{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2} \tag{6}$$

We define ρ_{xy} as the correlation coefficient between two variables. Also, we can use $E_0 = 1 - \rho_{xy}^2$ to describe the degree of similarity between the variables and E_0 eliminates the effect of different units of measurement of the variables compared to Q_0. The larger ρ_{xy} is, the more correlated x and y are. Similarly, the samller ρ_{xy} is, the less correlated x and y are.

3.4.2 Results of Correlation Analysis

The data source of age, height, weight, gender, ap_hi, ap_ho, cholesterol, gluc, smoke, alco, active, totaling of 11 variables, which suffer from cardiovascular disease correlation analysis, is listed as follows. This paper uses excel table CORREL function for the calculation of correlation coefficient. The correlation coefficient ranges from -1 to 1. When the value is close to -1, it means inverse correlation, similar to the inverse proportional function. Similarly, when the value is close to 1, it means positive correlation. The results of the correlation analysis are shown in Table 2.

Table 2: The Results of the Correlation Analysis

Factor	Correlation Coefficient	Absolute Value of Correlation Coefficient
Age	0.240	0.240
Height	0.005	0.005
Weight	-0.011	0.011
Gender	0.176	0.176
Ap_hi	0.433	0.433
Ap_ho	0.336	0.336
cholesterol	0.221	0.221
Gluc	0.088	0.088
Smoke	-0.019	0.019
Alco	-0.011	0.011
Active	-0.040	0.040

Analysing the above correlation coefficients, it is concluded that the correlation coefficients of ap_hi, ap_ho, age and cholesterol have relatively higher values. The correlation coefficients of height, weight, alco and active have relatively lower values. Therefore, we consider ap_hi, ap_ho, age and cholesterol to be a strongly correlated variable, gender, gluc and smoke to be a moderately correlated variable, and height, weight, alco and active to be a weakly correlated variable. From the analysis above, based on the given data, we believe that ap_hi, ap_ho, age and cholesterol play a decisive role in whether or not people have cardiovascular disease.

4 Task 2: Maching Learning Methods for Classification

Various machine learning methods have been employed to establish cardiovascular risk prediction model. Grid search and cross-validation techniques are utilized to optimize the model parameters, aiming to improve prediction accuracy and stability. In this paper, We adopt a total of six machine learning methods, including single model and integrated model, to predict cardiovascular disease in hemodialysis patients.

4.1 Cardiovascular Disease Prediction Using Single Models

Before establishing a machine learning model, it is necessary to pre-set certain parameters that can affect the model's construction and its predictive performance. This is done to find the optimal parameters that can make the model achieve the best results. A commonly used models is K-fold cross-validation combined with grid search. The grid search method originates from combinatorial optimization techniques, and its core idea is to enumerate all possible solutions in each parameter space, evaluate each solution using a certain criterion, and search for the optimal solution. Cross-validation is an effective approach to address issues like small dataset size and parameter optimization. It includes methods such as simple cross-validation, K-fold cross-validation, and bootstrapping, among which K-fold cross-validation is a commonly used method for model evaluation. The machine learning predictive model is constructed following the steps outlined in Figure 4.

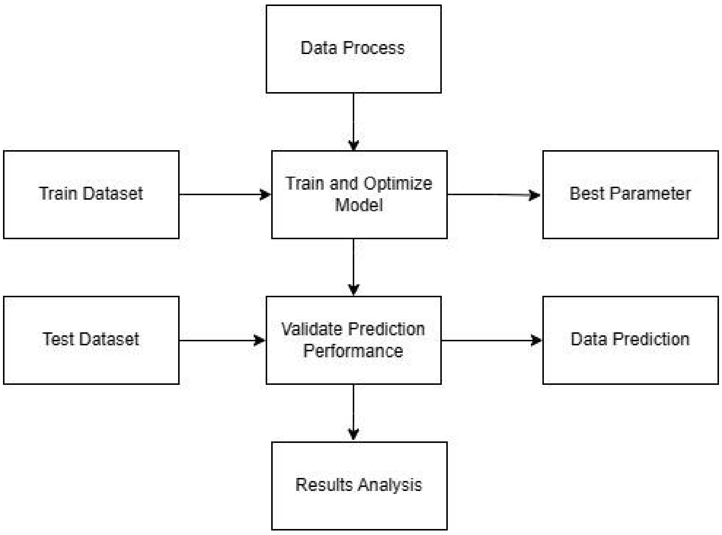

Figure 4: The Processing of Building Machine Learning Models

4.1.1 Logistic Regression Model

Logistic regression is a generalized linear regression model used to analyze and predict discrete variables based on either categorical or continuous independent variables. It is a common statistical method for handling qualitative variables. The mathematical representation of the narrow linear model can be expressed as equation(7).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{7}$$

Since the result of equation(7) is a specific value, and disease prediction requires classification, the results of equation(7) need to be transformed into corresponding probabilities. This requires the incorporation of the sigmoid function in the logistic regression algorithm. The sigmoid function introduces non-linearity and limits the output to the range $(0, 1)$, allowing for classification. The resulting probability formula after applying the sigmoid function is the equation(8).

$$p(X) = \frac{e^{a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n}}{1 + e^{a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n}} \tag{8}$$

By manipulating equation(8) through mathematical operations, we obtain the following equation(9).

$$a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n = \ln \frac{p(X)}{1 - p(X)} \tag{9}$$

The left side of equation(9) has a similar format to equation(7) and represents the odds ratio, which indicates the likelihood of the outcome occurring. As the value of the linear function tends toward positive infinity, the probability output of the model approaches 1, but does not exceed the upper limit. Conversely, as the value of the linear function tends toward negative infinity, the probability output of the model approaches 0.

When establishing a logistic regression model, it is important to choose the penalty term and optimization algorithm appropriately based on the specific circumstances, as these parameters determine the structure of the logistic model. Additionally, to avoid overfitting, an appropriate regularization coefficient should be selected. Table 3 shows the optimal parameters obtained through grid search and 5-fold cross-validation.

Table 3: Optimal Parameters of Logistic Regression Model

Model Parameter	Meaning	Value
Penalty	Regularization Penalty	$L2$
C	Inverse of Regularization Coefficient	10
Solver	Optimization Algorithm	liblinear

The results of logistic regression on the training set are as follows. Accuracy is 68.2%, $F1$ score is 65.5%, and AUC is 0.704. Testing the logistic model with the optimal parameters on the testing set yields the following results. Accuracy is 66.6%, F1 score is 61.6%, and AUC is 0.678. Figure 5 shows the ROC curve of the logistic model. It can be observed from the graph that the evaluation metrics of the logistic regression model on both the training and testing sets are relatively low, indicating limitations of the linear model in identifying complex relationships between variables.

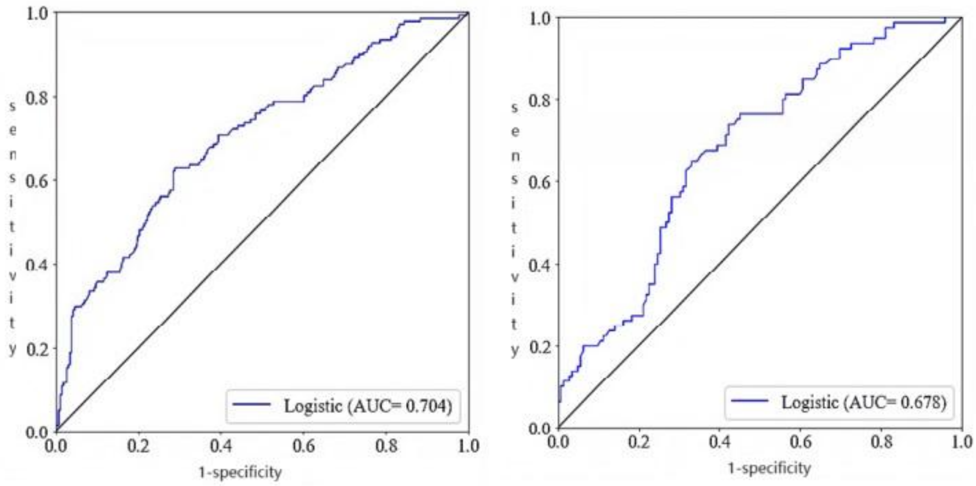

Figure 5: ROC Curve of the Logistic Model

4.1.2 BP Neural Network Model

The BP neural network algorithm possesses excellent nonlinear mapping capabilities and is commonly used to solve classification and prediction problems. The algorithm was conceptualized by Rumelhart and his colleagues in 1986 and functions through a combination of model training and error correction. It analyzes the error between each output and the expected result, continuously adjusts weights and thresholds through negative feedback, and ultimately obtains an appropriate model. The BP neural network model consists of an input layer, a hidden layer, and an output layer, as shown in Figure 6. Its characteristic is that the neurons in each layer are fully connected to the adjacent neurons, without connections between neurons in the same layer or feedback connections, thus forming a hierarchical feedforward neural network[12]. A single-layer feedforward neural network can only handle linearly separable problems, while a multi-layer feedforward neural network is required to handle nonlinear problems.

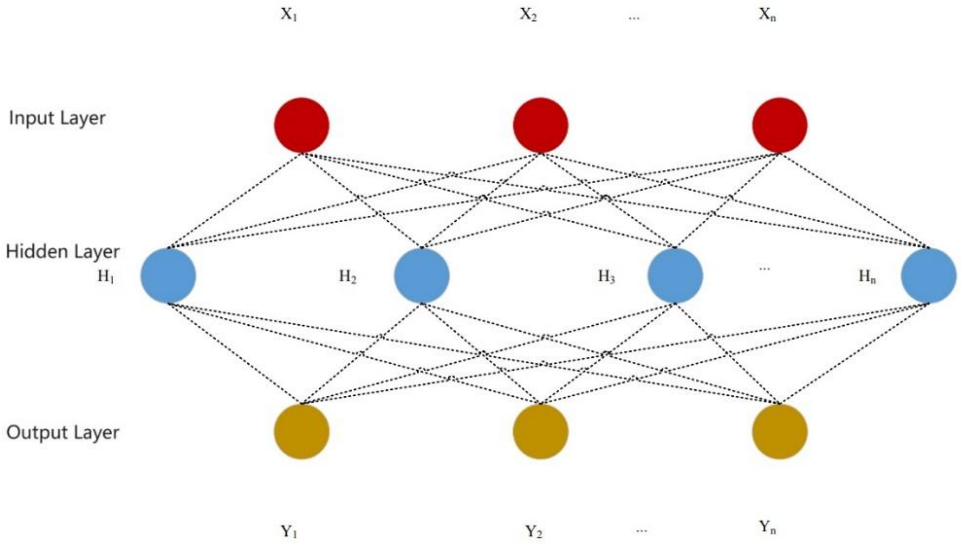

Figure 6: Schematic Diagram of the Basic Structure of the BP Neural Network

The BP neural network is a commonly used artificial neural network that can be applied to classification. A three-layer BP neural network structure was adopted, with only one hidden layer. To achieve optimal model performance, the neural network was trained using a training set, and the optimal parameter combination was obtained. The specific parameters are shown in Table 4. Through research, it has been found that the BP neural network model can improve the accuracy of prediction or classification.

Table 4: Optimization Parameter List of BP Neural Network

Model Parameter	Meaning	Optimal Value
lr	Learning Rate	0.1
n	Number of Hidden Layer Nodes	9

The results of the BP neural network model on the training set are as follows. Accuracy is 72.6%, $F1$ score is 71.5%, and AUC is 0.743. The results of testing the optimized parameter BP neural network model on the test set are as follows: accuracy is 71.6%, $F1$ score is 70.5%, and AUC is 0.729. Figure 7 shows the ROC curve of the BP neural network, indicating that the BP neural network model performs better than the logistic model in terms of prediction effectiveness.

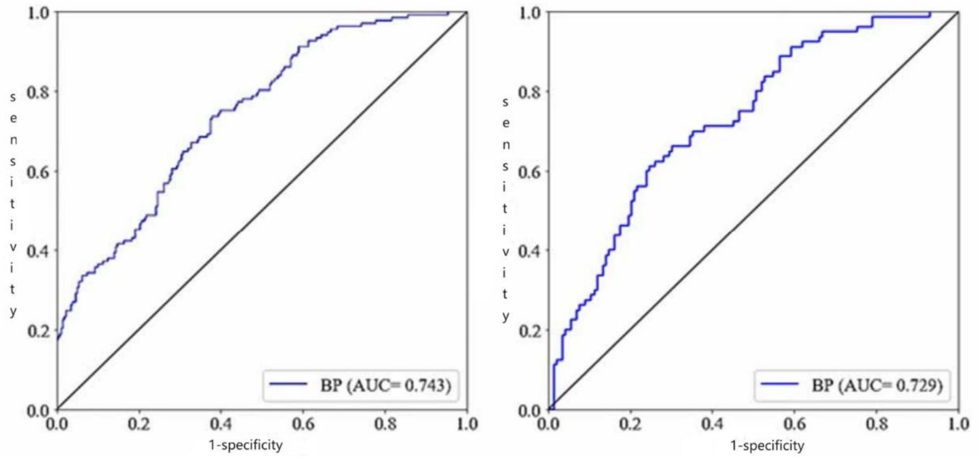

Figure 7: ROC Curve of the BP Neural Network Model

4.1.3 Support Vector Machine Model

Support Vector Machine(SVM) is a machine learning classification algorithm with a solid mathematical theory. It uses supervised learning to perform binary classification on data and is a useful tool for solving local minima and high-dimensional problems. The core idea of the SVM algorithm is to find a linear decision boundary that can separate binary data. In three-dimensional or higher-dimensional space, it seeks a hyperplane that can separate the two classes of data, with the goal of maximizing the distance between the data points closest to the separating hyperplane, denoted as $P1$ and $P2$, and the hyperplane itself.

Figure 8 illustrates two different classes of data represented by circles and squares in a two-dimensional plane.P represents the optimal decision boundary line, $P1$ and $P2$ are the samples closest to the boundary line after optimal division, and d is the distance between two parallel samples. The objective of SVM is to maximize the distance d between the support vectors $P1$ and $P2$ and the hyperplane. When the data becomes multidimensional, the goal is to find an optimal decision plane[13].

In cases where the data cannot be linearly separated, obtaining the solution to such problems is often challenging. Therefore, it is necessary to transform the data through nonlinear transformations to a corresponding high-dimensional space and then perform the separation. If the solution of SVM only involves inner product operations and there exists a function $K(x, x')$ in the low-dimensional space that is exactly equal to the inner product, i.e., $K(x, x') =< \phi(x), \phi(x') >$, then the SVM algorithm does not need to perform complex nonlinear transformations. Instead, it can directly obtain the inner product of the nonlinear transformation from the function $K(x, x')$, which allows mapping the low-dimensional data to a specific high-dimensional space, achieving linear separability of the data, and reducing computational complexity and memory overhead.

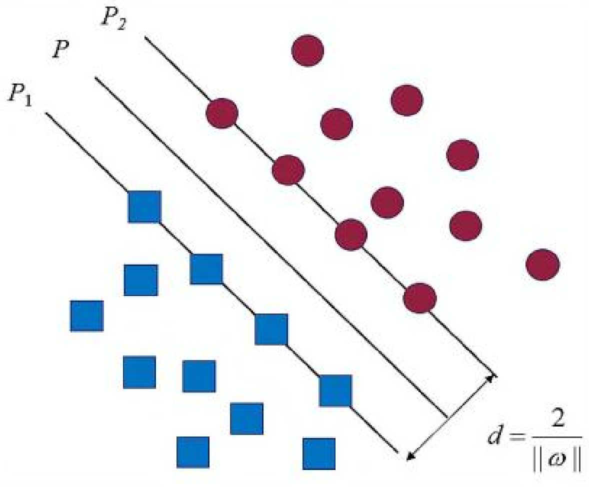

Figure 8: The Illustration of the Schematic Diagram of SVM

The commonly used kernel functions include linear, polynomial, and Gaussian functions. The expression for the linear kernel function is equation(10).

$$K(x, z) = x^T z + c \tag{10}$$

In equation(10), x and z represent vectors in the input space, and c is a constant. For the linear kernel function, the feature space F has the same dimensionality as the input space x and the vectors have the same number of features. The linear kernel function can be used when operations in the feature space are not required.

The expression for the polynomial kernel function is shown as equation(11).

$$K(x, z) = (\alpha x^T z + c)^d \tag{11}$$

The expression for the Gaussian kernel function is shown as equation(12).

$$K(x, z) = exp(-\gamma \|x - z\|^2) \tag{12}$$

A SVM prediction model based on default parameters was built using the training set. The parameters of the SVM prediction model were optimized using grid search combined and the optimal parameter combination is shown in Table 5.

Table 5: Optimal Parameters of Logistic Regression Model

Model Parameter	Meaning	Optimal Value
c	Penalty Parameter	1
kernel	Kernel Function Type	RBF
g	Coefficient for the Corresponding Kernel Function	0.1

The results on the training set for SVM is as follows. Accuracy is 73.7%, $F1$ score is 75.5%, and AUC is 0.768. The results on the test set for the optimized SVM model is that accuracy is 73.3%, $F1$ score is 74.4% and AUC is 0.760. Figure 9 shows the ROC curve of the SVM model, which indicates an improvement in performance compared to the Logistic model and the BP neural network model.

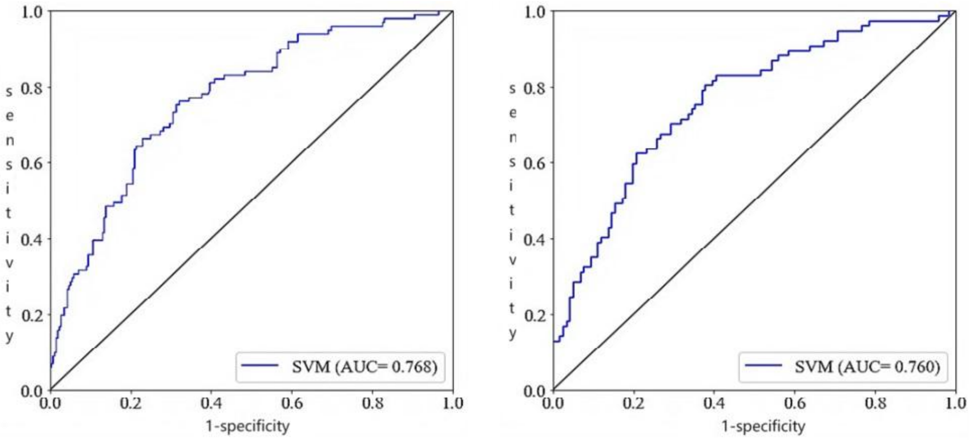

Figure 9: ROC Curve of the SVM Model

4.2 Cardiovascular Disease Prediction Using Integrated Models

Compared with a single model, the integrated model structure is more complex, because more parameters are introduced.The impact of parameters on the integrated model is greater than that of a single model, and the parameterization of the model becomes more difficult. This paper adopts the method of twice selection to determine the optimal parameters of the integrated model. The approximate range of the optimal parameters is determined and the combination of parameters with large intervals is adopted. Within the parameters selected for the first time, the optimal parameters of the model are determined by using the parameter network with small intervals. On the one hand, the optimal model parameters can be effectively selected. On the other hand, the time required for parameter selection can be effectively reduced and the computational efficiency can be improved.

4.2.1 Random Forest Prediction Model

Random forest is a method that uses multiple decision trees to train, classify and predict samples. In the process of data classification, its status in classification can be measured by the importance of each variable. "Random" in a random forest has two meanings: First, sample selection is based on sampling with retractions, which means that each sample has the potential to be selected multiple times or not selected. This sampling method can effectively increase the difference between decision trees and further reduce the risk of overfitting. Second, in the process of constructing the decision tree, the random forest does not use all the features to build each decision tree, but randomly selects a part of the features from all the features to build the decision tree. This method can effectively reduce the correlation between features and improve the performance of the model. In a random forest, each decision tree is built by iteratively partitioning the data. In the process of building the decision tree, the data is divided according to specific metrics until a predetermined stop condition is reached. In order to avoid overfitting problems, random forests can also limit the growth of decision trees by controlling parameters such as the depth of the tree and the minimum number of samples for nodes to stop splitting. The Settings of these parameters need to be adjusted according to the actual problem to obtain the best model performance.

The parameters of the random forest model were optimized by grid search method combined with 5-fold cross-validation and the optimal parameter combination was obtained, as shown in Table 6.

Table 6: Optimization parameter list of random forest model

Model Parameter	Meaning	Optimal Value
n_estimators	Number of Decision Trees	230
max_features	Maximum Number of a Single Decision Tree	4
min_samples_split	Minimum Number of Samples in a Leaf Node	3
min_samples_leaf	Minimum Number of Samples in Child Node	1
criterion	Evaluation Index	*gini*
max_depth	Maximum Depth of the Decision Tree	4

The results of random forest on the training set are as follows. Accuracy is 77.2%, $F1$ score is 78.5%, and AUC is 0.793. The results of Random Forest on the test set are as below. The accuracy rate was 76.2%, $F1$ score 77.6%, and AUC 0.788. Figure 10 shows the ROC curve of the random forest model. It can be seen from the figure that the evaluation indexes of the random forest model have been improved, and there is no significant difference between the results of the training set and the results of the test set.

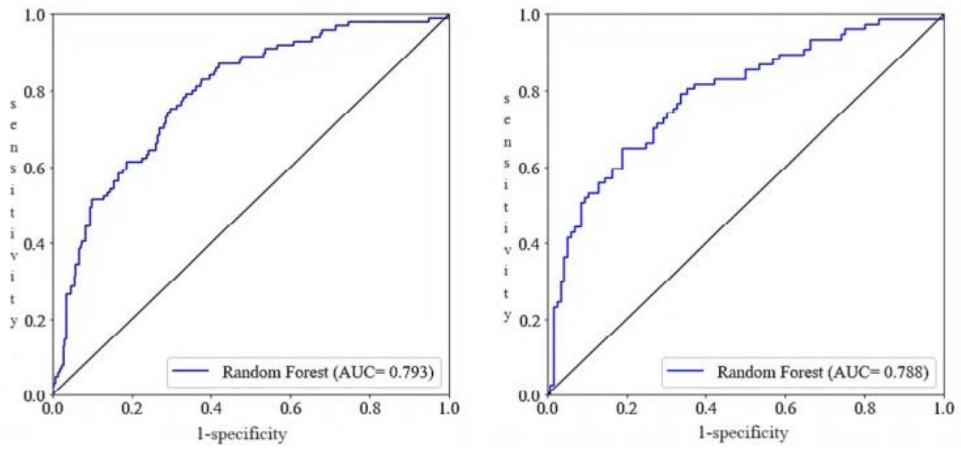

Figure 10: ROC Curve of Random Forest Model

4.2.2 XGBoost Prediction Model

XGBoost algorithm is under Boosting framework, whose essential difference is that the gain required for fitting the residual tree is different and the gain used by XGBoost is the difference of structure scores before and after splitting. An important feature of XGBoost is the introduction of a new splitting standard that minimizes segmentation losses at optimal segmentation points.

The core idea of XGBoost algorithm is three steps[14]. Firstly, the feature splitting method is adopted to continuously add trees and each additional tree is essentially learning a new function to fit the residual predicted last time. Secondly, after completing the training and obtaining k trees, the score of the sample should be predicted. Thirdly, the predicted value of the sample is the result of summing up the corresponding scores of each tree. After m iterations of XGBoost model, the definition of the objective function is shown in formula(13).

$$O^{(m)} = \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{(m-1)} + f_t(x_i)) + \Omega(f_i) \tag{13}$$

In the formula(13), l and Ω are loss functions and regular terms, respectively. y_i and $\widehat{y_i}$ represent the true value corresponding to the outcome and the predicted value of the model respectively. The number of samples is n.

The parameters of the XGBoost model are optimized by grid search method combined with 5-fold cross-validation and the optimal parameter combination is obtained, as shown in Table 7.

Table 7: Optimization Parameter List of XGBoost Model

Model Parameter	Meaning	Optimal Value
estimator	Maximum Number of Iterations	150
learning_rate	Learning Rate	0.2
min_child_sample	Minimum Amount of Data on a Leaf Node	12
gamma	Minimum Gain to be Added	0.5
subsample	Sample Percentage	0.9
colsample	Sample Proportion of the Variable	0.9
max_depth	Depth of the Tree	3

The results on the training set are as below. Accuracy is 79.7%, $F1$ score is 80.5%, AUC is 0.826. The results on the test set is that accuracy is 79.3%, $F1$ score is 79.6%, AUC is 0.820. Figure 11 shows the ROC curve of the XGBoost model. It can be seen from the figure that the ROC curve has improved compared with the random forest, indicating that the model has better performance in prediction. Therefore, the XGBoost model used performs well in parameter optimization and evaluation, which proves its effectiveness and reliability in solving classification problems.

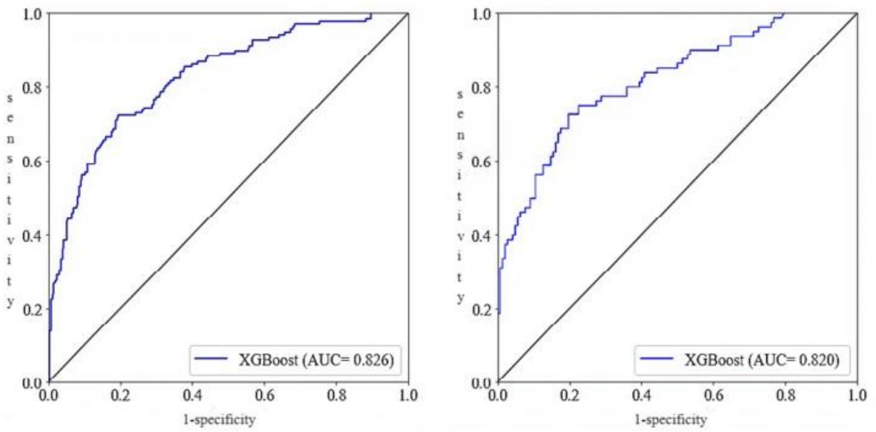

Figure 11: ROC Curve of XGBoost Model

4.2.3 LightGBM Prediction Model

LightGBM is designed to solve the challenges faced by GBDT when processing large-scale data, so that GBDT can be better and faster applied to industrial practices. By introducing histogram algorithm and adopting restricted leaf splitting strategy, lightGBM overcomes the shortcomings of XGBoost algorithm, such as excessive memory consumption and long training time.

The histogram algorithm uses the histogram to find the best segmentation points, processes the continuous variables, reduces the number of eigenvalues in the features and reduces the number of eigenvalues that need to be processed when the leaf nodes are split[15]. The basic idea is three steps. First, the continuous floating point feature values are discretized into k integers and a histogram with width of k is constructed. Then, when traversing the data, the discretized values are taken as exponents and accumulated in the histogram. On this basis, the traversal is carried out and the optimal segmentation point is found.

XGBoost algorithm adopts Level-wise as a growth strategy, as shown in Figure 12. This strategy traverses the data once and can split the leaves of the same layer at the same time, which is conducive to controlling the complexity of the model and achieving the effect of control fitting. However, in practical applications, the splitting gain of most leaves is relatively small, so there is no need to search and split the leaves, thus avoiding unnecessary calculations.

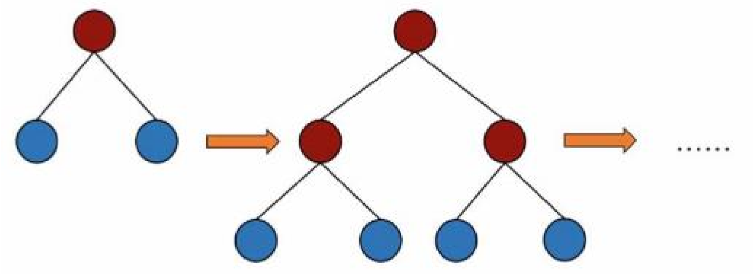

Figure 12: Level-wise Tree Growth

LightGBM algorithm uses Leaf-wise as a growth strategy. As shown in Figure 13, leaf-wise splits each time from all the current leaves, finds the Leaf with the greatest splitting gain and repeats the process. Compared with level-wise, leaf-wise algorithm has the following advantages. Under the same segmentation number, leaf-wise algorithm can effectively reduce the error and improve the accuracy of the algorithm. But the downside of Leaf-wise is that it will form a much deeper decision tree, leading to overfitting. Therefore, LightGBM algorithm adds a maximum depth limit to Leaf-wise to avoid overfitting and improve computing efficiency.

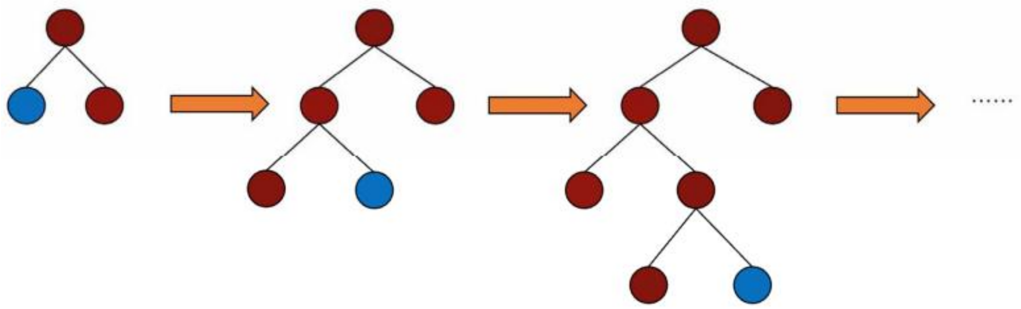

Figure 13: Leaf Splitting Constitutes a Decision Tree

Since LightGBM improves XGBoost by histogram algorithm and Leaf-wise strategy with depth restriction, the two are similar except for some parameter Settings. The optimal parameter combination obtained through optimization is shown in Table 8, where the parameter *num_leaves* reflects the complexity of trees in the model and is a unique parameter.

Table 8: Optimization Parameter List of XGBoost Model

Model Parameter	Meaning	Optimal Value
estimator	Maximum Number of Iterations	195
learning_rate	Learning Rate	0.1
max_depth	Depth of the Tree	5
min_child_sample	Minimum Amount of Data on a Leaf Node	20
num_leaves	Complexity of the Tree	26
subsample	Sample Percentage	0.8
colsample	Sample Proportion of the Variable	0.7

The results of the training set are that accuracy is 80.2%, $F1$ score is 81.3% and AUC is 0.836. The results of testing LightGBM model with optimal parameters with the test set are that accuracy is 80.2%, $F1$ score is 80.1% and AUC is 0.829. Figure 14 is the ROC curve of LightGBM model, from which we can see that the effect of LightGBM model is better than the previous two.

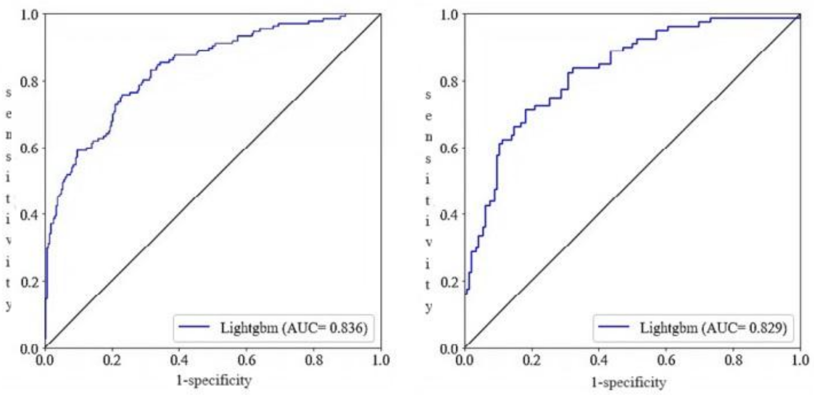

Figure 14: ROC Curve of LightGBM Model

5 Task 3: Comparison and Analysis of Results

In this paper, a single model and integrated model of cardiovascular disease prediction model are established based on the data after processing and feature selection, and the optimal prediction model was determined by various evaluation indicators. Table 9 shows the different evaluation indicators of the six models.

Table 9: Evaluation Indicators for Different Model Test Sets

Model	Accuracy	$F1.score$	AUC
Logistic Model	66.6%	61.6 %	0.678
BP Neural Network Model	73.3%	74.4 %	0.760
Support Vector Machine Model	71.6%	70.5 %	0.729
Random Forest Prediction Model	76.2%	77.6 %	0.788
XGBoost Prediction Model	79.3%	79.6 %	0.820
LightGBM Prediction Model	80.2%	80.1 %	0.829

Figure 15 shows the intuitive comparison of evaluation indicators.

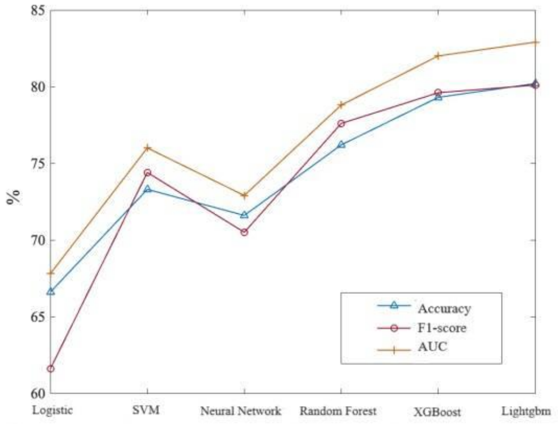

Figure 15: Comparison of Evaluation Indicators

As can be seen from the comparison of Figure 15, all indicators of the integrated model are better than that of the single model and the prediction ability and generalization ability of the model are enhanced. In addition, Logistic Model, as a linear model, has the worst prediction accuracy, which can also reflect that the nonlinear model is better in identifying variables. The results show that the LightGBM model with optimized parameters has the best results. However, non-linear ensemble models are more of a black box than linear models because they tend to consist of complex mathematical functions. This means that while nonlinear models can better identify complex relationships between variables and thus improve the model's prediction accuracy, it is difficult to find specific reasons for the high or low risk for a particular sample. This is a significant disadvantage for application scenarios where the model results need to be interpreted.

6 Conclusion

With the development of economy and the progress of society, people's lifestyle has undergone great changes and problems such as staying up late, smoking, irregular work and rest have become increasingly prominent, which have led to the increasing number of people in the world suffering from cardiovascular diseases. In recent years, with the popularity of computers, artificial intelligence, big data technology, Internet of things and other professions have developed rapidly and the communication with other industries has become increasingly frequent. The use of the advantages and characteristics of machine learning to predict cardiovascular diseases has become a topic sought after at home and abroad. In this paper, machine learning technology is used to construct a risk prediction model for cardiovascular diseases. The specific work of this paper is as follows.

- In terms of data processing, we carry out data quality analysis, data cleaning, data deletion and data conversion. In terms of exploratory data analysis, we use correlation analysis to quantitatively calculate the relationship between eleven indicators and cardiovascular diseases. Finally, we find that the four factors ap_hi, ap_ho, age and cholesterol play a major role in cardiovascular diseases.
- Grid search and cross-validation are used to determine the optimal parameters of the model. Evaluation indicators are compared to obtain the best predictive model. The results show that compared with the linear model, the nonlinear model has better prediction effect, and the LightGBM model with optimized parameters is the best among the nonlinear models.

Machine learning is widely used in disease prediction and it can process a variety of data formats in dynamic, high-volume and complex data environments, which can identify the impact of variables on outcomes more efficiently than past approaches[16]. Not only makes the final result highly accurate, but also can be used in multiple fields.

However, machine learning also has its drawbacks. Initial training is costly and time-consuming and requires a large amount of data support. At this stage, in the study of hemodialysis cardiovascular disease prediction research, machine learning is still in its infancy and it needs to be solved continuously in the future.

References

[1] Guo. , Li. , & Luo. (2013). Artificial intelligence and machine learning in cardiovascular health care. The Annals of thoracic surgery, 109(5), 1323-1329.

[2] Quer, G., Arnaout, R., Henne, M., & Arnaout, R. (2021). Machine learning and the future of cardiovascular care: JACC state-of-the-art review. Journal of the American College of Cardiology, 77(3), 300-313.

[3] Sun, W., Zhang, P., Wang, Z., & Li, D. (2021). Prediction of cardiovascular diseases based on machine learning. ASP Transactions on Internet of Things, 1(1), 30-35.

[4] Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European heart journal, 38(23), 1805-1814.

[5] Lu. (2009). Research on the Prediction Model of Cardiovascular Health Status Level Based on Data Mining. The Annals of thoracic surgery, 20(2), 12-15.

[6] Geng. (2012). Research on the Auxiliary Diagnosis Function of the Inspection Information System Based on Decision Tree. Journal of Heart Disease, 48(3), 162-171.

[7] Chen. (2014). Machine learning for cardiovascular biomechanics modeling: challenges and beyond. Annals of Biomedical Engineering, 50(6), 615-627.

[8] Pang. (2020). Machine learning based algorithm for risk prediction of cardio vascular disease (Cvd). Journal of critical reviews, 7(9), 836-844.

[9] Sitar, A., Zdrenghea, D., Pop, D., & Sitar-tut, D. (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. Age, 1(4), 4.

[10] Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020). Analysis and prediction of cardio vascular disease using machine learning classifiers. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 15-21). IEEE.

[11] Kagiyama, N., Tokodi, M., & Sengupta, P. P. (2022). Machine learning in cardiovascular imaging. Heart Failure Clinics, 18(2), 245-258.

[12] Friedrich, S., GroSS, S., König, I. R., Engelhardt, S., Bahls, M., Heinz, J., & Friede, T. (2021). Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. European Heart Journal-Digital Health, 2(3), 424-436.

[13] Ali, B., Gurbeta, L., & Badnjevic, A. (2017). Machine learning techniques for classification of diabetes and cardiovascular diseases. In 2017 6th mediterranean conference on embedded computing (MECO) (pp. 1-4). IEEE.

[14] Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE Access, 9, 106575-106588.

[15] Weikert, T., Francone, M., Abbara, S., Baessler, B., Choi, B. W., Gutberlet, M., & Leiner, T. (2021). Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges. European radiology, 31, 3909-3922.

[16] Kartal, E., & Balaban, M. E. (2018). Machine learning techniques in cardiac risk assessment. Turkish Journal of Thoracic and Cardiovascular Surgery, 26(3), 394.