Research on Diabetes Prediction Based on Multiple machine Learning Methods

Hui Yu *

School of Mathematics and Computer Science, Gannan Normal University, Ganzhou, Jiangxi, China

* Corresponding Author Email: 220700092@gnnu.edu.cn

Abstract. Diabetes falls within the category of chronic diseases, and the issue of its prevention and control has always been a health concern for all mankind. This study constructs a prediction model for diabetes based on four machine learning methods: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). In the distribution of important features in LR, the top five important features are selected. On this basis, a model is constructed through other algorithms, aiming to reduce the interference of other irrelevant features, and then compare the performance of models constructed by different machine learning algorithms. The dataset of this study contains 768 samples, covering 8 characteristics including metrics like the number of pregnancies and plasma glucose concentration. Finally, the accuracy rates of the models constructed by the four algorithms of LR, DT, RF, and XGBoost were 75.97%, 78.57%, 74.03% and 74.68% respectively. Combining the area under the receiver operating characteristic curve (AUC), as well as the precision rate, recall rate, and F1 score of the positive category, it is ultimately concluded that the predictive effect of DT is the best. However, to better integrate diabetes prediction models with practical applications, more data and resources are still needed to support them.

Keywords: Diabetes; Logistic Regression; Decision Tree; Random Forest; Extreme Gradient Boosting.

1. Introduction

Diabetes has always been a health issue of great concern to all sectors of society. Diabetes is a chronic disease. It occurs when insulin fails to function properly or is insufficient. If not detected and treated in time, it can also easily lead to other complications [1]. The latest data from the World Health Organization shows that the number of diabetes patients worldwide was 830 million in 2022, more than four times the 200 million in 1990. Moreover, 14% of 18-year-olds suffered from diabetes in 2022, which was twice the number in 1990 [2]. If potential patients can be identified through early diabetes prediction, diabetes can be treated and relieved in a timely manner, and the risk of other complications can also be reduced [3].

With the vigorous development of artificial intelligence, the application of machine learning methods in the field of disease prediction has gradually matured. The area under the receiver operating characteristic curve (AUC) of the diabetes prediction model obtained by scholars such as Ou through logistic Regression (LR) and lightweight gradient boosting was 0.906 and 0.910, respectively, indicating that the prediction effect of the model was good [4]. To further enhance the predictive power of the model and the accuracy of the test, scholars such as Zhao employed machine learning methods to construct multiple ensemble models. The model built based on the stacked ensemble method had the highest accuracy rate, reaching 93.83% [5]. In addition, Xiang also constructed three different diabetes prediction models in machine learning, analyzed and compared the accuracy of the three different algorithms, and finally concluded that the model constructed by Decision Tree (DT) had the best evaluation performance [6]. The above-mentioned research demonstrates the outstanding application of machine learning techniques for diabetes prediction.

This paper constructs a diabetes forecasting model utilizing four distinct algorithmic approaches: LR, DT, Random Forest (RF), and XGBoost. The top five important features are selected from the distribution of important features in LR. On this basis, the model is constructed through other algorithms, and the accuracy of the four different algorithms is compared. This study aims to explore

the advantages and disadvantages of different machine learning methods in diabetes prediction, providing certain references for the early prediction and clinical trials of diabetes.

2. Research Methods

2.1. Data Source and Introduction

The data for the diabetes prediction model is sourced from the Kaggle platform. This dataset contains 768 samples, covering eight related indicators: the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function, and age of the patient [7].

2.2. Data Preprocessing

Due to the large sample size of the dataset, there are multiple outliers in the dataset, and the dimensions of different features are also different. Therefore, in the data preprocessing stage of this study, the data were cleaned and standardized. The mean was used to replace the outliers, and the data of each feature was converted into a standard normal distribution with a mean of 0 and a standard deviation of 1, eliminating the influence of different feature dimensions. Subsequently, the dataset was split into features and target variables. To better evaluate the model's performance, this study segmented the data. The dataset was stochastically divided into a training set and a test set in an 8:2 ratio. To analyze the correlations among different features, this paper calculates the Pearson correlation coefficient of each feature to obtain the heat map of the feature correlation matrix, as shown in Fig. 1.

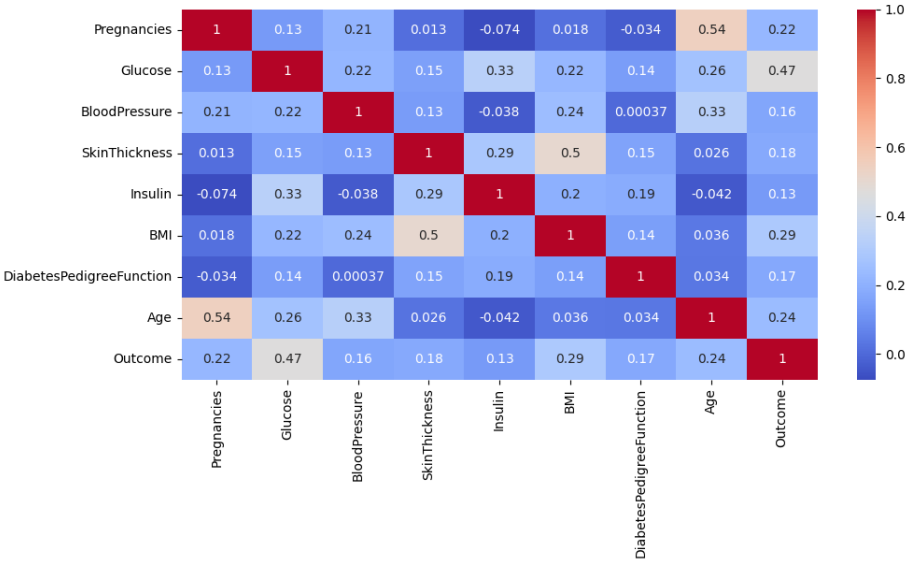


Fig 1. Feature correlation matrix heat map (Original)

As can be seen from Fig. 1, there is a strong correlation among some features. For instance, the correlation index between BMI and triceps skinfold thickness is 0.5, and the correlation index between age and the number of pregnancies is 0.54. Among the eight related features, plasma glucose concentration has the greatest impact on whether one has diabetes.

2.3. Logistic Regression

Logistic regression (LR) is a supervised learning algorithm for solving binary classification problems and is widely used in scenarios such as advertising click-through rate prediction and disease risk prediction. It can explain the importance of features and has the characteristics of high computational efficiency and good stability. Therefore, this study also applies LR to the diabetes prediction model. LR maps linear combinations to probability values between 0 and 1 through an S-

shaped curve function to predict the possibility of developing diabetes. The mathematical expression of the S-shaped curve function is as follows:

$$p(x)=\frac{1}{1+e^{-(\alpha_1 x_1+\alpha_2 x_2+...+\alpha_8 x_8)}} \quad (1)$$

Among them, $p(x)$ is the probability of developing diabetes under the feature x, with a value ranging from 0 to 1, x is the feature related to diabetes, and $\alpha_1,\alpha_2...\alpha_8$ is the model coefficient.

2.4. Decision Tree

Decision Tree (DT) is a supervised learning method that achieves classification and regression through a tree structure. It selects the optimal features through a recursive approach and segments the training data based on these features, thereby achieving correct classification of diabetes-related features. DT features automatic feature selection and interactive capture, strong visualization, and high flexibility. The common segmentation methods of DT include the Gini index and information gain. The Gini index can quantify the purity of node samples. The lower the value, the more samples in the node belong to the same category, providing a clear division basis for the construction of decision trees. Its calculation formula is:

$$Gini(Z)=1-\sum_{i=1}^{k} p_i^2. \quad (2)$$

Among them, $Gini(Z)$ is the Gini index, p_i is the probability of the category i, and Z is the current research dataset.

Information gain can preferentially select features that can significantly reduce the information entropy of the dataset for partitioning, thereby reducing classification errors. The calculation formula is:

$$IG=H(Z)-\sum_{i=1}^{n} \frac{|Z_i|}{|Z|} H(Z_i). \quad (3)$$

Among them, IG is the information gain, and $H(Z)$、$H(Z_i)$ is the information entropy of the dataset. The calculation formula is:

$$H(Z)=-\sum_{i=1}^{k} p_i \log(p_i). \quad (4)$$

Among them, p_i is the probability of the category i, Z is the data set under study, and Z_i is a subset divided according to features.

2.5. Random Forest

Random Forest (RF) is an ensemble learning method and a classifier that contains multiple decision trees. This paper improves the accuracy and generalization ability of the diabetes prediction model by constructing multiple decision trees and conducting voting or averaging the prediction results. Each tree of RF learns independently, which can achieve efficient parallelization. Even if a single tree is overfit, the overall prediction performance of the model can still remain stable. RF has the characteristics of integrating multiple decision trees, randomizing samples and features, and preventing overfitting.

2.6. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an efficient gradient boosting decision tree algorithm. It builds a strong learner by integrating multiple weak learners, expands the objective function to the second order using the Taylor formula to optimize the loss, and adopts a regularization mechanism to effectively prevent overfitting. During training, it uses a forward step-by-step algorithm to gradually add the tree model. When each tree is iterated, it learns the shortcomings of the previous tree, thereby improving the performance of the diabetes prediction model. XGBoost supports custom loss functions and can automatically adjust to find the optimal combination of hyperparameters. It

features high flexibility and strong interpretability. The objective function of XGBoost is the sum of the loss function and the regularization term, and its calculation formula is:

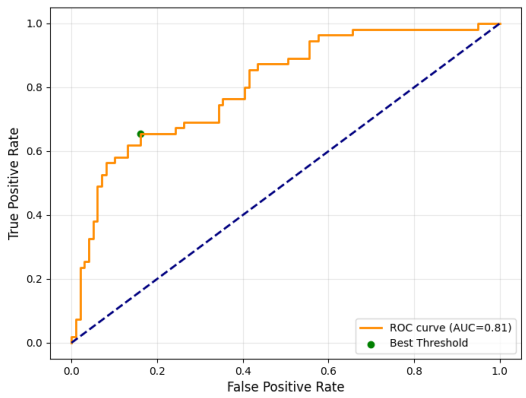
$$obj=\sum_{k=1}^{n} L(x_i,\hat{x_i})+\sum_{j=1}^{m} \Omega(g_j). \quad (5)$$

Among them, obj is the objective function, $L(x_i,\hat{x_i})$ is the loss function between the true value and the predicted value of the sample, x_i and $\hat{x}_i$ are respectively the true value and the predicted value of the sample i, and $\Omega(g_j)$ the complexity of the decision tree j.

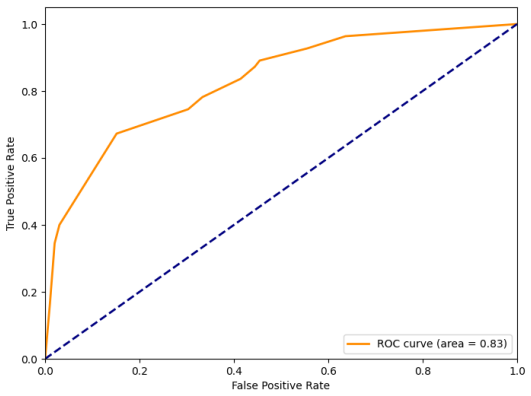
3. Results and discussion

3.1. The results of four algorithms: LR, DT, RF, and XGBoost

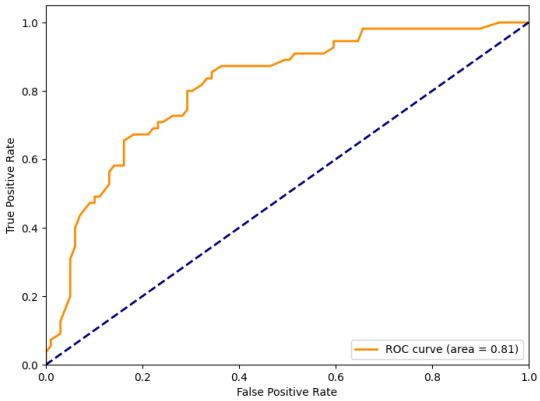
In this study, the top five important features selected from the distribution of LR's significance characteristics were plasma glucose concentration, BMI, age, diabetes pedigree function, and 2-hour serum insulin. On this basis, the model is reconstructed again through three algorithms: DT, RF, and XGBoost. The receiver operating characteristic curve (ROC) for evaluating the diabetes prediction performance of the classification model was ultimately obtained, as shown in Fig. 2. (a), (b), (c), and (d) are the ROC curves of LR, DT, RF, and XGBoost, respectively, with AUCs of 0.81, 0.83, 0.81, and 0.81, respectively. The range of AUC is between 0.5 and 1. The ROC curves of this study are all close to the upper left corner, indicating that the diabetes prediction models constructed in this paper all have good effects. Among them, the AUC of DT is the largest, and it is initially concluded that the model constructed by DT has the best performance.



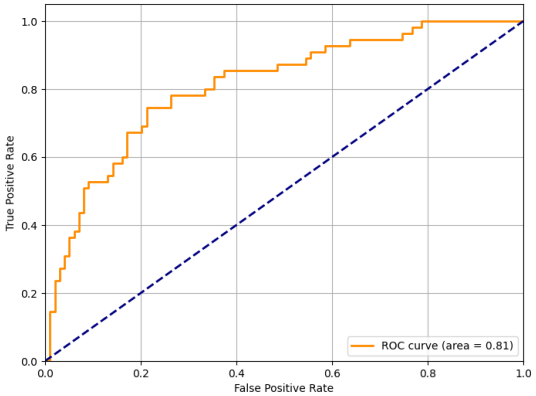
(a) The ROC curve of LR



(b) The ROC curve of DT



(c) The ROC curve of RF



(d) The ROC curve of XGBoost

Fig 2. ROC curves of LR, DT, RF and XGBoost (Original)

3.2. A comparison of four algorithms

To further compare the above four algorithms, the accuracy rates of each algorithm, as well as the precision, recall rate, and F1 score of the positive category, were calculated, respectively, as shown in Table 1:

Table 1. Model result

	accuracy	precision	recall rate	F1 score
LR	75.97%	67%	65%	66%
DT	78.57%	71%	67%	69%
RF	74.03%	62%	69%	66%
XGBoost	74.68%	61%	80%	69%

The F1 score is the harmonic mean of precision and recall. As shown in Table 1, DT and XGBoost have the largest F1 scores, and DT also has the highest accuracy and precision, indicating that the diabetes prediction model constructed by DT can capture more real confirmed cases. Although the recall of XGBoost led the way, its low precision rate would cause more healthy people to be misjudged as positive. Therefore, in terms of overall performance evaluation, the diabetes prediction model constructed by DT is the best.

3.3. Discussion

The experimental results show that the models constructed by different algorithms have both advantages and limitations. In practical applications, it is necessary to select appropriate models based on the different characteristics of the application scenarios and datasets. The sample set of this study was for female patients, and the number of features was limited, resulting in poor evaluation performance of the model [8]. By comparing the prediction models constructed by four algorithms, namely LR, DT, RF, and XGBoost, the coefficients of the logistic regression formula can directly reflect the influence and significance of different features on the risk of diabetes, but it is difficult to capture the complex interactions among different features [9]. The model constructed by DT has the best performance and can naturally handle the complex threshold effects between continuous variables. However, DT is prone to overfitting when dealing with large and complex datasets. RF can prevent overfitting by integrating multiple decision trees, thereby enhancing generalization ability. However, a smaller number of feature samples is easily overwhelmed by a larger number of samples, resulting in a low recall rate. XGBoost has advantages such as parallel computing acceleration and a regularization mechanism to prevent overfitting, but its training time is relatively long and its efficiency is low [10].

In the future, machine learning models can be further optimized by reducing their reliance on a single metric, enhancing the interpretability of the model, and introducing more robust feature selection and integration methods. The prevention of diabetes will also continue to advance due to technological breakthroughs and practical clinical applications, promoting the improvement of the public's health management level.

4. Conclusion

Ranked by the importance of characteristics, plasma glucose concentration, BMI, age, diabetes pedigree function, and 2-hour serum insulin are the top five important factors affecting diabetes. Therefore, in daily life, measures such as blood glucose concentration detection and enhanced physical exercise can be taken to prevent diabetes.

Although the prevention of diabetes is affected by the difficulty of modern lifestyle intervention and hormonal fluctuations, with the rapid development of science and technology and the continuous progress of human society, the excellent classification and prediction performance of machine learning in diabetes prediction models has made it widely applied in medical-related fields. In this study, the diabetes prediction model constructed by DT is the best. Compared with other algorithms,

it can capture more diabetic patients. However, the prediction of diabetes risk in clinical trials still needs further research. In the future, machine learning will be better applied in the field of diabetes prediction, and models can also be further optimized by transcending the reliance on a single indicator and enhancing interpretability. The prevention of diabetes will also continue to advance, promoting the improvement of the public's quality of life.

References

[1] Qu K. Diabetes prediction model based on machine learning [D]. Tianjin University, 2019.

[2] Zhou B., Rayner A. W., Gregg E. W., Sheffer K. E., Carrillo-Larco R. M., Bennett J. E., et al. Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. The Lancet, 2024, 404(10467): 2077-2093.

[3] Ma W., Wang K., Yu B., et al. The diabetes risk prediction model based on physical examination data contrast research. Journal of Modern Information Technology, 2020, 4(23): 72-75.

[4] Ouyang P., Li X., Leng F., et al. Application of machine learning algorithms in diabetes risk prediction of physical examination population. Chinese Journal of Disease Control and Prevention, 2021, 25(7): 849-853+868.

[5] Zhao X., Ji J., Wang L. Diabetes risk prediction model based on machine learning and empirical research. Journal of Huzhou Normal University, 2022, 44(8): 55-62.

[6] Xiang J. Application of machine learning in predicting diabetes. China Science and Technology Information, 2025(14): 77-80.

[7] Rahman M. H. Diabetes dataset. Kaggle, 2024-11-01. Accessed: 2025-12-08. Available: https://www.kaggle.com/datasets/hasibur013/diabetes-dataset

[8] Zou Q., Zhang Y., Wan Y., et al. Machine learning methods for constructing diabetes-related predictive models. Chinese Journal of Health Statistics, 2023, 40(4): 631-635+640.

[9] Liu Y., Jiang M., Li D., et al. Construction and validation of a prediction model for dysphagia in elderly stroke patients based on interpretable machine learning methods. Chinese Journal of Geriatric Cardiovascular and Cerebrovascular Diseases, 2020, 27(6): 698-704.

[10] Huang Y., Wu X., Yang J. Research and application progress of predictive model for diabetic kidney disease based on machine learning. Chinese Health Standard Management, 2025, 16(9): 194-198.