Diabetes Prediction Models Based on Machine Learning

Liu Bobo[1], Kang Xiaofei[1], Zhang Zongyue[2], Hou Liying[2,*], Wang Yidan[1,*]
[1]College of Science, North China University of Technology, Tangshan, Hebei, China
[2]School of Public Health, North China University of Technology, Tangshan, Hebei, China
**Corresponding Author*

Abstract: Objective to compare the predictive efficacy of random forest, BP neural network, gradient boosting tree and plain Bayesian models for the prevalence of diabetes. Practical application: by measuring the basic indicators such as individual height, weight, triglyceride, etc., the model can be used to predict the probability of individual disease, and then targeted to improve some indicators of the body, to achieve the effect of diabetes prevention intervention, and to provide new ideas for diabetes prevention research. Methods Using the 2009 survey data from the China Health and Nutrition Survey (CHNS), the data for men and women were statistically analyzed by dividing them into four groups according to the visceral fat index (VAI). Subsequently, the processed samples were divided into training sets and test sets by 4:1, and four machine learning models, namely, random forest, BP neural network, gradient lifting tree, and naive Bayes, were constructed. The experiment was conducted using a five-fold cross validation method, and the prediction effect was evaluated through indicators such as sensitivity, accuracy, and AUC. Results One-way ANOVA showed that the differences in height, weight, waist circumference, triglycerides, high-density lipoprotein cholesterol, body mass index, fasting blood glucose, and glycosylated hemoglobin among different VAI quartile groups were statistically significant (P<0.05). Comparison of prediction effects of four models: sensitivity 75.75%, 90.77%, 76.31%, 98.57%, accuracy 74.80%, 87.82%, 74.64%, 92.00%, AUC 0.713, 0.716, 0.668, 0.676, and Jorden index 0.34, 0.27, 0.22 and 0.21. Conclusion Based on the CHNS 2009 survey data, the BP neural network model has a better effect and stability in predicting diabetes.

Keywords: Visceral Adiposity Index; Diabetes; Random Forest; BP Neural Network; Gradien Boosting Decision Tree; Naive Bayes Model

1. Related Work

Diabetes is one of the top 10 causes of death worldwide and one of the most challenging public health issues [1], posing a considerable burden to the global economy and public health management. the risk of all-cause mortality in diabetes is 2-3 times higher than in non-diabetic patients [2]. In December 2021, the International Diabetes Federation released the 10th edition of the IDF Diabetes Atlas, stating that in 2021, there will be 537 million people aged 20-79 with diabetes, accounting for 10.5% of the global population in this age group [3]. As the largest developing country in the world, China's growing economy and the lack of attention to lifestyle and dietary habits have made China one of the countries with the most significant number of people with diabetes and the fastest-growing number of patients in the world [4], reaching 140 million patients with diabetes in 2021 [5]. Chapter 7 of the Health China 2030 Plan, "Strengthening public health services for the whole population", mentions the need to cover diabetes management interventions by 2030 fully. Therefore, searching for simple and easy indicators to predict the risk of diabetes and more accurate and effective prediction models is very important in preventing and controlling diabetes for the whole population.

Diabetes is a non-communicable disease caused by genetic and environmental factors [6], and obesity is the most crucial factor. Obesity may be caused by an unbalanced diet structure, excessive total calorie intake [7], and body mass index (BMI), waist-to-hip ratio (WHR), waist circumference (WC) and other indicators reflecting body obesity are mainly used to predict diabetes. However, it has been found that the visceral fat index (VAI) can better reflect the

degree of body obesity compared with traditional indicators such as BMI and WC [8], at the same time, it has a certain correlation with the blood sugar control of patients [9], which may be able to better prevent diabetes.

This article first introduces the background of diabetes. Section 1 lists the existing work related to VAI research on diabetes. Section 2 lists the shortcomings of existing work and the research innovation of this article. Section 3 introduces experimental methods, including data sources, research subjects and methods, statistical methods, and four machine learning methods. Section 4 lists the experimental results and conducts comparative analysis. Section 5 summarizes the entire text and proposes shortcomings and future work.

At present, the relationship between VAI and the prevalence of diabetes is often studied by using logistic regression models in clinical practice. For example, Yue Furong explored the correlation between VAI and diabetes through logistic regression analysis. the results show that VAI can effectively identify the early risk of diabetes in middle-aged and elderly people, and the prediction is better than BMI and WHtR [10]. Also, Miao Ying analyzed the impact of VAI on diabetes in the early stage of diabetes through single factor and multi factor logistic regression analysis, the results showed that VAI was a risk factor affecting the prognosis of pre diabetes subjects to diabetes, and with the increase of VAI, the prognosis risk would continue to increase [11].

2. Insufficient Previous Work and Innovative Research in This Article

Although the logistic regression model has the characteristics of fast training speed and easy to understand, it requires high data integrity and has certain limitations. While machine learning as a favourable tool to promote the era of arithmetic, such as Random Forest(RF), BP Neural Network(BPNN), Gradient Boosting Decision Tree(GBDT), and Naive Bayes Model(NBM) have significant advantages over traditional statistical models in data processing, for example, they are more inclusive of data, can reduce errors caused by missing data, and improve the accuracy of prediction. There are fewer reports on applying these four methods in diabetes, and there is a lack of relevant studies on the prediction of diabetes by these four methods. This study compared the predictive efficacy of random forest, BP neural network, gradient boosting tree, and naive Bayes model for diabetes regarding sensitivity, specificity, accuracy, AUC, and Jorden's index [12-14], so as to provide new ideas for the intervention and prevention of diabetes.

3. Experimental Methods

3.1 Data Source

All data in this study were obtained from the China Health and Nutrition Survey(CHNS), which aims to study the impact of social and economic transformation on the nutrition and health status of the Chinese population. Since 1989, nine data collections have been conducted in 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009 and 2011. Since blood lipids, a core indicator of VAI were only present in the 2009 collection, this paper's cross-sectional data from the 2009 CHNS were used as the baseline information.

3.2 Research Subjects and Methods

A total of 11, 929 healthy participants participated in the baseline interview in the 2009 survey, which included socioeconomic characteristics, lifestyle, general health status, medical history, physical measurements, and blood samples.

Inclusion criteria: diabetes diagnosis criteria were fasting blood glucose greater than or equal to 7.0 mmol/L and/or self-reported history of diabetes in the subject and/or use of insulin or oral hypoglycemic agents within two weeks [11], or glycosylated haemoglobin $\geq 6.5\%$ [15] could also be used as diagnostic criteria. Exclusion criteria: Exclusion of patients with unmeasured weight, height, waist circumference, triglyceride(TG) and high-density lipoprotein cholesterol(HDL-C); exclusion of patients with combined heart, liver and kidney diseases, malignant neoplasm diseases and pregnancy; exclusion of patients with mental illness or poor compliance. At the same time, for cases with severe missing data in the same sample, elimination was used to obtain baseline data for 9088 subjects. The sample size is relatively small, and the number of diabetes patients is also small (903 cases), which may affect the reliability and generalization of the results. Therefore, this paper uses the SMOTE method to expand the cases. We extracted the weight, height, waist circumference, triglyceride, high-density lipoprotein cholesterol, fasting

glucose, and glycated haemoglobin from the data and calculated the VAI accordingly. We constructed the classification models of RF, BPNN, GBDT, and NBM with VAI as the independent variable and diabetes as the dependent variable and compared the prediction effects of each model.

3.3 Research Methodology

3.3.1 Vai Indicator Calculation

The formula for calculating the visceral adiposity index was determined by a search of the relevant literature [11] and is given below.

$$VAI\,(Male) = \left(\frac{WC}{39.68 + 1.88 \times BMI}\right) \times \left(\frac{TG}{1.03}\right) \times \left(\frac{1.31}{HDL - C}\right) \quad (1)$$

$$VAI\,(Female) = \left(\frac{WC}{36.58 + 1.89 \times BMI}\right) \times \left(\frac{TG}{0.81}\right) \times \left(\frac{1.52}{HDL - C}\right) \quad (2)$$

$$BMI = Weight\ /\ Height^2 \quad (3)$$

In this study, the VAI calculated from the above formula was used to group the study subjects according to the quartiles of VAI.

3.3.2 Statistical Analysis

SPSS 22.0 was used for statistical analysis. Normality measures were expressed as $\bar{x} \pm s$, and comparisons between two groups were analysed by t-test. Non-normality measures were expressed as $M(P_{25}, P_{75})$, and the rank sum test was used to compare the two groups. the test standard $\alpha = 0.05$.

3.3.3 Machine Learning Methods

The RF model is a widely used and applicable machine learning algorithm. RF constructs multiple decision trees and then combines the voting results of each tree to predict the results. This voting mechanism combining a large number of decision trees can quickly generate training models with less human intervention and good robustness, thus improving prediction accuracy. the construction of RF mainly lies in the generation of decision trees. the training and test samples of the decision tree are extracted using a bootstrap aggregation algorithm, and the main rules used to split the data in the classification problem are the Gini coefficient, variance, etc. [16].

The BPNN model consists of an input layer, an output layer, and one or more implicit layers in between, where the neurons in each layer are connected only to the neurons in the adjacent layers. the information is transferred between the layers through different weights and thresholds. the training process of the BPNN algorithm consists of two parts: the forward transmission of information and the reverse transmission of error. the gradient descent technique minimises the mean squared error between the actual and desired outputs of the neural network. In the model's training process, the error decreases along the gradient direction by adjusting the strength of the connection between the input nodes and the hidden layer nodes and the strength of the connection between the hidden layer nodes and the output nodes as the threshold value. After repeated training to determine the network parameters corresponding to the minimum error to reach the desired error or set the number of learning times, the learning is terminated to obtain the trained prediction model [17].

The GBDT model was developed to guarantee classification or regression by continuously reducing the learning error rate arising from the training process and to face the imbalance in the type of actual production data. the algorithm has high prediction accuracy and can handle consistent and discrete forms of data [18].

The NBM model is widely used in data mining and machine learning, which first calculates the conditional probability that the sample to be judged belongs to each class based on Bayes' theorem and the assumption of conditional independence of attributes and then determines it as the class with the highest probability [19], which also has a relatively small computational overhead in the process of model training and data classification, and has the characteristics of simplicity and efficiency.

The process of this study can be referred to in the following flow chart.

The data were first screened by exclusion and inclusion criteria, followed by descriptive statistics and case expansion, and finally, four machine learning models for diabetes prediction were built.

4. Experimental Results and Analysis

4.1 Characteristics of the Interviewees

General demographic characteristics and clinical data. A total of 9088 study subjects were included in this study, including 4288(47.18%) males and 4800(52.82%) females; 8185(90.06%) had no diabetes, and 903(9.94%) subjects had

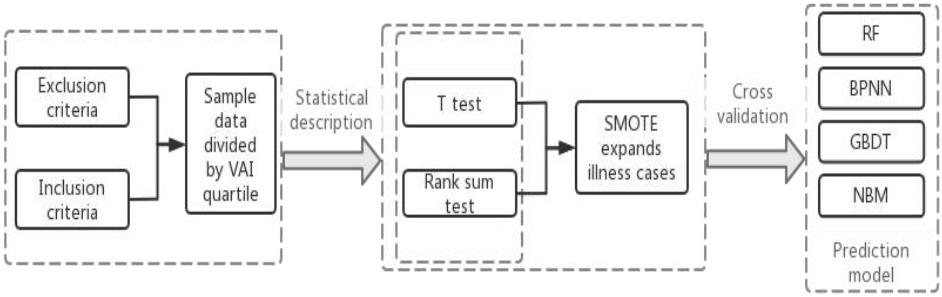

Figure 1. Overall idea

Table 1. Demographic characteristics of men based on visceral fat index (VAI) quartile score

Variables	Q1(< 0.676)	Q2(0.677 $- 1.183$)	Q3(1.184 $- 2.174$)	Q4(≥ 2.175)	Test statistic	P
Height	161.09 $\pm$ 13.21	165.12 $\pm$ 9.35	166.94 $\pm$ 7.49	167.46 $\pm$ 7.26	96.527	< 0.00
Weight	53.21 $\pm$ 13.37	60.53 $\pm$ 11.41	66.16 $\pm$ 11.56	71.43 $\pm$ 11.56	453.224	< 0.00
Waist Circumference	73.72 $\pm$ 10.90	80.36 $\pm$ 10.03	85.4 $\pm$ 10.15	90.2 $\pm$ 9.44	518.257	< 0.00
Triglyceride	0.67 $\pm$ 0.26	1.06 ± 0.27	1.59 ± 0.45	3.61 $\pm$ 2.39	1216.599	< 0.00
High-density lipoprotein cholesterol	1.76 $\pm$ 0.74	1.46 ± 0.31	1.28 ± 0.27	1.07 $\pm$ 0.24	473.746	< 0.00
BMI	20.13 $\pm$ 3.17	22.05 $\pm$ 3.12	23.63 $\pm$ 3.30	25.39 $\pm$ 3.41	510.296	< 0.00
Fasting blood sugar	5.02 $\pm$ 0.82	5.19 ± 1.22	5.37 ± 1.51	5.93 $\pm$ 2.17	74.688	< 0.00
Glycated haemoglobin	5.41 $\pm$ 0.54	5.54 ± 0.83	5.63 ± 0.89	5.89 $\pm$ 1.21	54.407	< 0.00

diabetes. the study subjects were divided into four groups according to VAI quartiles according to gender, and the data comparison results among the groups are shown in Tables 1 and 2. the results of one-way ANOVA revealed statistically significant differences between height, weight, waist circumference, triglycerides, high-density lipoprotein cholesterol, BMI, fasting glucose, and glycosylated haemoglobin of the study subjects ($P < 0.05$).

4.2 Comparison of Four Models for Predicting Diabetes

There were 8185 cases without diabetes mellitus (90.06%) and 903 cases with diabetes mellitus (9.94%). the sample with and without disease was highly unbalanced, and the sample was expanded by SMOTE method [20] to 4583 cases with the disease.

The analysis was performed with VAI as the independent variable and whether or not they had diabetes as the dependent variable, and the variable assignments are shown in Table 3.

The sample set was divided into a training set and a test set by 4:1 and a five-fold cross-validation method was used for the test set of the samples, as shown in Figure 2. the test set of the samples was predicted using the RF, BPNN, GBDT, and NBM classification models. the sensitivity, specificity, accuracy, Jorden index, and AUC values were compared and analysed in five aspects, as shown in Table 4.

Based on the analysis of the available data, it was concluded that the sensitivity of the neural network classification model was 90.77%, which was more sensitive and had a higher predictive

Table 2. Female Demographic Characteristics Based on Visceral Fat Index (Vai) Quartile Score

Variables	Q1(< 0.934)	Q2(0.95 $- 1.544$)	Q3(1.545 $- 2.684$)	Q4(≥ 2.685)	Test statistic	P
Height	154.17 $\pm$ 9.57	155.28 $\pm$ 7.13	155.57 $\pm$ 6.97	155.69 $\pm$ 6.54	9.921	< 0.001
Weight	50.02 $\pm$ 10.36	53.71 $\pm$ 9.87	57.51 $\pm$ 10.15	61.29 $\pm$ 9.86	279.624	< 0.001
Waist Circumference	72.88 $\pm$ 10.04	77.69 $\pm$ 10.03	82.19 $\pm$ 10.06	86.98 $\pm$ 9.99	436.105	< 0.001
Triglyceride	0.67 $\pm$ 0.19	1.03 $\pm$ 0.24	1.47 $\pm$ 0.35	2.93 $\pm$ 1.65	1613.667	< 0.001
High-density lipoprotein cholesterol	1.79 $\pm$ 0.67	1.53 $\pm$ 0.29	1.4 $\pm$ 0.28	1.18 $\pm$ 0.26	471.313	< 0.001
BMI	20.85 $\pm$ 3.18	22.18 $\pm$ 3.36	23.70 $\pm$ 3.59	25.24 $\pm$ 3.52	370.902	< 0.001
Fasting blood sugar	5.03 $\pm$ 0.93	5.06 $\pm$ 0.82	5.29 $\pm$ 1.22	5.68 $\pm$ 1.79	70.15	< 0.001
Glycated haemoglobin	5.40 $\pm$ 0.61	5.48 $\pm$ 0.95	5.62 $\pm$ 0.80	5.84 $\pm$ 1.03	61.024	< 0.001

Table 3. Classification Assignment Table

Variables	Assignment
Fasting blood glucose or Glycated haemoglobin	$< 7.0\text{mmol/L} = 0,\ \geq 7.0\text{mmol/L} = 1$ or $< 6.5\% = 0,\ \geq 6.5\% = 1$

Figure 2. Illustration of 50% Discount Cross-Validation

Table 4. Comparative Analysis Results of Models

Parameters	RF	BPNN	GBDT	NBM
Sensitivity (%)	75.75%	90.77%	76.31%	98.57%
Specificity(%)	58.39%	37.23%	45.99%	25.55%
Accuracy rate (%)	74.80%	87.82%	74.64%	92.00%
AUC	0.713	0.716	0.668	0.676
Yoden Index	0.34	0.27	0.22	0.21

value for the actual positive rate of diabetes classification. the accuracy of the neural network and the plain Bayesian classification models were 87.82% and 92.00%, respectively, significantly higher than the other models. the sensitivity and accuracy of the neural network model were not significantly different from the plain Bayesian classification model, while the AUC value of the neural network model was 0.716. the AUC value of the plain Bayesian model was 0.676, and the former was higher than the latter, see Figure 3.

Therefore, the BP neural network model has a better predictive effect and stability in diabetes.

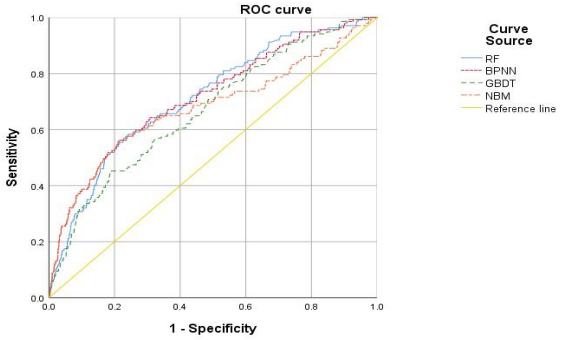

Figure 3. Comparison of ROC Curves of Four Models

5. Summary

By analyzing the relationship between visceral fat index and diabetes, this paper proposes a diabetes prediction model based on random forest, BP neural network, gradient lifting tree and naive Bayes, and obtains the optimal prediction model through comparative analysis. First, by observing the general demography characteristics of men and women, it is found that there are statistically significant differences in height, weight, waist circumference, triglycerides, high-density lipoprotein cholesterol, body mass index, and glycosylated hemoglobin in different VAI quartile groups. Then four prediction models of diabetes were constructed. Conclusion: Based on the open data of CHNS in 2009, it was found that the BP neural network model had good prediction effect and stability in diabetes. the traditional diagnostic standard of diabetes needs to measure the fasting blood sugar or glycosylated hemoglobin of individuals to determine whether they have diabetes. This study can measure some basic indicators, including height, weight, triglyceride, etc., and then combined with the BP neural network model, it can predict the probability of individuals suffering from diabetes, and then targeted to improve some indicators of the body to achieve a preventive intervention effect on diabetes, It provides certain value for the prevention of diabetes.

Some shortcomings and shortcomings of this paper.

1) This article uses the SMOTE method to expand disease cases. Although it effectively eliminates the impact of imbalanced disease proportions, this method also has certain shortcomings, such as the authenticity and representativeness of the expanded sample may be affected, which affects the reliability of the research results.

2) The analysis of the stability and application value of the model is not sufficient, and there is no discussion on the challenges and limitations that may be faced in the specific application of the model. In addition, only data from 2009 was used for the study, and further verification is needed to determine whether it will be affected by social, environmental and other factors in the future.

3) The selection of samples and data processing methods were not fully explained and discussed, and other factors that may affect the prediction effect, such as lifestyle, genetics, etc., were not explored.

4) In this paper, four models for predicting diabetes based on machine learning are discussed and compared. However, it is not clearly pointed out what shortcomings exist in the comparison of the four models for diabetes prediction. It may be necessary to further explore the limitations and improvements of these models in practical application.

5) This paper does not explore the limitations and limitations of visceral fat index itself, which may affect the accuracy and applicability of the model.

The future work of this article includes: firstly, further exploring the authenticity of predictions before and after case expansion and conducting specific analysis; Second, further study other factors that may affect the prediction of diabetes except visceral fat index to achieve more accurate prediction; Third, combining the algorithm itself and the reality of diabetes, we developed an algorithm improved diabetes prediction model, making the model more stable and practical, and providing a more practical plan for diabetes prevention.

Acknowledgments

Fund programs: Ministry of Education Supply and Demand Matching Employment Education Project (20220105939); Innovation and Entrepreneurship Training Program for University Students of North China University of Technology(R2022120).

References

[1] Lin XL. Analysis of global, regional and national disease burden of diabetes from 1990 to 2017 [D]. Zhejiang University, 2020. DOI:10.27461/d. cnki. gzjdx. 2020.003332.

[2] Yang JJ, Yu D, Wen W, et al. Association of Diabetes With All-Cause and

Cause-Specific Mortality in Asia: A Pooled Analysis of More Than 1 Million Participants. JAMA Netw Open. 2019; 2(4):e192696. Published 2019 Apr 5. doi:10.1001/jamanetworkopen. 2019.2696

[3] Zhang LW, Ruan MH, Liu JL, et al. Analysis of research and development trend in the field of diabetes [J/OL]. HEREDITY:1-26[2022-10-21]. DOI:10.16288/j. yczz. 22-272.

[4] Ma SL, Xv YJ, Meng RL, et al. Effects of blood pressure and overweight/obesity on diabetes in Guangdong residents aged 40 and above [J]. Chinese Journal of Disease Control, 2022, 26(04):397-400+429. DOI:10.16462/j. cnki. zhjbkz. 2022.04.006.

[5] Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract. 2022; 183:109119. DOI:10.1016/j. diabres. 2021.109119

[6] Yv C, Wang MZ, Jin YF, et al. Influence of interaction between family history and obesity on the incidence of diabetes in pre diabetes population [J]. Chinese Journal of Disease Control, 2020, 24(09):997-1002. DOI:10.16462/j. cnki. zhjbkz. 2020.09.002.

[7] Zou DJ, Zhang Z, Ji LN. Chinese Expert Consensus on Alleviating Type 2 diabetes [J]. China General Practice, 2021, 24(32):4037-4048.

[8] AMATO M C, GIORDANO C, GALIA M, etal. Visceral adiposity index:a reliable indicator of visceral fat function associated with cardiometabolic risk [J]. Diabetes Care, 2010, 33(4):920-922. DOI:10.2337/dc09-1825.

[9] Cao YY, Tang X, Sun KX, et al. Relationship between glycemic control and visceral adiposity index in patients with type 2 diabetes mellitus [J]. Journal of Peking University (Medical Edition), 2017, 49(03):446-450.

[10] Yue FR, Tian YH. Visceral fat index and its predictive value in middle-aged and elderly patients with diabetes [J]. Ningxia Med J, 2021, 43(10):923-926+864. DOI:10.13621/j. 1001-5949.2021.10.0923.

[11] Miao Y, Chen P, Yan PJ, et al. Study on the correlation between visceral fat index and the prognosis of pre diabetes patients to diabetes [J]. Journal of the Third Military Medical University, 2020, 42(21):2154-2161. DOI:10.16016/j. 1000-5404.2202006124.

[12] Deng LF, Liang J, Lu D, et al. Analysis and comparison of risk prediction models for intractable postpartum urinary retention constructed by three statistical methods [J]. Journal of Guangxi Medical University, 2022, 39(09):1442-1447. DOI:10.16190/j. cnki. 45-1211/r. 2022.09.015.

[13] Li LW, Huang Q, Shi JC, et al. Analysis and Comparison of Hypertension Incidence Prediction Models for Overweight and Obese People Based on Three Statistical Methods [J]. Modern Preventive Medicine, 2021, 48(11):2061-2066.

[14] Liu YH, Song J, Li MJ, et al. Study on the Classification Model of Hypoproliferative Myelodysplastic Syndrome and Aplastic Anemia Based on Data Mining [J]. Modern Preventive Medicine, 2021, 48(17):3254-3258.

[15] Mei Z. HbA1c included in the diagnostic criteria of diabetes [J]. Jiangsu Health Care, 2021(07):50.

[16] Sun J, Fan M, Cui XD, et al. A Multi beam Seabed Sediment Classification Method Based on the Combination of ReliefF and Stochastic Forest Model [J]. Marine Science Bulletin, 2022, 41(02):131-139.

[17] Yu F, Wang KJ, Zhang WL, et al. Prediction of coagulant dosage for in-situ turbidity control in water ecological restoration based on genetic algorithm optimized BP neural network [J/OL]. Journal of Environmental Engineering:1-12[2022-10-14].

[18] Ma LF, Xiao HM, Tao JW, et al. Intelligent classification of lithology based on gradient lifting decision tree algorithm [J]. Petroleum Geology and Recovery Efficiency, 2022, 29(01):21-29. DOI:10.13673/j. cnki. cn37-1359/te. 2022.01.003.

[19] Li SQ, Lv WY, Deng X, et al. Naive Bayesian Classification Algorithm Based on Improved PCA [J]. Statistics and decision-making, 2022, 38(01):34-37. DOI:10.13546/j. cnki. tjyjc. 2022.01.007.

[20] Tong R, Kan LH, Zhu ZS. Prognostic Modeling of Heart Failure Based on Logistic Regression and Random Forest [J/OL]. Journal of Fudan University (Medical Edition):1-9[2022-10-14].

