Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/27/20240731

Comparison of dimensionality reduction methods and evaluation of prediction effect based on cardiovascular disease data

Shuyue Jing

School of Management, Guangdong University of Technology, Guangzhou, Guangdong, 510520, China

3221003390@mail2.gdut.edu.cn

Abstract. Cardiovascular diseases have become the leading cause of death in the world, and the mortality rate caused by them is still on the rise, which is a major challenge facing the world. In the study of cardiovascular diseases, there are many kinds and quantities of factors leading to the disease, so how to screen more effective factors for research and accurate prediction is an important problem. The aim of this study is to conduct an in-depth comparative analysis of multiple dimensionality reduction methods for cardiovascular disease data. This paper evaluates the results produced by various dimensionality reduction methods and models, and uses Accuracy Rate, Confusion Matrix and Area Under Curve (AUC) and other indicators to evaluate their prediction effects. The study finds that the decision tree model has the best performance and is superior to LDA linear discriminant analysis in terms of accuracy and data dimension. The principal component analysis method is relatively complex, although it can effectively reduce the data dimension, its accuracy in forecasting decreases, and this method is not conducive to horizontal and vertical comparison and the accumulation of statistical data.

Keywords: Cardiovascular diseases, dimensionality reduction methods, prediction, accuracy, data dimensions.

1. Introduction

Cardiovascular diseases (CVDs), as the world's leading fatal diseases [1], have shown a sharp increase in recent years [2], and are a major global challenge to public health. Effective preventive measures against cardiovascular diseases are essential [3], and predicting the occurrence of cardiovascular diseases can provide great theoretical significance and applied value for the treatment of clinical cardiovascular diseases [4]. At present, in the medical field, facing the identification, diagnosis and detection of complex problems such as cardiovascular diseases, the use of dimensionality reduction methods to deal with the relevant datasets is conducive to the elimination of research interfering information, which in turn helps to predict the relevant problems [5].

In order to better understand and predict the occurrence of cardiovascular diseases, various data-driven methods have been applied to data containing many complex features. Wang and others used the Join point regression model to analyse the average annual percentage change (AAPC) of cardiovascular disease (CVD) mortality in the elderly in China, and applied the GM(1,1) model to predict the crude rate of CVD deaths among the elderly in China from 2020 to 2030, pointing out that males are higher than females, and rural areas are higher than urban areas, and predicting that the crude rate of CVD deaths among the elderly in China will decline year by year, but still remain at a high level[3]. Liu used the use of survival analysis method of cardiovascular disease risk factors for single-factor analysis, and the use of Cox proportional risk model to carry out risk factors and the risk of morbidity of multifactorial analysis. At the same time, a prediction model for cardiovascular morbidity risk factors was established, pointing out that the risk factors in the problem of synergistic effect between each other, not only to pay attention to the number of risk, but also to pay attention to the combination between different risk factors [6]. Malovini et al. pointed out that because some traits are the result of a combination of rare and common variants and non-genetic factors, multivariate methods can help to analyse the genetic characteristics of cardiovascular diseases by compensating to some extent for informative genetic variants that had been overlooked by traditional univariate methods [7]. According to the Coronary Heart Disease (CHD) formula, Anderson K M and others showed that it is potentially important to control multiple risk factors instead of focusing one single risk factor in the studies related to cardiovascular disease. Parametric Model is better than some existing standard regression models. It can predict different time lengths and its probability expression is more direct than Cox proportional risk model [8]. Based on public health data, scholar Cui built a dynamic risk prediction model of cardiovascular disease using a longitudinal submodel connected by association structure and a Multivariate joint model of survival submodel, and obtained predictive indicators of the risk of cardiovascular disease in men and women. It was also pointed out that the area under the curve (AUC value) of the prediction model constructed by this model method is higher than that of the prediction model constructed by the multi-factor Cox proportional risk regression model method, confirming that this method is superior to the Cox proportional risk regression model [9]. VanDer et al. pointed out that the dimensionality reduction methods remove redundant and irrelevant features through feature extraction, and convert the original dataset containing high dimensional data (HDD) into a new dimensionality reduction dataset [10]. However, different dimensionality reduction techniques (DRTS) can produce significantly different results. Therefore, it is particularly important to choose the best dimensionality reduction method based on a given data set and analysis task. Ayesha S and others believed that the variation of results may come from differences in basic assumptions, mathematical formulas, and adaptability to data features [11].

Therefore, this study applies a variety of dimensionality reduction methods to the cardiovascular disease data set, and evaluates the results produced by these methods to interpret the reasons behind the observed differences in results, using the comparison of their AUC value, Accuracy Rate, Confusion Matrix and other indicators.

2. Methods

2.1. Data source and description

The data set used in this paper comes from KAGGLE website, with about 70,000 pieces of data. The variables and forms of the data set are shown in Table 1, including 11 independent variables and 1 target variable. The data set is of high quality and suitable for the subsequent research methods of this paper.

Table 1. Data description.

Variable	Symbol	Characteristic	Unit
Age	age	Objective Feature	days
Height	height	Objective Feature	cm
Weight	weight	Objective Feature	kg
Biological Gender	gender	Objective Feature	1-Female;2-Male
Systolic Blood Pressure	ap_hi	Examination Feature	mmHg
Diastolic Blood Pressure	ap_lo	Examination Feature	mmHg

Table 1. (continued).

Cholesterol	cholesterol	Examination Feature	1-normal 2-above normal 3-well above normal
Glucose	gluc	Examination Feature	1-normal 2-above normal 3-wel above normal
Smoking	smoke	Subjective Feature	0-no;1-yes
Alcohol Intake	alco	Subjective Feature	0-no;1-yes
Physical Activity	active	Subjective Feature	0-no;1-yes
Presence of CVDs	cardio	Target Variable	0-no;1-yes

2.2. Index selection and description

This study will first introduce the dimensionality reduction methods under consideration, including principal component analysis (PCA) in linear dimensionality reduction, linear discriminant analysis (LDA), and decision tree model. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss [12]. Linear discriminant analysis (LDA) is the most widely used supervised dimensionality reduction approach. After removing the null space of the total scatter matrix St via principal component analysis (PCA), the LDA algorithm can avoid the small sample size problem [13]. Decision tree classification algorithm is one of the most widely studied and applied topics in data mining, with high application value. However, this method may over-fit training data and be easily affected by noisy data, so the application of pruning method is very important [14].

2.3. Method introduction

In this study, the missing data and abnormal data will be reasonably filled and cleaned to meet the prerequisite requirements of different dimensionality reduction methods. Various dimensionality reduction techniques mentioned above will be applied to the pre-processed data set to produce dimensionality reduction results.

Secondly, this study will display the dataset qualitatively and quantitatively. Qualitative analysis includes data visualizations, such as boxplots and bar charts, to identify how is the data distributed. More in-depth quantitative analysis will be used to identify differences in capture variance and feature discrimination. In order to analyze the reasons for the observed differences in results, this study will conduct an in-depth investigation, including the inherent assumptions, mathematical foundations, and algorithmic procedures of each dimensionality reduction method, and reference the relevant literature to understand the nuances that contribute to the differences in results.

Finally, in order to quantitatively evaluate the performance of each dimensionality reduction method, this study adopts a series of indicators, including accuracy, confusion matrix, receiver operating characteristic curve (ROC) and the values of Area Under Curve (AUC). In the calculation of the above indicators, appropriate cross-validation strategies are adopted to ensure robustness and generalization.

By using the above methods, this study aims to comprehensively and systematically analyze the effects of various dimensionality reduction methods on the analysis of CVD data sets. By comparing results, exploring the causes of differences, and evaluating quantitative performance, it will help to deepen the understanding of the utility of different techniques in enhancing cardiovascular disease prediction.

3. Results and discussions

3.1. Data preprocessing

There are 70,000 data in the initial dataset. Figure 1 shows the situation of cardiovascular diseases of different genders. It can be seen that among the men involved, the number of people whether suffering from cardiovascular diseases are both about 23,000, and the number of women related to those are about 12,500. It can be preliminarily determined that gender has a small effect on the risk of cardiovascular disease.

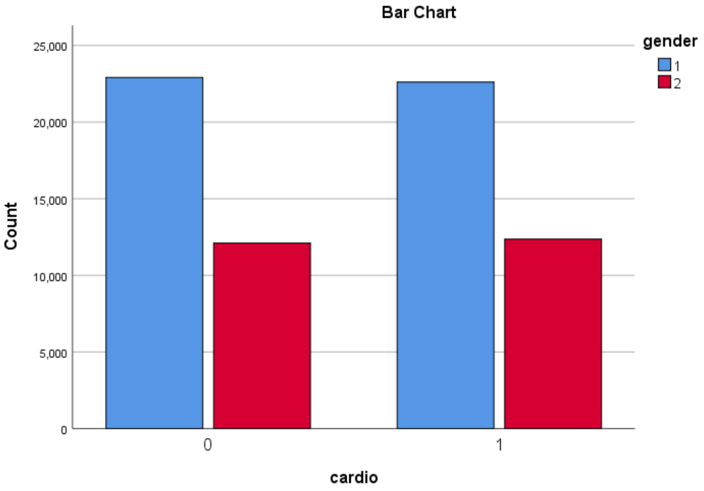

Figure 1. Gender bar chart.

As shown in Figure 2, the average age of people without cardiovascular disease is about 19,000 days, while the other is about 20,500 days, which shows that the latter is higher than the former. It can be preliminarily inferred that cardiovascular disease is more common in people with higher age.

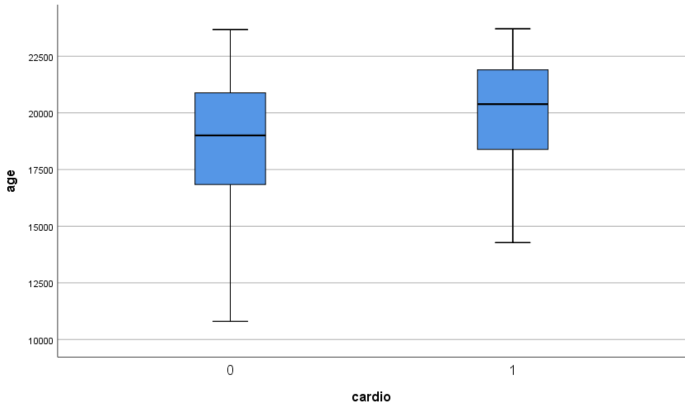

Figure 2. Age boxplots.

In this study, the values of age, height, weight, systolic blood pressure and diastolic blood pressure were standardized. Different normal value ranges were selected for different variables. Abnormal values were eliminated according to this range, leaving 66,753 valid data.

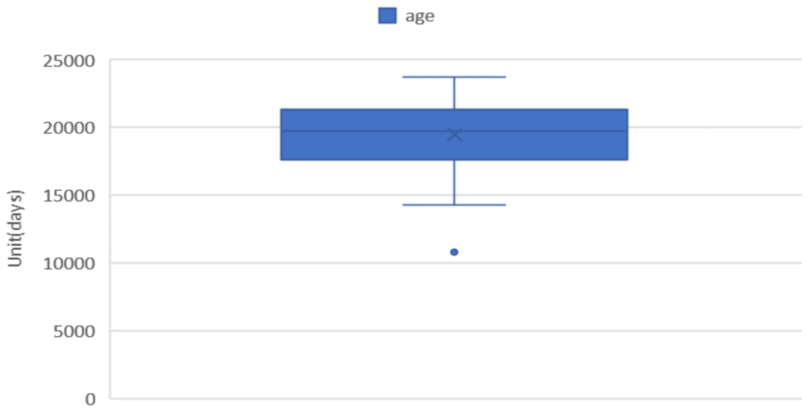

Figure 3. Adjusted age boxplots.

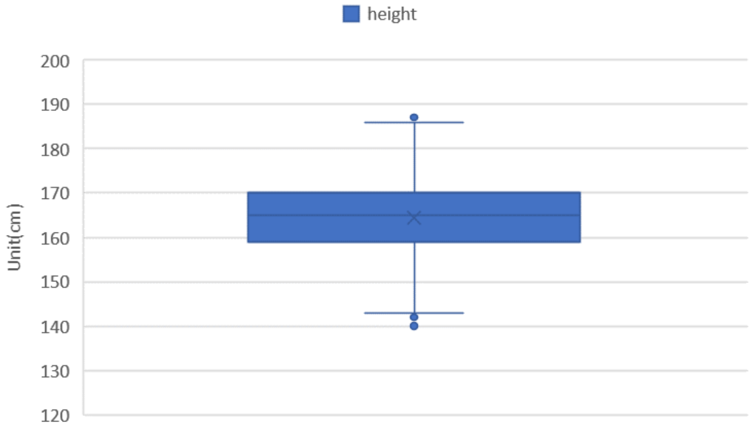

Figure 4. Adjusted height boxplots.

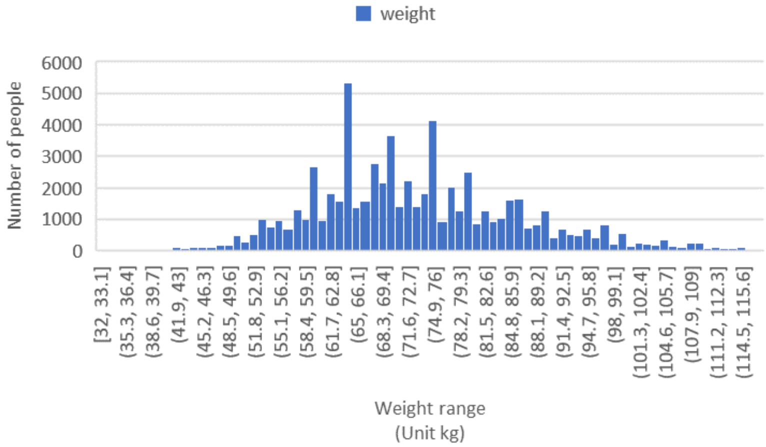

Figure 5. Adjusted weight histograms.

After adjustment, there is no extreme value greater than 3 times the box distance in the data set, but there is an outlier value greater than 1.5 times the box distance. Figure 3, 4 and 5 show the current characteristics of the new data set. For the values of age, height and weight, it can be concluded that these outliers can be retained. Therefore, there are no outliers that need to be processed, and this data set can be used to carry out follow-up research. Table 4 show the descriptive statistics of 11 independent variables and 1 dependent variable in the adjusted data set. It can be seen from Table 2 that compared with the data before adjustment, the variables such as systolic blood pressure and diastolic blood pressure in the processed data set are in the normal range, and the overall data quality is higher.

Table 2. Descriptive statistics.

	N	Minimum	Maximum	Mean	Std. Deviation
age	66753	10798	23713	19454.90	2469.135
gender	66753	1	2	1.35	.476
height	66753	140	188	164.38	7.675
weight	66753	32.00	117.00	73.4562	13.14647
ap_hi	66753	80	170	125.71	15.160
ap_lo	66753	50	111	80.94	8.913
cholesterol	66753	1	3	1.36	.674
gluc	66753	1	3	1.22	.567
smoke	66753	0	1	.09	.282
alco	66753	0	1	.05	.223
active	66753	0	1	.80	.397
cardio	66753	0	1	.49	.500
Valid N (listwise)	66753				

The data set was then divided into a training set and a test set in a ratio of 8:2.

3.2. Dimensionality reduction technique

3.2.1. PCA principal component analysis. PCA principal component analysis maps M-dimensional features to N-dimensional features, which are new orthogonal features, also known as principal components. It should be noted that each principal component is a linear combination of all the original variables, and each principal component is not related to each other. The principal components obtained by this method are more representative than the original independent variables.

After standardizing the data, the correlation matrix and total variance interpretation are obtained by calculating covariance, eigenvalue, information contribution rate and cumulative contribution rate. Then the principal component function is obtained by calculating the coefficients of each original variable standardized in each principal component, and then the principal component is determined. The weight of each principal component is obtained through the variance percentage of the sum of squares of loads extracted from the total variance interpretation table. After normalization, the equation for calculating the comprehensive score of each data can be obtained, and then the comprehensive evaluation score can be obtained.

Table 3. Total Variance Explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.199	19.990	19.990	2.199	19.990	19.990
2	1.883	17.120	37.110	1.883	17.120	37.110
3	1.305	11.862	48.973	1.305	11.862	48.973
4	1.114	10.129	59.102	1.114	10.129	59.102

Table 3. (continued).

5	.991	9.011	68.113	.991	9.011	68.113
6	.893	8.118	76.231	.893	8.118	76.231
7	.771	7.010	83.241	.771	7.010	83.241

Table 4. Component Matrix[a].

	Component						
	1	2	3	4	5	6	7
ap_hi	.732	-.355	-.367	.169	-.015	-.114	-.137
ap_lo	.713	-.319	-.405	.168	-.012	-.176	-.172
gender	.425	.652	.027	-.157	.033	.217	-.343
height	.393	.630	-.036	-.451	.099	.066	-.031
smoke	.307	.539	.236	.442	-.134	.035	-.156
gluc	.297	-.324	.702	-.194	.024	-.104	-.147
cholesterol	.401	-.397	.615	-.078	.030	-.074	-.061
alco	.250	.374	.266	.603	-.179	-.101	.388
active	.003	.029	.045	.282	.956	.059	.021
age	.280	-.361	.011	.075	-.081	.865	.157
weight	.580	.102	-.064	-.377	.092	-.133	.615

On the basis of having passed the significance test, the following is used to illustrate the parameters. Table 3 shows the extraction of the square sum of variance percentage of load, and the principal component analysis method is used to extract 7 principal components from the original 11 independent variables, and the proportion of these principal components in explaining the results is 83.241%. Meanwhile, Table 4 shows the factor load matrix. The following is an explanation of these 7 principal components, which are $Y_n (n = 1,2,\dots,7)$.

$$Y_1 = 0.494 \times \mathrm{Zap}_{\mathrm{hi}} + 0.481 \times \mathrm{Zap}_{\mathrm{lo}} + \cdots + 0.391 \times \mathrm{Zweight} \tag{1}$$

Y is the comprehensive score of each piece of data. The weight of each principal component is known by extracting the square of load and the percentage of variance, divided by the cumulative contribution rate. After normalization of these weights, the expression of the comprehensive evaluation score can be obtained.

$$Y = \frac{\begin{pmatrix} 0.19990 \times Y1 + 0.17120 \times Y2 + 0.11862 \times Y3 + 0.10129 \times Y4 \\ +0.09011 \times Y5 + 0.08118 \times Y6 + 0.07010 \times Y7 \end{pmatrix}}{0.83241} \tag{2}$$

Finally, let the new data set obtained by principal component analysis be dataset 1.

3.2.2. LDA linear discriminant analysis. When training samples, LDA linear discriminant analysis needs to find a straight line and project the training samples to a line that can make the projection points of the same type of samples as close as possible, while the projection points of different types of samples are as far away as possible. During the prediction, the data to be predicted is projected onto the line, the category is determined according to the position of the projection point, and the prediction effect is evaluated by the accuracy of the prediction, so as to test whether the current independent variable can explain the corresponding results, and then determine the valuable dimensionality reduction data set.

Table 5. Variables Entered/Removed

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	ap_hi	.811	1	1	53400.000	12454.790	1	53400.000	.000
2	age	.788	2	1	53400.000	7180.993	2	53399.000	.000
3	cholesterol	.773	3	1	53400.000	5227.104	3	53398.000	.000
4	weight	.770	4	1	53400.000	3983.362	4	53397.000	.000
5	active	.769	5	1	53400.000	3211.392	5	53396.000	.000
6	ap_lo	.768	6	1	53400.000	2684.717	6	53395.000	.000
7	smoke	.768	7	1	53400.000	2307.908	7	53394.000	.000
8	gluc	.767	8	1	53400.000	2023.961	8	53393.000	.000
9	height	.767	9	1	53400.000	1801.808	9	53392.000	.000
10	alco	.767	10	1	53400.000	1623.878	10	53391.000	.000

As can be seen from Table 5, an independent variable named 'gender' is eliminated by this method, and the significance of the remaining independent variables is 0.00, so there are 10 remaining factors. At the same time, the discriminant equation has passed the validity test, and its significance is 0.00, which can be considered that the discriminant ability of the equation is significant.

Table 6. Standardized Canonical Discriminant Function Coefficients.

	Function
	1
age	.314
height	-.042
weight	.143
ap_hi	.729
ap_lo	.078
cholesterol	.300
gluc	-.053
smoke	-.034
alco	-.040
active	-.087

A standardized typical discriminant equation expression can be obtained from Table 6, that is $\mathrm{Y} = 0.314\mathrm{x}_1 - 0.042\mathrm{x}_2 + 0.143\mathrm{x}_3 + 0.729\mathrm{x}_4 + 0.078\mathrm{x}_5 + 0.3\mathrm{x}_6 - 0.053\mathrm{x}_7 - 0.034\mathrm{x}_8 - 0.04\mathrm{x}_9 - 0.087\mathrm{x}_{10}$. Through this method, 72.6% of originally grouped cases were correctly classified. Finally, let the new dataset obtained by LDA linear discriminant analysis be dataset 2.

3.2.3. Decision tree model. The Chi square automatic interaction detection (CHAID) decision tree algorithm optimally segments the sample according to the given target variable and explanatory variable, performs automatic judgment grouping of multivariate contingences according to the significance of Chi-square test, and divides the dataset into the subset with the least impurity until the stopping criterion is met.

The original training data set was divided into the training set and the test set of the decision tree model at a ratio of 7:3 to verify the explanatory degree of the extracted new independent variable to the dependent variable. The growth method of CHAID can be used to quickly and efficiently mine the main influencing factors, that is, to select the best attribute and segmentation point and judge the priority of the attribute by calculating the Chi-square statistic between each variable and the target variable. The chi square statistic measures the correlation between the two variables, and the variable with a larger Chi-square value has a higher correlation with the target variable. After calculating the Chi-square value corresponding to the target attribute and each attribute, the optimal attribute is found according to the classification of the parent node, and the first level of the decision tree is obtained.

Table 7. Model Summary.

	Growing Method	CHAID
Specifications	Dependent Variable	cardio
	Independent Variables	age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active
	Validation	Split Sample
	Maximum Tree Depth	3
Results	Independent Variables Included	ap_hi, age, weight, cholesterol, active, gluc
	Number of Nodes	70
	Number of Terminal Nodes	46
	Depth	3

As can be seen from Table 7, the original 11 independent variables can be analyzed through the decision tree model to determine whether there is cardiovascular disease, and 6 independent variables (systolic blood pressure, age, weight, cholesterol, physical activity and glucose) can be analyzed, and the result is obtained, the depth of the decision tree model is 3.

Table 8. Risk.

Sample	Estimate	Std. Error
Training	.185	.001
Test	.188	.002

According to the risk table, as shown in Table 8, the training estimate is 0.185, and 18.5% of the 70% training samples are misclassified. The test estimate was 0.188, with 18.8% of the 30% tested samples misclassified. It can be seen that the decision tree model does not overfit when more than 80% of the correct rate is guaranteed, indicating that the model has certain significance. Finally, let the new dataset obtained from the decision tree model be dataset 3.

3.3. Comparison of dimensionality reduction methods

When the dimensionality reduction results obtained by principal component analysis are used for prediction, the output variable is the comprehensive score, which cannot be unified with the dependent variable ‘cardio’. Therefore, this study uses the dimensionality reduction data set obtained by principal component analysis to conduct logistic regression and obtain the predicted ‘cardio’ variable, which can be compared equally with other dimensionality reduction methods.

First of all, from the above, we can get the characteristics of the explanatory variables in the data set processed by each method and model. In the dimensionality reduction data set obtained by principal component analysis, there are 7 explanatory variables. After combining with the forward stepwise regression analysis of logistic regression, the variables are filtered, so there are 6 explanatory variables, which are linear combinations formed by the original independent variables. Second, when the forward stepwise regression is performed alone, the independent variable ‘gender’ is not included in the prediction equation. Furthermore, the independent variable ‘gender’ was also not included after the LDA linear discriminant analysis, which is the same as the data set obtained when the logistic regression was performed alone. Although LDA linear discriminant analysis and logistic regression are similar in form, their assumptions, parameter estimation methods and model meanings are different. Finally, the decision tree model is used to classify the six main independent variables, and the number of explanatory variables generated by this model is the least in the dataset obtained by the dimensionality reduction methods in question.

Secondly, in terms of fit degree, after principal component analysis combined with logistic regression, the overall percentage of correct prediction of both the training set and the test set was 72%. The accuracy of the LDA linear discriminant analysis training set is 72.6%, and its accuracy is also close to the accuracy of the training set when using the test set for verification, but the difference between them is slightly increased compared with the first method. A 7:3 ratio of training set and validation set was adopted when the decision tree model was constructed, and the misclassification ratio of the two are almost the same, 18.5% and 18.8% respectively. Therefore, neither the dimensionality reduction methods nor the prediction models have overfitting.

Then, the ROC curve and its area under the curve, AUC, were used to evaluate the prediction effects of the above methods and models. In this study, the following four dimensionality reduction methods and prediction models were used to compare with the actual value of the ‘cardio’ variable of the original training set. They were respectively principal component analysis method combined with forward stepwise regression, and directly carried out forward stepwise regression of logistic regression, LDA linear discriminant analysis method and decision tree model on the original training data set, and then drew the ROC curve. The AUC value is obtained by calculating the area under the curve.

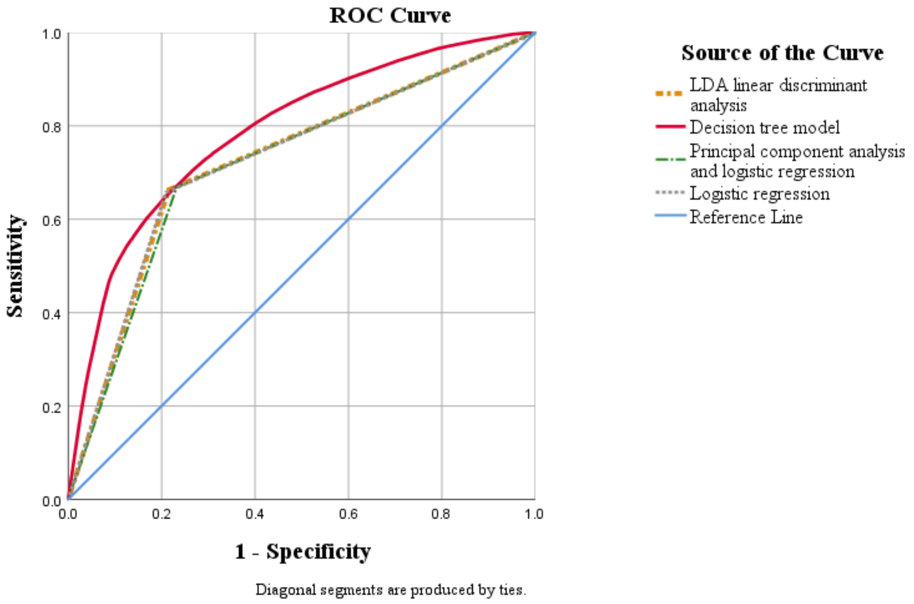

Figure 6. ROC Curve.

When the ROC curve as a whole is closer to the point (0,1), that is, the closer it is to the top left of the coordinate system, the corresponding method or model is better. Figure 6 shows the ROC curve. The intersection point of these four methods and the ROC curve of the model is roughly (0.22,0.65). Among them, the decision tree model has the best effect, and the other three kinds of prediction effect are very close. It is worth noting that the prediction effect of logistic regression alone is better than that of principal component analysis combined with logistic regression, and the difference between the two is not large. It also indirectly reflects the relationship of explanatory variables in the dimensionality reduction dataset obtained by LDA linear discriminant analysis and forward stepwise regression of logistic regression.

Table 9. Area Under the Curve.

Test Result Variable(s)	Area	Std. Errora	Sig.	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
LDA linear discriminant analysis	.724	.002	.000	.720	.728
Decision tree model	.788	.002	.000	.784	.792
Principal component analysis and logistic regression	.718	.002	.000	.714	.723
Logistic regression	.724	.002	.000	.720	.729

a.Null hypothesis: true area = 0.5

Table 9 describes the AUC indexes of the above four methods and models. In this study, the AUC values of LDA linear discriminant analysis, decision tree model, principal component analysis combined with forward stepwise regression and separately forward stepwise regression are 0.724, 0.788, 0.718 and 0.724 respectively. Therefore, the diagnosis accuracy of decision tree model is higher, followed by LDA linear discriminant analysis and forward stepwise regression, and the lowest is principal component analysis combined with forward stepwise regression, and the AUC values of the four are relatively close. It can be roughly inferred that after dimensionality reduction of principal component analysis, its accuracy is not as high as that of direct logistic regression prediction on the original training data set.

Finally, in terms of computational complexity, the combination of principal component analysis and forward stepwise regression has the highest processing complexity. This method not only does not reduce the number of raw variables to be collected, but also requires linear combination calculation of the original variables, and then it may be necessary to compare the results with other predictive models. LDA linear discriminant analysis and forward stepwise regression require 10 explanatory variables with moderate complexity. After the decision tree model is processed, only 6 explanatory variables are retained, which greatly reduces the indicators to be collected. Moreover, due to the reduction of dimensions, the number of samples to be collected can be reduced to a certain extent, and its complexity is moderate. Therefore, the decision tree model performs better in this respect.

To sum up, the performance of decision tree model is the best, followed by LDA linear discriminant analysis, but its fit degree is relatively low and unstable. Principal component analysis is relatively complex and its prediction performance is relatively weak.

4. Conclusion

For these three methods and models, there are no overfitting cases. Among them, decision tree model has higher accuracy and can greatly simplify the burden of data collection in terms of sample size and data dimension. The complexity of the LDA model is moderate, and the independent variable gender can be omitted in terms of sample requirements, and its accuracy is also high, but the fitting degree may be unstable, and its accuracy is relatively low. Although the principal component analysis method can reduce the data dimension, it also needs to collect the original data of the same dimension, and only after linear combination operation can the data set with lower dimension be obtained. Besides, its accuracy is relatively low, and it requires high manpower and material resources input in the early sampling, and it also has high complexity in the later calculation and analysis. Therefore, this study believes that LDA linear discriminant analysis or decision tree model can be used for data reduction and result prediction when dealing with cardiovascular disease or similar data sets and predicting related results, so as to obtain more efficient processing and more reliable prediction results.

Similarly, by analyzing the characteristics of the three dimensionality reduction datasets obtained, the present study concluded that age, blood pressure (systolic and diastolic), cholesterol, weight, physical activity, and glucose are seven important measures of cardiovascular disease. Therefore, for cardiovascular disease researchers, they can focus on analyzing and screening the above indicators to preliminarily predict and determine whether someone has the risk of cardiovascular disease. For the general public or healthy people, they can enhance their health awareness based on these seven indicators. For example, with the growth of age, the time interval of physical examination should be appropriately shortened, the possible cardiovascular health problems should be detected early, and the treatment time should be better. They can also control and monitor their own weight and blood pressure, as well as the intake of cholesterol and glucose, and carry out appropriate physical exercise. To help reduce your risk of cardiovascular disease.

This study also has the following deficiencies. First of all, this paper does not subdivide cardiovascular diseases, but only classifies all cardiovascular diseases and takes whether to suffer from cardiovascular diseases as its attribute value. The specificity and pertinence of these diseases need to be improved. Secondly, in order to compare the prediction accuracy, this study combined principal component analysis with logistic regression as a prediction model to compare with other independent methods or models, which may lead to errors caused by different comparison baselines. Therefore, when comparing methods and models, a more suitable uniform comparison baseline should be sought.

References

[1] Mc Namara K, Alzubaidi H and Jackson J K 2019 Cardiovascular disease as a leading cause of death: how are pharmacists getting involved. *Integrated pharmacy research and practice*, 1-11.

[2] Laslett L J, et al. 2012 The worldwide environment of cardiovascular disease: prevalence, diagnosis, therapy, and policy issues: a report from the American College of Cardiology. *Journal of the American College of Cardiology*, 60(25), 1-49.

[3] Wang X F, et al. 2023 Trend and prediction of cardiovascular disease mortality in the elderly in China from 2009 to 2019. *Modern Preventive Medicine*, 50(1), 39-45.

[4] Chen M M, Fang Z H, Tu W Y, et al. 2023 Construction and effect analysis of heart disease prediction model based on Logistic regression model. *Hospital Management Forum*, 39(02), 32-35.

[5] Liu C X, Shi D M and Song W J 2023 Research context and recent progress of dimensionality reduction methods for high dimensional data. *Journal of Statistics*, 4(03), 11-21.

[6] Liu M L 2017 Prospective study on risk factors and risk prediction model of cardiovascular disease in 35-64 years old population in China. *China Health Industry*, 2.

[7] Malovini A, Bellazzi R, Napolitano C, et al. 2016 Multivariate methods for genetic variants selection and risk prediction in cardiovascular diseases. *Frontiers in cardiovascular medicine*, 3, 17.

[8] Anderson K M, Odell P M, Wilson P W F, et al. 1991 Cardiovascular disease risk profiles. *American heart journal*, 121(1), 293-298.

[9] Cui Z Z 2023 Research on dynamic risk prediction model of cardiovascular disease based on public health data. *Suzhou University*.

[10] Van Der Maaten L, Postma E O and Van Den Herik H J 2009 Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66), 13.

[11] Ayesha S, Hanif M K and Talib R 2020 Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59, 44-58.

[12] Jolliffe I T and Cadima J 2016 Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

[13] Zhao X, Guo J, Nie F, et al. 2020 Joint principal component and discriminant analysis for dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2), 433-444.

[14] Zhang Y 2010 Research on Decision Tree Classification and Pruning Algorithm. *Harbin University of Science and Technology*.