Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
A		JS06012024140

Cardiovascular Disease Prediction Based on Machine Learning

Cardiovascular diseases are a class of diseases that pose a serious threat to human health, and the diagnosis of CVDs and the analysis of their causes are important references for the development of more effective CVD prevention measures and the reduction of the global burden of disease. To address this need, we explored the prediction of CVDs by different machine learning methods.

For **Task 1**, we performed three steps of **checking data quality**, **downscaling** and **data cleaning**. The data were transformed from the initial 70,000 to 68,585; in the exploratory analysis part, we performed three basic tasks: **descriptive statistics**, **correlation analysis** and **visualisation**. We solved the correlation coefficient as well as the standardised regression coefficient for continuous variables, and for categorical variables, we set up hypothesis tests based on the column table and chi-square statistic. The results showed that among the continuous variables, systolic blood pressure had the greatest contribution to the occurrence of cardiovascular diseases. Meanwhile, at a certain level of confidence, the occurrence of cardiovascular disease was associated with cholesterol level, blood glucose concentration, patients' smoking, alcohol consumption and exercise.

For **Task 2**, based on the nine relevant indicators identified in Task 1, we proposed three different models for CVD risk assessment and diagnosis. Firstly, we established a CVD risk assessment model based on **decision tree**, which used a multivariate test to improve the decision tree construction method to establish the model, and went through the operations of extracting behavioural norms, cross-validation and pruning in turn, and finally predicted the classification results, while obtaining the maximum decision parameter of 5. Secondly, we established a CVD risk model based on **logistic regression**, which is centred on the use of the sigmoid function to convert linear combinations into probability values and then make classification decisions based on the probability values. The parameters of the model are estimated by minimising the loss function in the training stage of the model; finally, a cardiovascular disease diagnostic model based on **support vector machine** is established. The model separates the samples of different categories by finding an optimal hyperplane, and the optimal parameter of this support vector machine is 'C' = 1 after training.

For **Task 3**, based on the outputs of three different CVD risk assessment and diagnosis models from Task 2, we analysed the evaluation metrics including **accuracy**, **precision**, **recall**, **F1 score**, and also compared and analysed the performance differences of the models by means of **confusion matrix** and **ROC curve**. The results show that the overall performance of the three models is similar, and each of the three classifiers has its own relative advantages in terms of evaluation metrics.

Key Words: Cardiovascular disease risk prediction; Machine learning; SVM; Decision tree; Logistic regression

Contents

1 Introduction . . . 1
- 1.1 Problem Background . . . 1
- 1.2 Restatement of the Problem . . . 1
- 1.3 Our Work . . . 2

2 Model Preparations . . . 2
- 2.1 Assumption . . . 2
- 2.2 Data Preprocessing . . . 3

3 Exploratory Data Analysis . . . 5
- 3.1 Descriptive Statistics . . . 5
- 3.2 Correlation Analysis . . . 7

4 Modal I : Decision Tree Based Cardiovascular Disease Risk Model . . . 10
- 4.1 Decision Tree Selection . . . 10
- 4.2 The Basic Process of Decision Tree . . . 11
- 4.3 Extract the Corresponding Behavior Rules . . . 11
- 4.4 Cross-validation . . . 11
- 4.5 Decision Tree Pruning . . . 12
- 4.6 Predictive Performance of Decision Tree Models . . . 13

5 Model II:Logistic Regression Based Cardiovascular Disease Risk Model . . . 14
- 5.1 Introduction to Basic Model . . . 14
- 5.2 Development of Model . . . 14
- 5.3 Evaluation Results . . . 16

6 Model III :SVM Based Cardiovascular Disease Diagnosis Model . . . 17
- 6.1 Model Building . . . 17
- 6.2 Model Designing . . . 18
- 6.3 Model Results . . . 19

7 Model Performance Comparison . . . 19

8 Model Evaluation, Improvement and Promotion . . . 21
- 8.1 Model Evaluation . . . 21
- 8.2 Model Improvement and Promotion . . . 22

References . . . 22

Appendices . . . 23

1 Introduction

1.1 Problem Background

Cardiovascular diseases (CVD) is a serious threat to human health, and in 2021, an estimated 20.5 million people worldwide will die of CVD, accounting for 33% of the total number of deaths, making it the leading cause of death worldwide. In China Cardiovascular Health and Disease Report 2022, it is pointed out that in the composition of urban and rural deaths in China, two out of every five deaths are attributed to CVD; it is projected that the global prevalence and mortality rate are still increasing, and the inflection point of the decline in the burden of disease has not yet appeared, and it has become a major public health problem affecting the economic and social development of the world.

The diagnosis of CVD and the analysis of its causes are important references for the development of more effective CVD prevention measures and the reduction of the global disease burden.

Figure 1: Cardiovascular Disease

1.2 Restatement of the Problem

The requirements given out are restated to help better orientate our focus.

- In Task 1, we need to perform preprocessing and exploratory analysis of the given data.
- In Task 2, we need to select a classification method in machine learning to predict whether a patient has cardiovascular disease based on the given information about patient metrics.
- In Task 3, we need to comprehensively evaluate the performance of the model, compare the prediction performance of different classifiers, and give conclusions and recommendations on the prediction of cardiovascular diseases, which can provide useful references for the medical industry.

1.3 Our Work

This problem essentially requires us to predict whether a patient has cardiovascular disease or not based on the patient's physiological indicators, medical testing indicators, and subjective information provided by the patient using classification methods or variable methods in machine learning. To avoid complex descriptions and to visualise our workflow, the flowchart is shown in Figure 2.

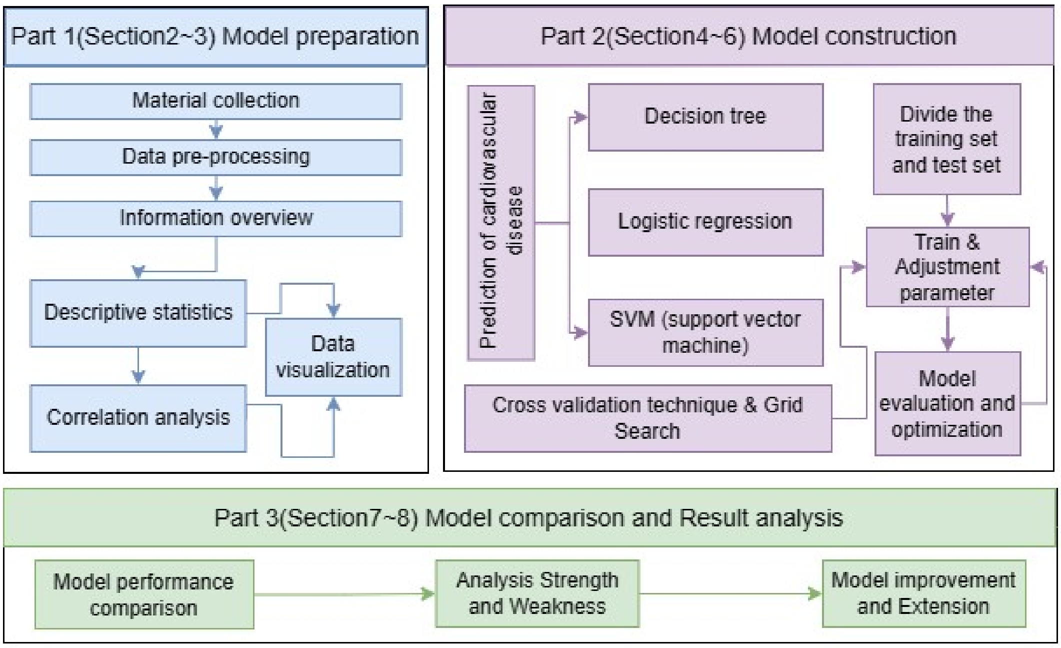

Part 1(Section2~3) Model preparation
Material collection
Data pre-processing
Information overview
Descriptive statistics
Data visualization
Correlation analysis
Part 2(Section4~6) Model construction
Prediction of cardiovascular disease
Decision tree
Logistic regression
SVM (support vector machine)
Cross validation technique & Grid Search
Divide the training set and test set
Train & Adjustment parameter
Model evaluation and optimization
Part 3(Section7~8) Model comparison and Result analysis
Model performance comparison
Analysis Strength and Weakness
Model improvement and Extension

Figure 2: Flow Chart

2 Model Preparations

2.1 Assumption

Assumption 1 The data given is true and reliable. In Task 1, the pre-processing and exploratory analysis of the given data needed to ensure the accuracy and completeness of the data. Therefore, it is assumed that the raw data has been collected and collated correctly, with no obvious errors or omissions.

Assumption 2 The selected machine learning algorithm is applicable to the problem. In Task 2, we select the classification method in machine learning for prediction. This hypothesis assumes that the selected algorithm is applicable to the problem and is able to learn the underlying patterns and regularities from the given data.

Assumption 3 Reliable performance evaluation of models. In Task 3, we need to fully assess the performance of the model and compare the prediction performance of different classifiers. This assumption assumes that the selected performance evaluation metrics and methods are reasonable and reliable, providing an objective evaluation of the model's predictive ability.

2.2 Data Preprocessing

2.2.1 Check data quality

Because the original data has the possibility of containing outliers, in order to avoid including outliers into the calculation and analysis process of the data without eliminating, which will have adverse effects on the results, the outliers are first judged. Because age, gender, and subjective information provided by patients are highly personalized, only an excel sheet was used for preliminary screening to confirm that there were no outliers beyond common sense. At the same time, considering the characteristics of height and weight vary from person to person, in order to facilitate the processing of abnormal values, the dimension is reduced to BMI, and the formula is

$$BMI = \frac{height(cm)}{weight(kg)^2} \tag{1}$$

The criteria for judging outliers in box plots are based on quartiles Q1, Q3 and interquartile distance IQR=Q3-Q1. The quartile has a certain resistance, as many as 25% of the data can become arbitrarily far away without greatly disturbing the quartile, so outliers cannot affect this criterion. The results of box plots to identify outliers are relatively objective. Therefore, box plots were used to analyze the abnormal values of systolic blood pressure, diastolic blood pressure and BMI.

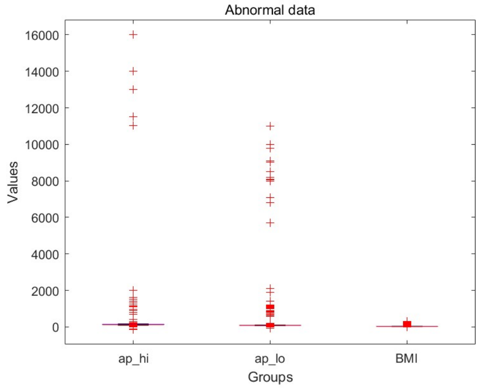


Figure 3: Box Plot of Unprocessed Data

Boxplots provide us with a criterion for identifying outliers: outliers are defined as values less than Q1-1.5IQR or greater than Q3+1.5IQR. However, considering that the data contained patients, that is, systolic blood pressure, diastolic blood pressure and BMI may not be within the normal range, the range of abnormal values was defined by referring to the Chinese Clinical Guidelines for Hypertension and the body mass index method of Lambert-Adolph Jacques Kettler on the basis of box plot data, and the specific range was as follows:

Table 1: The Filter Range of Given Data

index	scopemin	scopemax
ap_hi	90mmHg	220mmHg
ap_lo	40mmHg	140mmHg
BMI	10	60

2.2.2 Data cleaning

Due to the particularity of the collected data, that is, modifying an index will change the nature of the sample, and the outlier data only accounts for 2% of the total data, so the outlier data is deleted.

By browsing the data as a whole, duplicate and missing items are not included. Outliers here include values beyond the above specified range and cases where systolic blood pressure is less than diastolic blood pressure.

The specific deletion is as follows:

Table 2: Number of Deletion

index	amount
ap_hi	255
ap_lo	1023
aphi < aplo	10
height	1
totol	1415

The box plot of the deleted data shows that the outlier data is significantly reduced and the retained outliers are within a reasonable range and evenly distributed on both sides of the box.

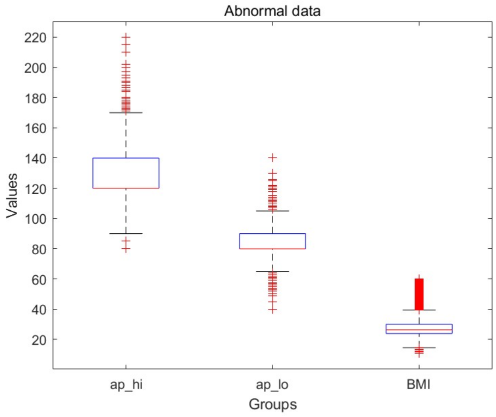


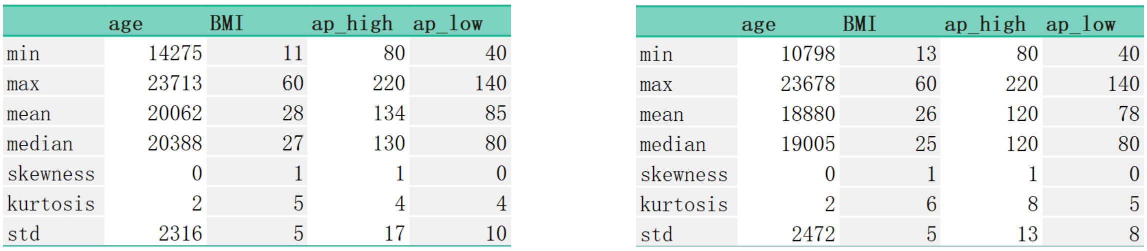
Figure 4: Box Plot of Processed Data

3 Exploratory Data Analysis

3.1 Descriptive Statistics

We calculated the maximum, mean and standard deviation of age, BMI, ap_hi and ap_lo of all samples, diseased samples and non-diseased samples, and obtained the following result figure:

The results in Figure 5 show that the skewness of both diseased and non-diseased samples is around 0, which is roughly symmetrical. The kurtosis is all around 3, which is close to normal distribution and can be tested for normality. When considering the dimension, the standard deviations of the four indicators are not very different, which indicates that the dispersion of the two types of samples is similar.



	age	BMI	ap_high	ap_low
min	14275	11	80	40
max	23713	60	220	140
mean	20062	28	134	85
median	20388	27	130	80
skewness	0	1	1	0
kurtosis	2	5	4	4
std	2316	5	17	10

	age	BMI	ap_high	ap_low
min	10798	13	80	40
max	23678	60	220	140
mean	18880	26	120	78
median	19005	25	120	80
skewness	0	1	1	0
kurtosis	2	6	8	5
std	2472	5	13	8

Figure 5: Descriptive Statistical of Continuous Variables[Sick (left); No disease (right)]

Since cholesterol and blood sugar levels, as well as smoking, drinking and exercise were categorical variables that were significantly associated with the presence or absence of cardiovascular disease in the sample, these indicators were statistically compared between the diseased and non-diseased samples.

In Figure 6,for all samples, blood sugar levels and cholesterol levels were divided into three grades,normal,above normal and a lot more than normal, The samples with normal levels accounted for the vast majority, while the samples with residual blood sugar levels and cholesterol levels accounted for a relatively small number and the number of samples of these two categories was similar.

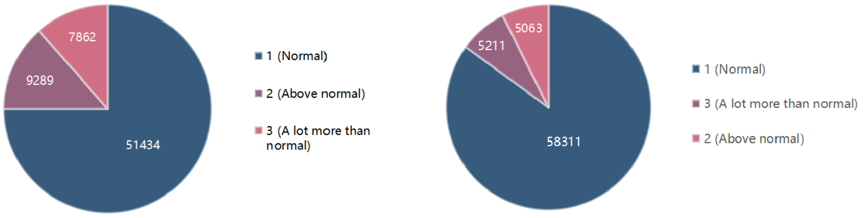


Figure 6: Visualization of Statistical Results of Categorical Variables

As for Figure 7, the majority of the samples do not smoke, do not drink and often exercise, and the subjective indicators of most of the samples are relatively positive.

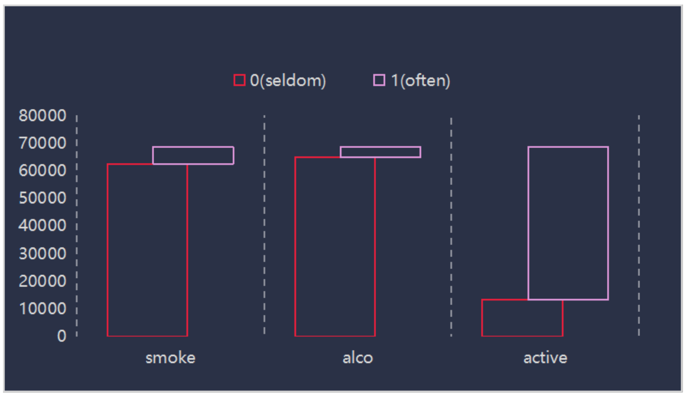


Figure 7: Smoking Drinking and Exercise Count Statistics

3.1.1 Test of Normality

Considering the correlation coefficient between the test indicators, it is necessary to take the normal distribution of the data as the premise, and understanding the distribution of the data is conducive to establishing a better model for exploration. We conducted a normal test on the samples

The main subjects were quantitative variables in the indicators: age,BMI, diastolic blood pressure, systolic blood pressure.

If the confidence level is 95 percent, the four indexes can be considered as normal distribution.

The QQ map is drawn to show the normal distribution more intuitively through the visual way.

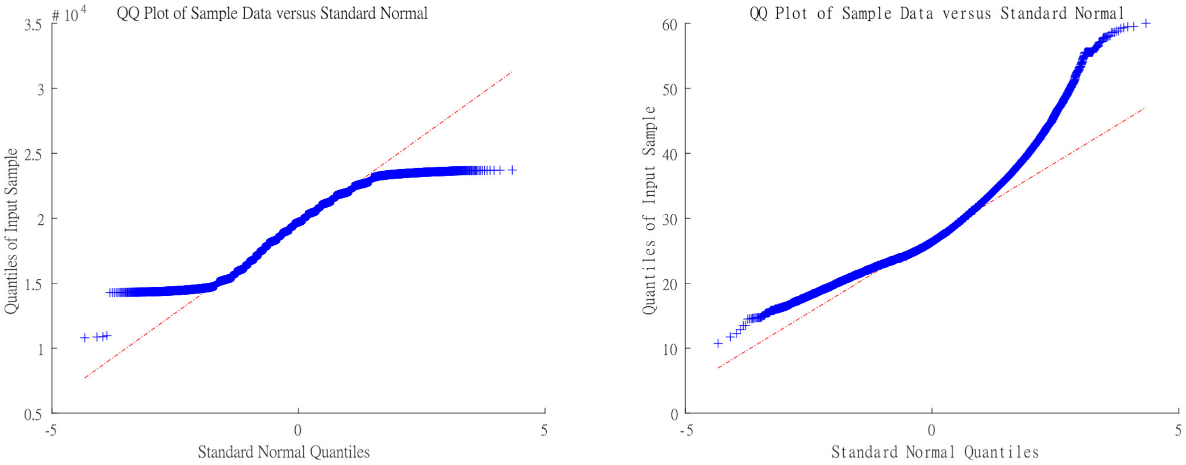


Figure 8: Normality Test of Age and BMI - QQ Chart

Due to the fact that some outliers are retained in the data preprocessing process of the data processing system, the two ends of the image are somewhat off the straight line, but the general trend is straight line. Under the premise of allowing a certain error, it can be considered that the four quantitative indicators are normally distributed.

3.2 Correlation Analysis

3.2.1 The correlation between the continuous independent variable and the dependent variable

The continuous variables after preprocessing include: physiological indicators: BMI; Two medical indicators: systolic blood pressure (the highest value of arterial blood pressure when the heart is contracting), diastolic blood pressure (the lowest value of arterial blood pressure when the heart is diastolic). Heat maps of these three measures (as dummy variables) are shown in Figure 9.

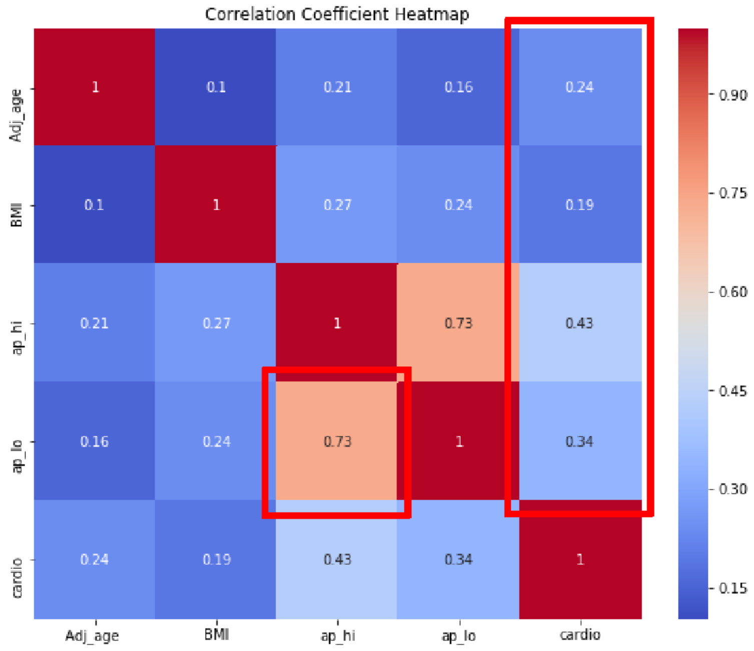


Figure 9: Correlation Heatmap

In the heat map, the red part on the right is the correlation coefficient between the occurrence of cardiovascular diseases and each index, which can be concluded as follows:

1. In a continuous variable, $\beta_{\text{ap\_}hi} > \beta_{ap\_lo} > \beta_{\text{Adj\_}age} > \beta_{BMI}$. Systolic blood pressure (the highest arterial blood pressure when the heart is contracting) has the largest contribution to the occurrence of cardiovascular disease among the continuous variables.

2. Among the independent variables, there is a large correlation between systolic blood pressure and diastolic blood pressure, which may lead to certain autocorrelation problems in the subsequent model establishment. If necessary, diastolic blood pressure can be removed to ensure the accuracy of the prediction results.

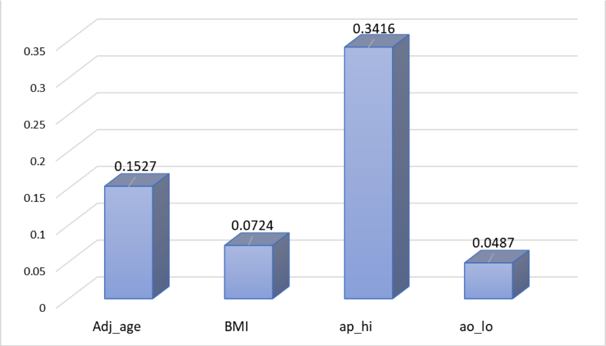


Figure 10: Visualize Results With Standardized Regression Coefficients

Because the data is standardized, the estimated coefficients can be replaced by the height of the columns, and the regression coefficients are comparable, allowing us to visually distinguish the impact of the independent variable on the dependent variable. Therefore, the following conclusion is drawn: in continuous variables, $\beta_{\text{ap\_}hi} > \beta_{Adj\_age} > \beta_{BMI} > \beta_{ao\_lo}$. Systolic blood pressure (the highest arterial blood pressure when the heart is contracting) has the largest contribution to the occurrence of cardiovascular disease among the continuous variables.

3.2.2 Correlation between categorical independent variables and dependent variables

Categorical variables after preprocessing include: physiological indicators: disease sex; Two medical indicators: cholesterol level, blood glucose concentration (the amount of glucose in the blood); Subjective information three: patient smoking, alcohol consumption and exercise.

Firstly, by integrating the incidence of cardiovascular diseases in different genders, as shown in Figure 10, we can easily observe that the proportion of patients with and without cardiovascular diseases in different genders is similar, about 50%, with no obvious difference. Therefore, the effect of gender on prevalence will not be covered below.

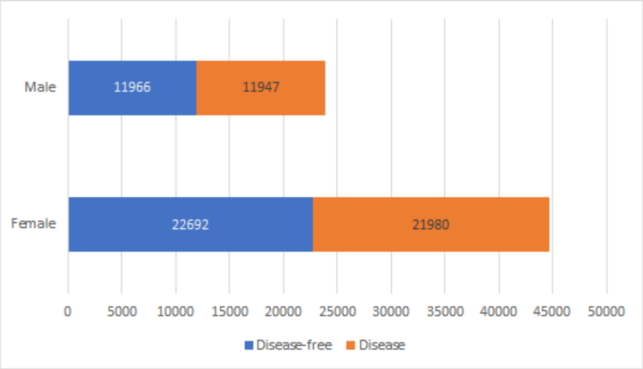
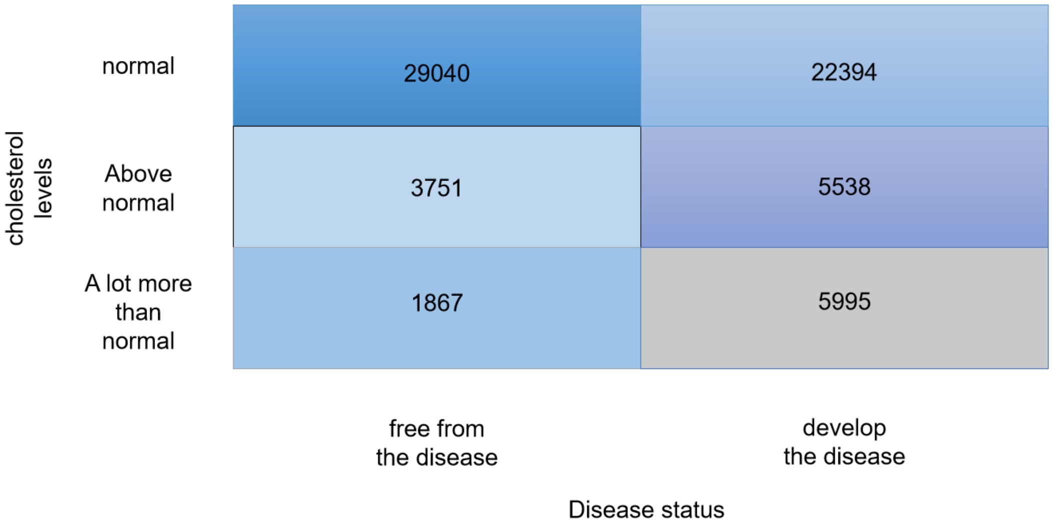


Figure 11: The Incidence of Cardiovascular Ddiseases in Different Genders

Next, we'll set up hypothesis tests based on contingency tables and chi-square statistics to investigate the relationship between cardiovascular disease and cholesterol levels, blood glucose levels (the amount of glucose in the blood), and how much the patient smoks, drinks, and exercises.

Due to the limitation of the text, the cardiovascular disease and the patient's cholesterol level was taken as examples.

- Test:
  Null hypothesis: Whether a patient has cardiovascular. disease and patient cholesterol level are independent.
  Alternative hypothesis: Whether the patient has cardiovascular disease and the patient's cholesterol level are not independent.



cholesterol levels	free from the disease	develop the disease
normal	29040	22394
Above normal	3751	5538
A lot more than normal	1867	5995

Disease status

Figure 12: Contingency Tables for Cholesterol Levels

Using R software, the following results were obtained:
Here is a contingency table of expected frequencies:

Table 3: Expected Frequency Contingency Tables for Cholesterol Levels

	No ill	ill
Normal	27822.220	6755.780
Above Normal	19520.934	4740.066
Above Normal Greatly	7841.846	1904.154

Each expected frequency in the contingency table is significantly greater than 5, and the results of the chi-square test are credible.

$$X\_squared = 13359.86$$
$$p\_value = 0$$

At the significance level of 0.05, the p-value is significantly less than the significance level of 0.05, so the null hypothesis is rejected at the 95% confidence level, that is, the patient has cardiovascular disease and the patient's cholesterol level is not independent, and the patient's cholesterol level has an effect on whether he has cardiovascular disease.

Table 4: Chi-square test results for other categorical variables

	X_square	P_value
Concentration of sugar in blood	13067.19	0
Alcohol consumption	4.808979	0.02831182
Sports situation	97.11327	6.547273e-23
Smoking situation	17.88664	2.344618e-05

From the output, we can see that, with 95% confidence, the occurrence of cardiovascular disease is related to cholesterol levels, blood glucose levels (the amount of glucose in the blood), the patient's smoking, and exercise. At the 90% confidence level, the occurrence of cardiovascular disease was also associated with patient exercise. The discussion and research in the following article can be continued.

4 Modal I : Decision Tree Based Cardiovascular Disease Risk Model

Based on the patient's physiological indicators, medical test indicators and subjective information provided by patients, we construct a decision tree model based on fine sets.

4.1 Decision Tree Selection

The model is also suitable for univariate decision tree, but the decision tree with single variable as the test attribute will make the generated decision tree have problems such as subtree duplication and some condition attributes need to be tested many times. Therefore, this paper will use multivariate tests to improve the decision tree construction method to build the model.

4.2 The Basic Process of Decision Tree

The specific process is as follows:

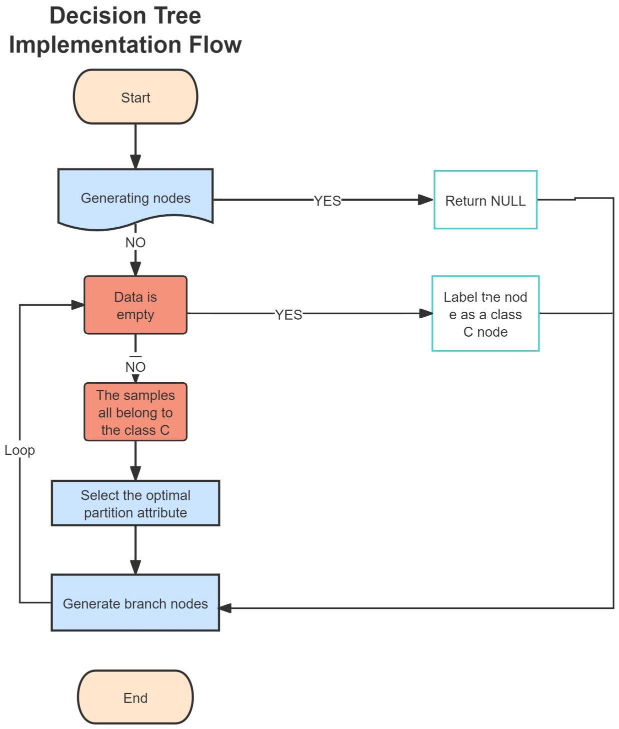


Figure 13: The Basic Process of Decision Tree

4.3 Extract the Corresponding Behavior Rules

In the decision tree, each leaf node represents a rule, that is, the left condition of the rule indicates that the root node starts to reach all the intermediate nodes on the path to the leaf node, and the right condition of the rule indicates the type of the leaf node.

According to the available data, the decision tree divided the attributes into 9 items: age, BMI, systolic blood pressure, diastolic blood pressure, cholesterol level, blood glucose concentration, smoking, drinking and regular exercise. Among them, gender was ignored because it had little effect on the disease outcome. At the same time, the response of the decision tree is whether the person has cardiovascular disease.

4.4 Cross-validation

In order to improve the accuracy of the decision tree model prediction, the decision tree should be trained with input data before testing it. In this paper, k-fold cross validation is used, and the specific steps are as follows:

1. Divide all the data into K folds.

2. Without repetition, take one sample at a time for the test set and the other K-1 samples for the training set to train the model, and then compute the MSE*i* of the model on the test set.

3. Take the average of the K-1 MSE*i* run to get the final MSE

$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSEi \tag{2}$$

After many tests, this model was cross-validated using 5-fold cross validation, that is, k=5.

4.5 Decision Tree Pruning

The resulting decision tree is visualized by matlab, and the following figure is obtained:

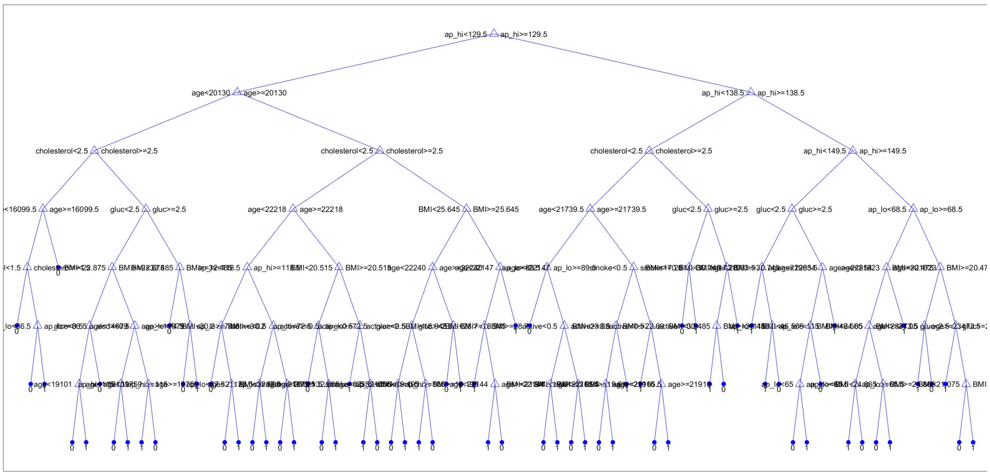


Figure 14: Visual Decision Tree

In order to improve the readability of the decision tree, the visualized decision tree is pruned.

If $e^t(t) \le e^t(Tt) + Se(e^t(Tt))$ is existed, the subtree (Tt) should be pruned.

Among them,

$$e^t(Tt) = \sum e(t) + \frac{Nt}{2} \tag{3}$$

$$e^t(t) \leq e(t) + \frac{1}{2} \tag{4}$$

$$Se(e^t(Tt)) = \sqrt{e^t(Tt)\frac{n(t) - e^t(Tt)}{n(t)}} \tag{5}$$

e(t) is the number of error samples caused by the pruning operation of node t, *n(t)* is the number of data samples of the decision tree at node t, and *Nt* is the number of leaves of the subtree *Tt* of the decision tree. The decision tree model after pruning is visualized, and the following is obtained:

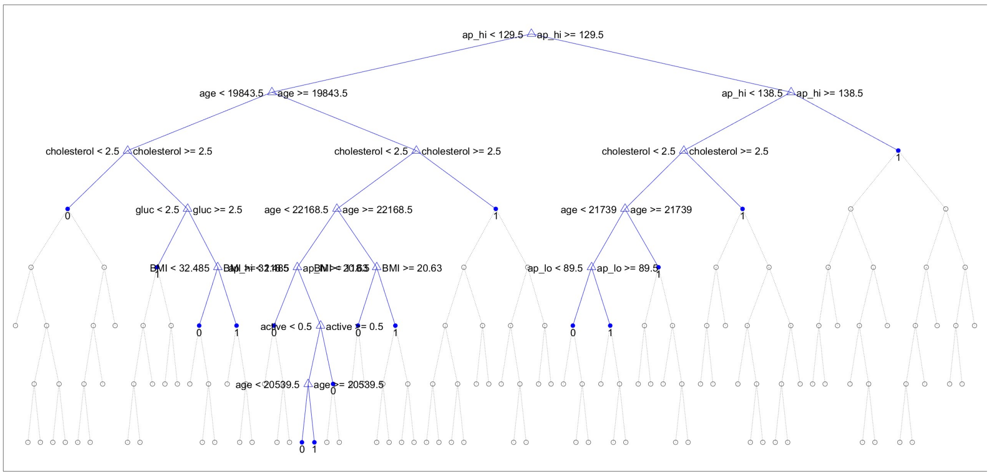


Figure 15: Visual Decision Tree after Pruning

4.6 Predictive Performance of Decision Tree Models

The prediction accuracy of the decision tree is 73.0% by matlab Classification Learner.

It can be seen from Figure 17 that the ROC curve is close to the upper left corner, which proves that TPR is significantly larger than FPR, that is, TPR is close to 1, indicating that the model prediction results have a high degree of fitting, and the AUC value is 0.7904, indicating that the cardiovascular disease risk assessment model based on decision tree is relatively complete.

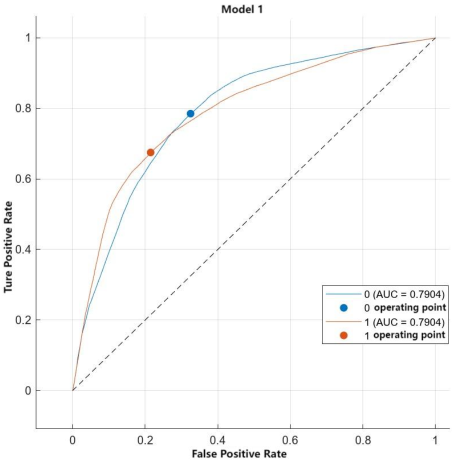


Figure 16: ROC Curve

5 Model II:Logistic Regression Based Cardiovascular Disease Risk Model

5.1 Introduction to Basic Model

Logistic regression is a statistical learning method used to solve binary classification problems, where the goal is to predict the probability that an input variable is associated with a particular class. The core principle of logistic regression is to use the sigmoid function to convert the linear combination into a probability value, and then make a classification decision based on the probability value. The steps are as follows:

1. The input features X are linearly combined with the model parameters: z $= X\theta$
2. Convert z to a probability using the Sigmoid function: h$_\theta(X) = \frac{1}{1+e^{-z}}$
3. The classification decision is made based on the value of he(X), usually 0.5 is used as the smell value. If he(X) > 0.5, then the prediction is positive class (1), otherwise the prediction is negative class (0) determines the probability of class association.

Based on this, we developed a logistic regression based CVD risk model.

5.2 Development of Model

Based on the patient's physiological indicators, medical indicators and subjective information provided by patients, we constructed a logistic regression based cardiovascular disease risk model.
The process is as follows:

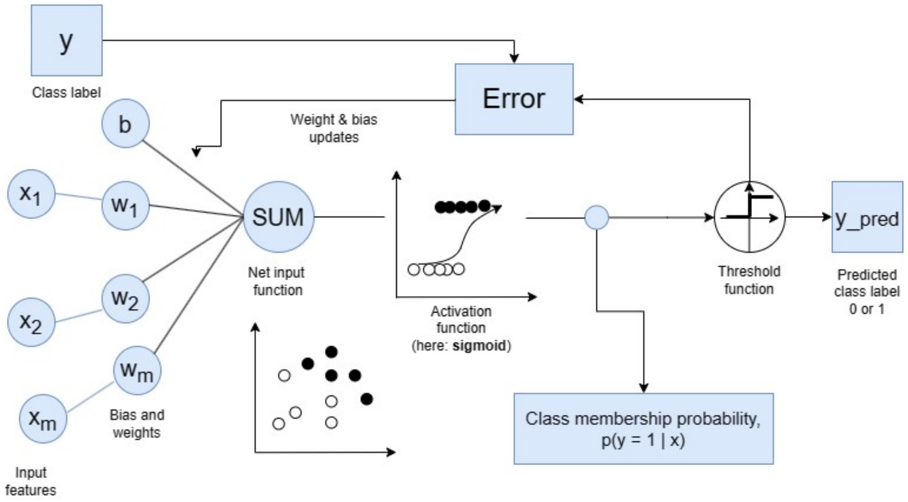


Figure 17: Machine Learning - Logistic Regression Internal Implementation Flowchart

From the regression coefficients, we can draw the following conclusions:

- Systolic blood pressure (the highest arterial blood pressure when the heart is contracting) has the greatest impact on cardiovascular disease, with age and cholesterol levels being the second and third factors.
- Among all the indicators, patient exercise is a negative indicator, that is, patients who exercise regularly have a lower probability of cardiovascular disease than those who do not exercise regularly.
- It is worth noting that although the regression coefficients of subjective information (smoking, drinking and exercise) of patients are small, they cannot be accurately measured due to the categorical variables, so the influence of individual behavioral norms on the occurrence of cardiovascular diseases cannot be ignored.

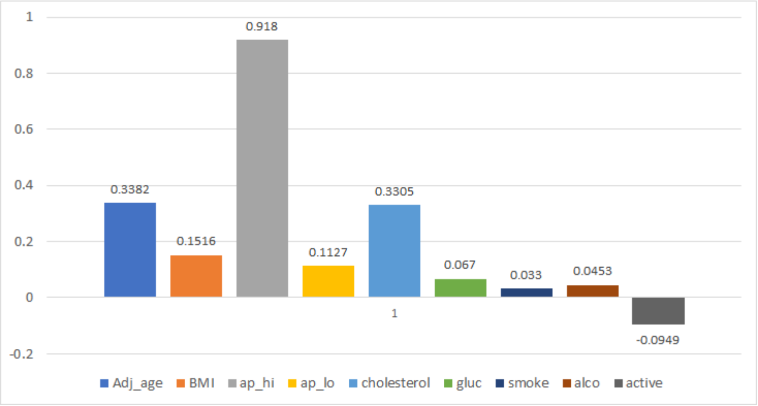


Figure 18: Regression Coefficient Visualization

5.3 Evaluation Results

Based on the confusion matrix of this model , the model correctly predicted 5467 positive samples in the test set as positive, but incorrectly predicted 1480 positive samples as negative. Meanwhile, the model correctly predicted 4543 negative class samples as negative class, but incorrectly predicted 2227 negative class samples as positive class.

The Classification Report in the Python software provides an evaluation of the precision, recall, and F1 score of the logistic regression model on each class.
See Table 5

Table 5: Evaluation Metrics

	Precision	Recall	F1-score	Support	Accuracy
0	0.71	0.79	0.75	6947	0.73
1	0.75	0.67	0.71	6770	0.73
Avg/total	0.73	0.73	0.73	13717	0.73

Interpretation of evaluation indicators:

- "precision" is the proportion of samples predicted to belong to the class. For class 0, 71% of the samples predicted as 0 were true zeros, and 75% of the samples predicted as 1 were false predictions. For class 1, 75% of the samples predicted as 1 were true ones, and 67% of the samples predicted as 0 were false predictions.
- "recall" represents the proportion of samples that actually belong to the class that were correctly predicted as that class. For class 0, 79% of the 0 labels were correctly predicted as 0, while 21% of the 0 labels were incorrectly predicted as 1. For class 1, 67% of the 1 labels are correctly predicted as 1, while 33% of the 1 labels are incorrectly predicted as 0.
- "f1-score" is a weighted average of precision and recall and can be considered as a comprehensive evaluation metric. The f1-score is 0.75 for class 0 and 0.71 for class 1.
- "support" indicates the number of samples per class. There are 6947 samples for class 0 and 6770 samples for class 1.
- "avg/total", gives the average over all classes. The average precision is 0.73, the average recall is 0.73, and the average f1-score is 0.73. This means that the overall performance of the model is good, but there are differences in the prediction results for different classes.

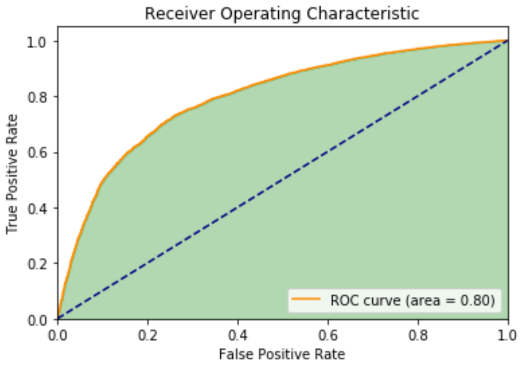


Figure 19: Logistic Regression ROC Curve

Receiver Operating Characteristic curve is a graphical tool used to evaluate the performance of binary classification models.

- The curve is closer to the top left, indicating better model performance. The top left point indicates that the model has a low false positive rate while maintaining a high recall, that is, mispredicting negative class samples as few as possible.
- the Area Under the Curve (AUC) is a measure of the performance of the model. An AUC of 0.8 indicates that the model distinguishes well between positive and negative cardiovascular disease samples.

6 Model III :SVM Based Cardiovascular Disease Diagnosis Model

6.1 Model Building

support vector machines (SVM) is a binary classification model. Its basic model is a linear classifier with the largest margin defined on the feature space, which distinguishes it from perceptrons.

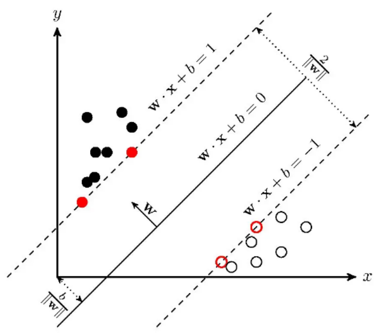

y
x
$\mathbf{w} \cdot \mathbf{x} + b = 1$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = -1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}$
$\frac{b}{\|\mathbf{w}\|}$

Figure 20: SVM

For a given dataset T and hyperplane $w^{*}x+b=0$, define the geometric distance of the hyperplane with respect to the sample point (x_i, y_i) as

$$\gamma_i = y_i\left(\frac{w}{\|w\|}\cdot x_i + \frac{b}{\|w\|}\right) \tag{6}$$

The minimum value of the geometric interval of the hyperplane with respect to all sample points is

$$\gamma = \min_{i=1,2,\ldots,N} \gamma_i \tag{7}$$

The problem of finding the maximum splitting hyperplane for SVM model can be formulated as the following constrained optimization problem:

$$\min_{w,b} \theta\left(w\right) = \min_{w,b}\max_{a_i\geq 0}\left(w,b,\alpha\right) = p^{*} \tag{8}$$

Therefore, for the second problem, we establish a classification model based on support vector machine, which takes nine indicators as input to classify and predict whether the sample has cardiovascular disease.

6.2 Model Designing

Input Training set data

Among them, we initially set the ratio of training set and test set to 80% and 20%.

Self-changing face X is age,BMI, diastolic blood pressure, systolic blood pressure, and other nine indicators that are significantly related to the presence or absence of cardiovascular disease.

The presence or absence of disease is a categorical variable y, with 0 indicating no disease and 1 indicating cardiovascular disease.

Output Separating hyperplanes and classification decision functions

(1) (1) Choose the penalty parameter C > 0, construct and solve the convex quadratic programming problem.

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{N}\alpha_i \tag{9}$$

$$s,t,\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{10}$$

Among them, $0\leq x_i\leq C, i=1,2,\ldots,N$
Get the optimal solution $\alpha^{*}=\left(\alpha_{11}^{*}\alpha_2^{*},\ldots,\alpha_N^{*}\right)T$

(2) computation

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \tag{11}$$

Choose a component $\alpha*$ of α_j^*,which satisfies the condition $0 < \alpha_j^* < C$ and calculate

$$b^* = yi - \sum_{i=1}^{N} \alpha i^* yi(xi \bullet xj) \tag{12}$$

(3) Solving for hyperplanes

$$W^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \tag{13}$$

Classification decision function

$$f\left(x\right) = sign\left(w^* * x + b^*\right) \tag{14}$$

6.3 Model Results

The data were imported into matlab for training, and finally the training results were obtained. After several training, the highest accuracy of support vector machine is 72.6%.

The following confusion matrix shows the more refined accuracy and confusion rates between the two classification results, where the highest accuracy can reach 82%, and the lowest mixing efficiency is only 18%, which verifies the feasibility of the model establishment.

At the same time, the ROC chart of the current model is also obtained. According to the ROC chart, the FPR of the current model is 0.37, the TPR is 0.82, and the AUC value is generally significantly greater than 0.5, which indicates that the diagnostic criteria are correct and feasible.

7 Model Performance Comparison

The confusion matrices of the three prediction models are shown in Fig. 27.The confusion matrices of the three prediction models demonstrate the confusion matrices of the different prediction models under the training and validation sets of the dataset of this study.From the confusion matrices, it can be seen that the true-positive (TP) and true-negative (TN) accounts for the majority of the dataset both in the training and validation sets, which suggests that the prediction models of machine learning used in this study are relatively efficient for the cardiovascular disease prediction is relatively efficient.

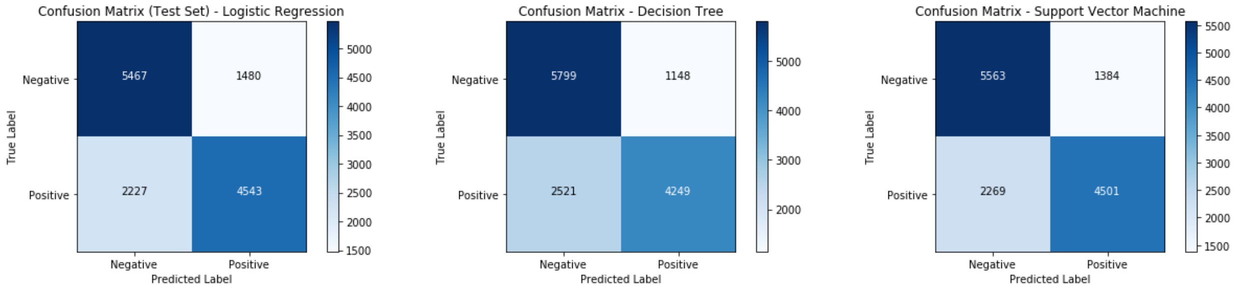


Figure 21: Confusion matrix of test set model

The subject operating characteristic (ROC) curves of the three machine learning prediction models in the training and validation sets are shown in Figure 2.The ROC curves of the three models have similar results in the training set, but the logistic regression model and the decision tree model have a larger AUC compared to the SVM in the results of the test set (the AUC metrics, in turn, are capable of assessing the overall performance of the classifiers). Nonetheless, the ROC lines of decision tree, support vector machine and logistic regression do not differ much in terms of AUC and have similar results both in the training set and in the validation set. Further metrics are needed to compare the predictive performance of the three classifiers.

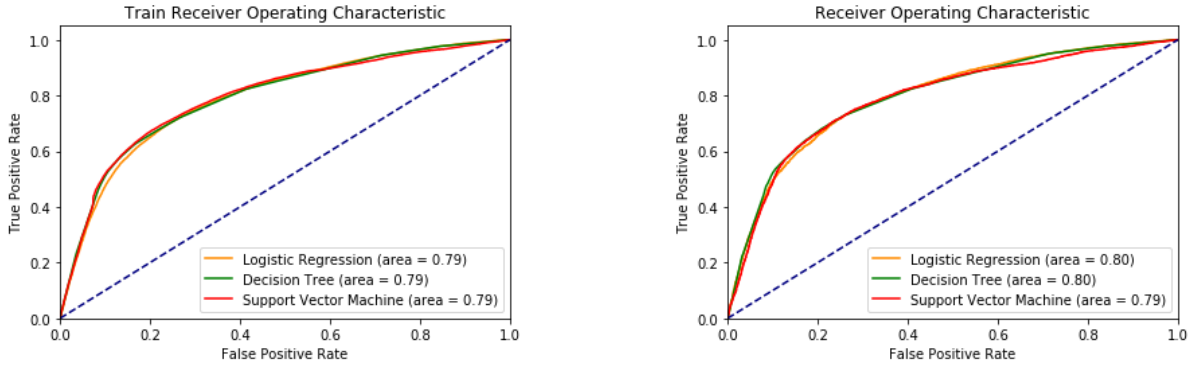


Figure 22: ROC Curves[the training set(left) ; the test set(right)]

The confusion matrix gives the accuracy, recall, precision, F1 value and AUC of the three prediction models as shown in Table 6.The test set is used to test the generalisation ability of the models.

1. Decision tree model shows the highest accuracy, precision and AUC on the test set. this means that the model has high accuracy in predicting positive and negative cases and can distinguish between positive and negative cases well.

2. The SVM model has the highest F1 value on the test set. the F1 value is an indicator that takes accuracy and recall into consideration, so it means that the model has a good balance between accuracy and completeness.

3. The logistic regression model has the highest recall and AUC on the test set. recall measures the model's ability to identify positive examples, while AUC measures the comprehensive performance of the model's predictive ability. Therefore, the logistic regression model performed well in identifying positive examples and had better overall prediction performance.

Based on the above results, each of the three models has its own relative advantages, so the final choice depends on the importance and requirements of the actual scenario for the model performance metrics. In cardiovascular disease prediction, if the accuracy of diagnosis is more important, the precision rate and accuracy rate are more important, and it is more appropriate to choose the decision tree model at this time; if it is more concerned about avoiding the missed diagnosis, the recall rate and F1 index may be more important, and it is more appropriate to choose the SVM model or the logistic regression model at this time.

Table 6: Comparison of Three Machine Learning Models

Method	Accuracy	Recall	Precision Radio	F1 Index	AUC
Decision Tree	0.7352	0.6276	0.7873	0.6984	0.80000
SVM	0.7337	0.6648	0.7648	0.7113	0.7900
Logistic Regression	0.7297	0.6710	0.7543	0.7102	0.8000

8 Model Evaluation, Improvement and Promotion

8.1 Model Evaluation

8.1.1 Advantages

1. In the second problem, we train three machine learning algorithms: logistic regression, decision tree, and support vector machine multiple times to obtain the best parameter state of each model.

2. In the third question, we conduct a comprehensive evaluation of the three selected models, from the five indicators of accuracy, recall, precision,F1 index and AUC performance evaluation.

8.1.2 Disadvantages

In real life, the evaluation and prediction of a problem involves many factors, even factors that cannot be predicted by the model, but in order to ensure the integrity of the mathematical modeling process, we selected several important indicators through correlation tests, which may cause a certain degree of bias in the evaluation and prediction.

8.2 Model Improvement and Promotion

1. We can collect more extensive and accurate data, build a more complete and effective mathematical model, and more accurately predict whether a patient has cardiovascular disease.

2. By establishing the mathematical model of physiological indicators, medical indicators and subjective information, it can be extended to other diseases, and the best classification model can be selected according to the indicators of different diseases.

References

[1] World Heart Federation. World Heart Report 2023:Confronting the World's Number One Killer[M]. Geneva,Switzerland:World Heart Federation,2023:3-5.

[2] Hu Shengshou, Wang Zengwu. Summary of China Cardiovascular Health and disease Report 2022 [J]. Chinese Journal of Cardiovascular Disease, 2023,21 (7) : 577-600. (in Chinese) DOI: 10.3969/j.issn.1672-5301.2023.07.001.

[3] Hang li. (2019). Statistical Learning Methods (2nd ed.). Tsinghua University Press.

[4] Perly. (2023, June 24). Machine Learning with Python - Classification Model Evaluation & Classification Algorithms (k-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Svm). CSDN. https://www.csdn.net/

[5] Mert ozcanse , & serhat peker. (2023). A Classification and Regression Tree Algorithm for Heart Disease Modeling and Prediction. Healthcare Analytics, 3(100130). https://doi.org/10.1016/j.health.2022.100130

[6] Matt gifford, & Tuncay bayrak. (2023). A Predictive Analytics Model for Forecasting Outcomes in the National Football League Games Using Decision Tree and Logistic Regression. Decision Analytics Journal, 8(100296). https://doi.org/10.1016/j.dajour.2023.100296

Appendices

Appendix A: Programmes Codes

Here are simulation programmes we used in our model as follow.

Input Python source:

```
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt

data = pd.read_excel('D:\heart_disease.xlsx')
X = data[['Adj_age','BMI','ap_hi','ap_lo','cholesterol','gluc','smoke','alco','active']]
y = data['cardio']
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)


X_train,X_test,y_train,y_test=train_test_split(X_scaled,y,test_size=0.2,random_state=42)
clf = LogisticRegression()


param_grid = {'C': [0.001, 0.01, 0.1, 1, 10]}

grid_search = GridSearchCV(clf, param_grid, cv=5)
grid_search.fit(X_train, y_train)
best_params = grid_search.best_params_

clf = LogisticRegression(**best_params)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

acc = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)

print('Accuracy:', acc)
print('Confusion Matrix:\n', cm)

classification_metrics = classification_report(y_test, y_pred)
print('Classification Report:\n', classification_metrics)

y_pred_prob = clf.predict_proba(X_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', label='ROC curve (area = %0.2f)'
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.fill_between(fpr, tpr, color='blue', alpha=0.3)
plt.show()

print(clf)
print("Logistic regression coefficients:", clf.coef_)
```

Input matlab source:

```
function [trainedClassifier, validationAccuracy] = trainClassifier(trainingData)
inputTable = trainingData;
predictorNames = {'age', 'BMI', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'smoke',
    'alco', 'active'};
predictors = inputTable(:, predictorNames);
response = inputTable.cardio;
isCategoricalPredictor = [false, false, false, false, false, false, false, false,
    false];
classNames = [0; 1];
classificationTree = fitctree(...
    predictors, ...
    response, ...
    'SplitCriterion', 'gdi', ...
    'MaxNumSplits', 100, ...
    'Surrogate', 'off', ...
    'ClassNames', classNames);
predictorExtractionFcn = @(t) t(:, predictorNames);
treePredictFcn = @(x) predict(classificationTree, x);
trainedClassifier.predictFcn = @(x) treePredictFcn(predictorExtractionFcn(x));
trainedClassifier.RequiredVariables = {'age', 'BMI', 'ap_hi', 'ap_lo', 'cholesterol',
    'gluc', 'smoke', 'alco', 'active'};
trainedClassifier.ClassificationTree = classificationTree;
inputTable = trainingData;
predictorNames = {'age', 'BMI', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'smoke',
    'alco', 'active'};
predictors = inputTable(:, predictorNames);
response = inputTable.cardio;
isCategoricalPredictor = [false, false, false, false, false, false, false, false,
    false];
classNames = [0; 1];
partitionedModel = crossval(trainedClassifier.ClassificationTree, 'KFold', 5);
[validationPredictions, validationScores] = kfoldPredict(partitionedModel);
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

Report on Use of AI

1. OpenAI ChatGPT (Jan 16, 2024 version, ChatGPT-4)

Query1 What are the visualization processing of support vector machine?

Output1 Support Vector Machine (SVM) is a commonly used machine learning algorithm, which has a wide range of applications in classification, regression and anomaly detection. SVM can be visualized to help us understand the performance and characteristics of the model and make further adjustments and optimizations.
Here are some common SVM visualization methods:

Scatter plot Two-dimensional or three-dimensional scatter plot can be used to display the distribution and classification of data. For binary classification problems, different colors or shapes can be used to distinguish the two types of samples, which can intuitively show the classification effect of SVM.

Classification hyperplane SVM can divide data into two categories and build a maximally spaced hyperplane to distinguish the two categories. Two-dimensional or three-dimensional graphics can be used to show the hyperplane and data distribution, which can better understand the classification principle and effect of SVM.

Decision boundaries SVM can use decision boundaries to divide data into two categories. Contour lines or three-dimensional surfaces can be used to show decision boundaries and data distribution, which can better understand the classification principle and effect of SVM.

Kernel functions SVM can use different kernel functions to handle nonlinear classification problems, such as Gaussian kernel functions, polynomial kernel functions, etc. Two - or three-dimensional graphics can be used to show the effects of kernel functions and data distribution, so that you can better understand the nonlinear processing principle and effect of SVM.

Support vectors SVM can use support vectors to determine classification hyperplanes and decision boundaries. Scatterplot or contour plot can be used to show support vectors and classification effects, so that you can better understand the classification principle and effect of SVM.

The above are some commonly used SVM visualization methods, which can help us better understand the classification principle and effect of SVM, and then adjust and optimize the model.

Query2 Decide what pruning methods the tree has?

Output2 Decision tree pruning is the process of simplifying the decision tree in order to avoid overfitting. Here are some common decision tree pruning methods:

Pre-pruning Evaluation is carried out on each node during the construction of the decision tree. If the continued division of nodes cannot significantly improve the performance of the model, the division is stopped and the node is marked as a leaf node. The evaluation indexes of pre-pruning methods are information gain, Gini index, etc.

Post-pruning Build a complete decision tree first, then prune the tree from the bottom up. The specific method is to evaluate each non-leaf node, replace the node with a leaf node, and prune the tree if there is no significant decline in tree performance after pruning. Cross-validation and validation set error can be used to evaluate the performance after pruning.

Minimal Error Pruning In the post-pruning process, the subtree with the smallest classification error is selected for pruning.

Pessimistic Error Pruning In the process of post-pruning, statistical methods are used to calculate the confidence interval of each node, and pruning is carried out according to the confidence interval.

Cost Complexity Pruning In the post-pruning process, a cost complexity function is introduced, which takes into account the complexity and performance of the model comprehensively. The optimal pruning point is found by adjusting the pruning parameters.

2. OpenAI Ernie (Jan 17, 2024 version, Ernie 4.0)

Query1 What machine learning algorithms can make classification predictions?

Output1 There are many algorithms in machine learning that can be used for classification prediction, and here are some of the common ones:

Decision tree A decision tree is a common classification algorithm that classifies instances by building a tree graph. Decision trees can be used to solve various classification problems, such as mail filtering, spam identification, credit scoring, etc.

Naive Bayes Naive Bayes is a classification algorithm based on Bayes' theorem that classifies instances by calculating the probability of each class. Naive Bayes assumes that features are independent of each other, so it is suitable for text classification, spam recognition and other scenarios.

Support Vector Machine (SVM) SVM is a supervised learning algorithm that implements classification by finding decision boundaries that maximize the separation of instances of different classes. SVM is suitable for solving various classification problems, such as face recognition, handwritten digit recognition, etc.

K-nearest Neighbor (KNN) KNN is an instance-based learning algorithm that implements classification by assigning an instance to the class of the K training instances closest to it. KNN is suitable for solving various classification problems, such as text classification, image recognition, etc.

Neural network A neural network is a complex machine learning algorithm that performs classification by simulating the way neurons in the human brain are connected. Neural networks can be used to solve various classification problems, such as image recognition, speech recognition, natural language processing, etc.

These are some common classification algorithms, but not all of them. In fact, there are many other algorithms and models in the field of machine learning that can be used for categorical predictions, and which one is chosen depends on the nature of the problem and the characteristics of the data.

Query2 Is there any way to evaluate the model?

Output2 There are many ways to evaluate machine learning models, but here are some common ones:

Accuracy Accuracy is the most basic index in classification model evaluation, which represents the proportion of the number of samples correctly predicted by the model to the total number of samples.

Accuracy rate Accuracy rate refers to the proportion of samples that the model predicts to be positive examples that are actually positive examples.

Recall rate Recall rate refers to the proportion of all samples that are truly positive cases that are predicted to be positive cases by the model.

F1 value The F1 value is the harmonic average of the accuracy rate and the recall rate, and is used to consider the performance of the accuracy rate and the recall rate.

ROC curve and AUC The ROC curve is drawn with the false positive rate as the horizontal axis, the true positive rate as the vertical axis, and the AUC is the area under the ROC curve. ROC curves and AUC can be used to evaluate the performance of classification models, especially for unbalanced classification problems.

Mean square error, mean absolute error, R-square value These indicators are used to evaluate the performance of the regression model, mean square error represents the mean of the square of the difference between the predicted and the true value, mean absolute error is the mean of the absolute value of the difference between the predicted and the true value, and R-square value represents the degree of variability that the model can account for.

Cross-validation Cross-validation is a method of splitting a data set into multiple parts and using one part of the data for model training and the other part for testing the model. By repeating the validation many times and averaging the results, the performance of the model can be more accurately evaluated.

Learning curve A learning curve is a method of evaluating the performance of a model by observing how the model's performance on the training and test sets changes as the number of samples changes, helping to determine whether there is overfitting or underfitting.

These are some common methods for evaluating machine learning models, which one you use depends on the nature of the problem and the characteristics of the data. In practical application, it is usually necessary to use a variety of evaluation indicators and methods to comprehensively evaluate the performance of the model.