Comparative Research on Diabetes Influencing Factors Based on Random Forest and Decision Tree Models

Liping Li *

College of Financial Mathematics and Statistics, Guangdong University of Finance, Guangdong, 510521, China

* Corresponding author: 201527219@m.gduf.edu.cn

Abstract. In tandem with society's rapid progress, the prevalence of diabetes has risen sharply due to factors such as changes in eating habits, bad lifestyle and serious aging problems. Therefore, it is of great significance to forecast the influencing factors of diabetes mellitus. In this paper, the Pima Indian Diabetes data set in UCI is taken as the experimental data, the Random Forest and Decision Tree methods are used for modeling. The effect of these two models is analyzed according to four indicators, comprised of rate of accuracy, ratio of precision, rate of recall and F1-score. The F1-score ratio of the Random Forest model is as high as 79%, while the F1-score ratio of the Decision Tree model is 72%. Hence, the Decision Tree model is slightly inferior to the Random Forest model. Based on the predictive variables selected in this article, the most important predictive variables of the factors causing diabetes are blood glucose, body mass index, age and diabetes pedigree function.

Keywords: Diabetes, Machine learning, Random Forest, Decision Tree, Influencing factor.

1. Introduction

Diabetes is an endocrine disease in which the pancreas does not secrete insulin properly. According to the survey, diabetes comprises over 170 different types of complications, which have a high fatality rate, thus patients' physical health is much at danger. The International Diabetes Federation (IDF) estimates that 537 million people worldwide have diabetes in 2021, with 783 million anticipated to have the disease by 2045, and the proportion of adults with diabetes may reach one in eight. The majority of people worldwide who have diabetes reside in China. Nowadays, China has 140 million diabetics, up from 90 million in the previous ten years, a 56% increase [1].

As the third major chronic non-communicable disease to threaten human health after cardiovascular disease and cancer, the costs associated with treating diabetes are rising at an alarming rate. Based on projections from the World Health Organization (WHO), between 2005 to 2015, diabetes led to an economic loss of about 557.7 billion dollars, of which about 109 billion dollars was spent in 2019 alone, undoubtedly bringing a huge social and economic burden to China [2]. At present, since diabetes cannot be totally eradicated, adopting active intervention strategy for it and other risk factors is essential. Effective diabetes management can lessen the expense of preventative and treatment as well as the potential of being diagnosed with diabetes. The cost of treating individuals will decrease with early diagnosis of the illness.

Owing to the rise in the number of diabetics and substantial losses in the economy caused by the disease, a large number of scholars at home and abroad actively engaged in the study of diabetes and its complications, using various methods to establish models to analyze and predict the disease. In foreign countries, the early construction of diabetes risk assessment models was mainly based on statistics. Smith used Logistic Regression (LR) method to classify and predict the Pima Indian diabetes (PID) data set, and the corresponding accuracy was 79.17% [3]. Aiswarya used J48 Decision Tree and Naive Bayes as classifiers, utilizing PID data set to classify the diagnosis of diabetes, and the classified outcomes were 74.8% and 79.5% respectively [4]. Anuja proposed a Support Vector Machines (SVM) based diabetes prediction model by using PID data set as the captcha data set for the model, the model's accuracy reached 78% [5]. Vigneswari used several Decision Tree (DT) algorithms such as C4.5, Logical Model Tree, Random Forest (RF) and Random Tree to predict the

influencing factors of diabetes. After comparing the performance of several machine learning tree classifiers, it was found that Logical Model Tree had the best expression effect [6].

Scholars from all walks of life in China also committed to apply various statistical methods to carry out profound research on the disease. Qin W used a Logistic algorithm and RF model to forecast the blood glucose of type II diabetes patients in Beijing, among which the RF model has a better effect [7]. Zhang C F optimized the Extreme Gradient Boosting model by genetic algorithm and compared it with several other machine learning models, the optimal classification prediction model was obtained through the fusion of Stacking models finally [8]. Lin X built a diabetes risk prediction model based on Back Propagation (BP) neural network, RF and SVM, and evaluated the prediction results by accuracy, recall rate and other indicators, among which the accuracy rate of the RF algorithm reached 91.04% [9]. Zhang Q constructed diabetes prediction models based on Artificial Neural Network, SVM, RF and LR algorithms, and combined the models with the integrated learning methods [10]. Chen S J used the disease factors of BP neural network to quantitatively analyze the effects of blood glucose and screened out 10 factors such as gender and weight [11].

These findings suggest that many factors contribute to diabetes. There may be a complex relationship between diabetes risk and people's basic conditions, living habits, routine physical examination indicators, etc. However, people's awareness of diabetes prevention is relatively weak, and thus it is the responsibility of contemporary society to detect diabetes risk as early as possible and carry out preventive interventions. The establishment of an effective diabetes diagnosis model is extremely crucial for both preventing and treating diabetes. In this paper, two methods which contain RF and DT were used to conduct a comparative study on the influencing factors of diabetes, aiming to enhance the performance of illness forecast model, increase the screening rate of diabetes disease, and reduce the degree of disease deterioration.

2. Research Methods Design

2.1. Data Source and Description

The data in this paper are derived from the PID dataset in the UCI dataset, which is mainly about the characteristic information of Pima Indian women over 21 years old. There are a total of 768 samples, including 9 characteristic variables, respectively 8 predictors and 1 label variable (label value 1 represents diabetes, label value 0 represents no diabetes). The eight attribute variables were pregnancies, blood glucose, blood pressure, triceps skin thickness, serum insulin, body mass index, diabetes pedigree function and age. According to the Target variable label value category statistics, there are 268 samples with a value of 1 (having diabetes) and 500 samples with a value of 0 (not having diabetes) (Fig. 1).

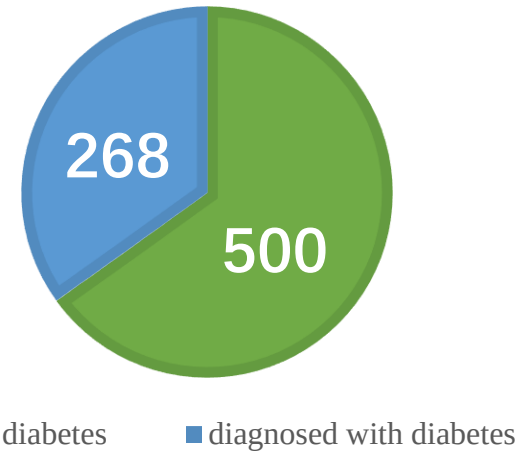


Figure 1. Target variable label value classification (Photo credit: Original)

2.2. Variable Selection

Table 1. Predictors in the Pima Indian diabetes dataset

Characteristic name	Minimum	Maximum	Mean	Variance
pregnancies	0.000	17.000	3.845	11.354
blood glucose	0.000	199.000	120.895	1022.248
blood pressure	0.000	122.000	69.105	374.647
triceps skin thickness	0.000	99.000	20.536	254.473
serum insulin	0.000	846.000	79.799	13281.180
body mass index	0.000	67.100	31.993	62.160
diabetes pedigree function	0.078	2.420	0.472	0.110
age	21.000	81.000	33.241	138.303

Table 1 shows the descriptive statistics of various characteristic variables in the diabetes data set of Pima Indians. Due to the high incidence of diabetes among the local population, samples of the population in this region are selected for investigation and research. The above variables are significant influencing factors for diabetes, because meaningless information and overly complex information characteristics should be avoided while measuring numerous influencing factors, so that the analysis is based on the above 8 predictors.

2.3. Research Method

A method of machine learning called Random Forest depends on Decision Tree generalization classification. The sampling method is adopted in this model to train and predict multiple trees in a random way, each tree will make corresponding judgment results according to the samples. Its advantages include excellent model generalization capabilities, rapid training speed and straightforward implementation techniques when handling highly dimensional data. It has a high tolerance for noise and abnormal data, which can simplify data information and remove redundant information. RF is extremely popular data mining method, and it can perform data regression as well as classify data according to data attributes. The algorithm's job is to determine how significant a variable is. The capacity of a variable to classify result variables is better when it has a higher significance score, and interactions between characteristic variables in the diabetes data set can be detected during training.

A method for categorizing data based on a set of rules is the Decision Tree. It primarily uses inductive algorithms to produce understandable rules and decision trees, after which it employs decisions to examine fresh data. The mapping between object properties and object values is represented by this model. Its advantages are that the algorithm is easy to understand, simple explanation mechanism, and less time complexity. It can clearly illustrate how the entire problem's decision-making process worked in various stages according to time and order. The phases and levels are evident when applied to complicated multi-stage decision-making, which is useful for collaborative investigation of decision-making institutions. It has the capacity to carefully evaluate a variety of factors, which is helpful for making proper choices.

Therefore, this paper chose to use RF and DT methods to process diabetes data sets, and establish corresponding diagnostic models. According to the results of the models, the factors causing diabetes are further analyzed, and the important factors affecting diabetes are obtained, and then the operation results of the two models are compared and analyzed.

3. Results and Discussion

3.1. Descriptive Statistics

3.1.1. Pearson Correlation Test

Table 2. Pearson correlation table

Independent Variable (X)	Dependent Variable (Target)
pregnancies	0.222**
blood glucose	0.467**
blood pressure	0.065
triceps skin thickness	0.075*
serum insulin	0.131**
body mass index	0.293**
diabetes pedigree function	0.174**
age	0.238**

Comment: * $p<0.05$ ** $p<0.01$

Pearson correlation coefficient can be used to obtain the strength of the correlation between Target variable and 8 predictors. As demonstrated by Table 2, the coefficients of correlation between Target and pregnancies, blood glucose, serum insulin, body mass index, diabetes pedigree function, and age are 0.222, 0.467, 0.131, 0.293, 0.174 and 0.238 respectively, and they all indicate a significance level of 0.01, reflecting an extremely positive correlation. The phase relation value between Target and blood pressure is 0.065, close to 0, proving that there is no connection between blood pressure and Target. The phase relation between Target and triceps skin thickness is 0.075 and significant at 0.05 level, demonstrating a positive correlation between Target and triceps skin thickness. In general, except that there is no correlation between the variable blood pressure and the Target variable, all other predictors are positively correlated with the target variable.

In conclusion, the correlation coefficient between Target and blood glucose is the highest, followed by the correlation coefficient between Target and body mass index, age, pregnancies, diabetes pedigree function decreasing successively. The correlation coefficient between Target and serum insulin, blood pressure, and triceps skin thickness was lower. It can be preliminarily learned that the variables of blood glucose, body mass index, age and pregnancies have more significant weights than other variables in predicting whether a sample has diabetes, and they can be recorded as influential factors that are more likely to cause the disease for prediction.

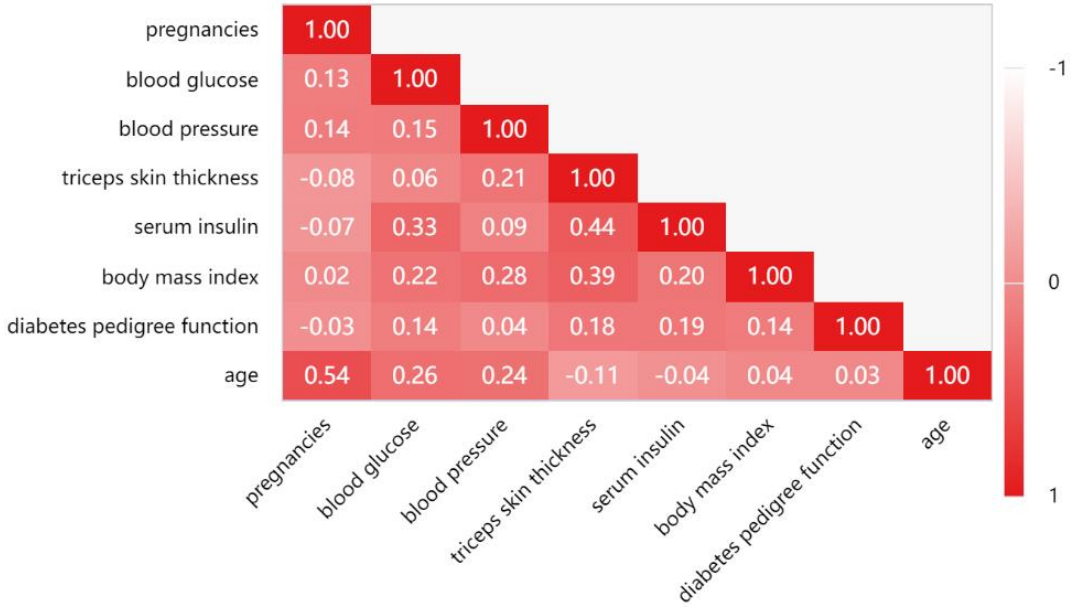
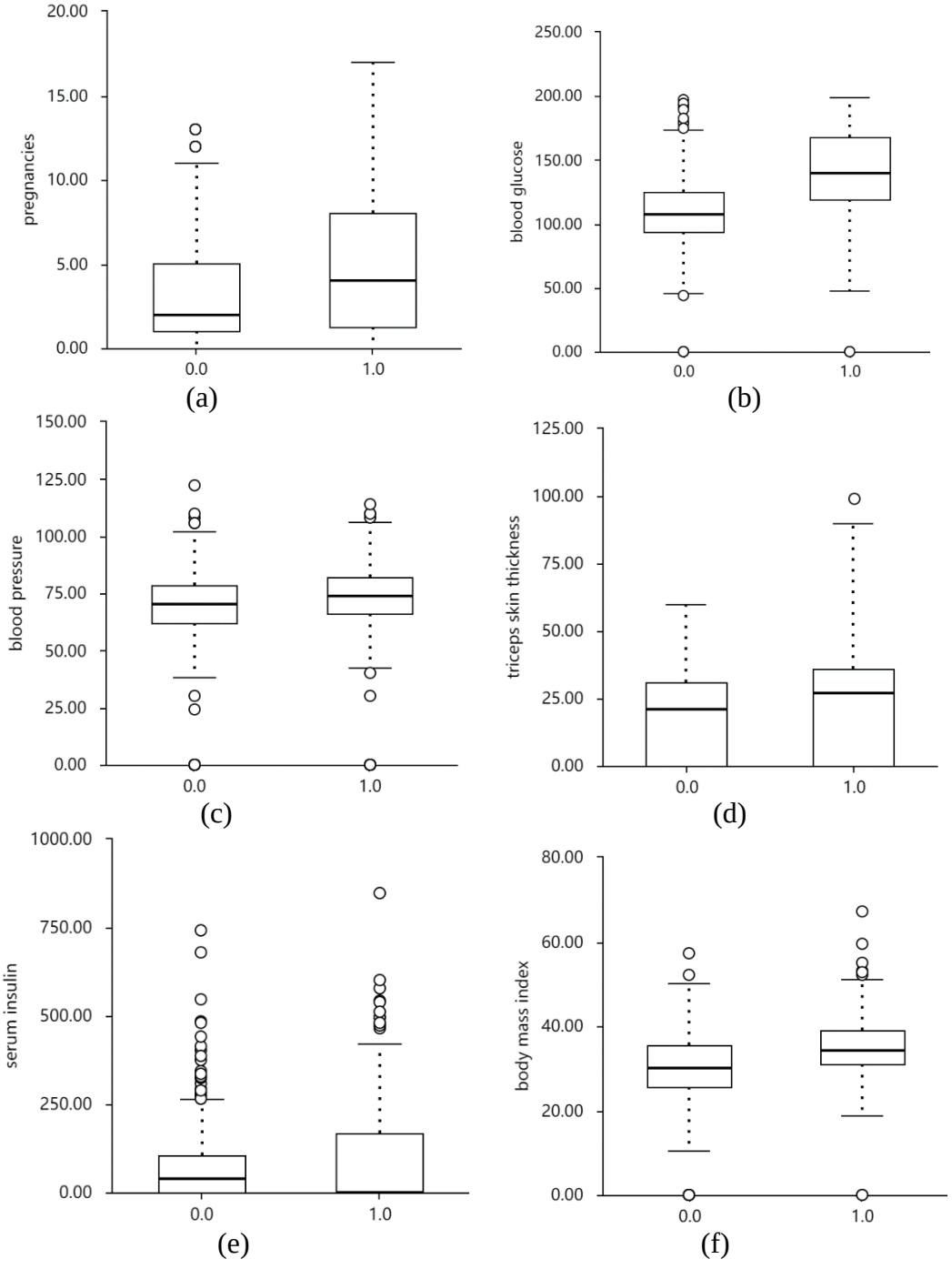


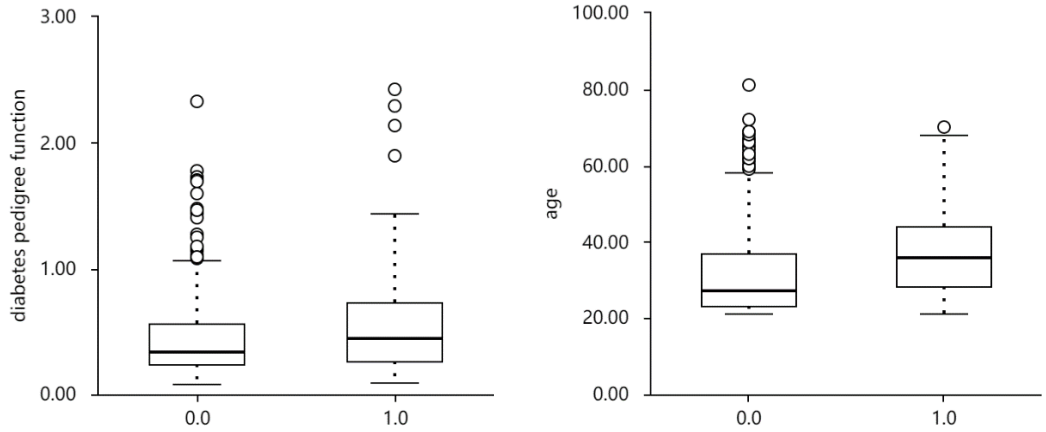
Figure 2. Pearson correlation coefficient diagram (Photo credit: Original)

Fig. 2 shows the correlation coefficient between variables. Specific analysis shows that the correlation values between pregnancies and blood glucose, blood pressure and age are 0.13, 0.14 and 0.54 respectively, all of which are greater than 0. The results indicate that pregnancies are positively correlated with the three items. At the same time, the correlation values between pregnancies and triceps skin thickness, serum insulin, body mass index, diabetes pedigree function are close to 0, indicating that there is no correlation between the four items. Analyze the remaining variables in the same way, The relationship between pregnancies and age, blood glucose and serum insulin, triceps skin thickness and serum insulin can be obtained the correlation between them are stronger than other variables.

3.1.2. Box Plots

In this study, samples with diabetes in the data set were labeled as factor 1, and samples without diabetes were labeled as factor 0. Then, the two classified samples and each influencing factor were visually analyzed in the form of boxplot, and the results were shown in Fig. 3(a)-(h) below.



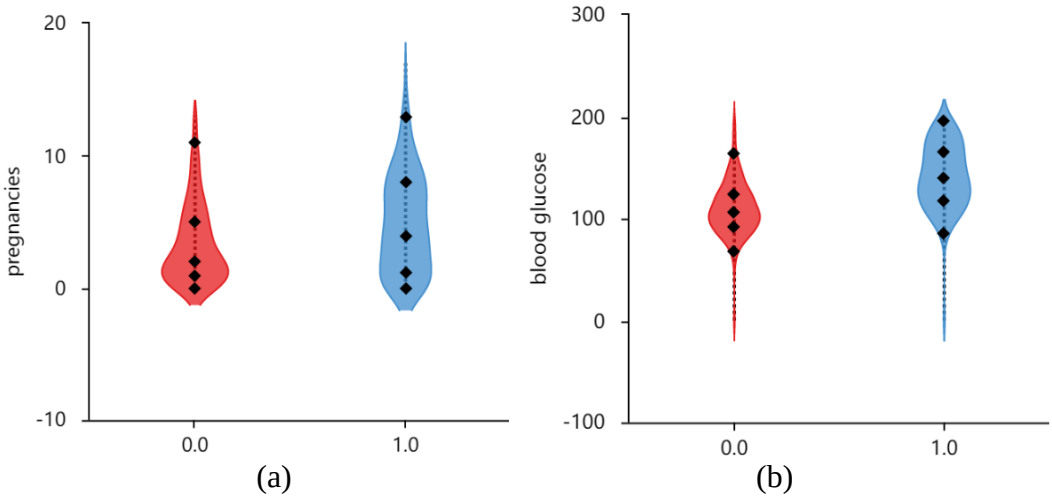


(g) (h)

Figure 3. Box plot of predictor variables (Photo credit: Original)

As shown in Fig. 3, the sample population with and without diabetes showed some differences in the predictive variables. Among them, the values of pregnancies, blood glucose, diabetes pedigree function and age were significantly different among people with or without the disease. As shown in Fig. 3(b), the blood glucose of non-diabetic individuals was generally between 90 and 125, while that of diabetic individuals was between 120 and 170. As shown in Fig. 3(h), the age range of patients with diabetes was generally 30-50 years old, and the age range of those without diabetes was 21-38 years old, indicating that the blood glucose and individual age of patients were higher than those of non-patients. Second, there was little difference in blood pressure values between people with and without diabetes, suggesting that blood pressure may play little role in predicting whether an individual has diabetes. In addition, with respect to the determination of serum insulin and triceps skin thickness in different groups, the maximum values of these two variables were compared between patients and non-patients. Therefore, these special factors ought to be fully given consideration when predicting the possibility of an individual developing diabetes.

3.1.3. Violin Plots



(a) (b)

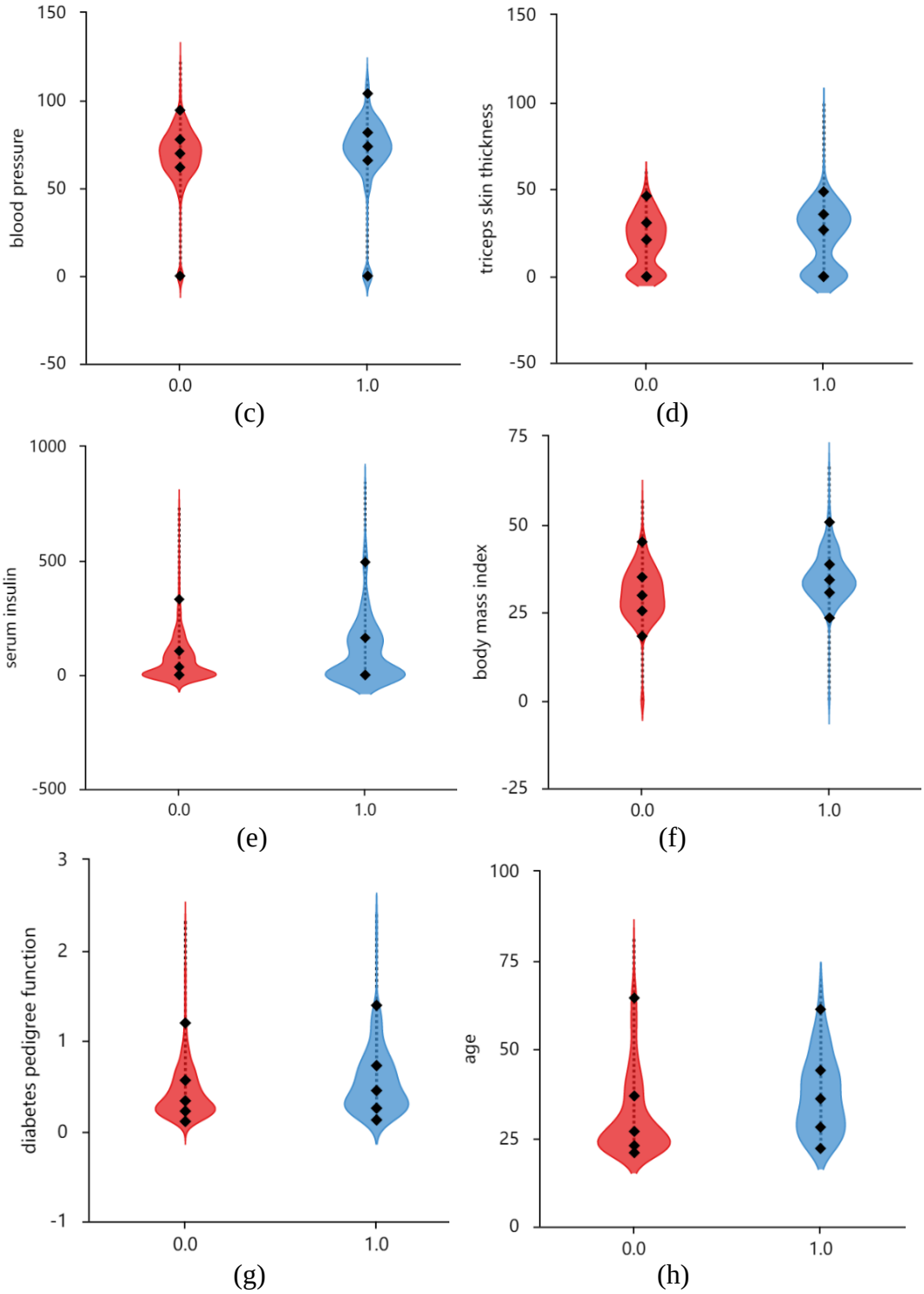


Figure 4. Violin plot of predictor variables (Photo credit: Original)

As viewed by Fig. 4, there were certain differences in the numerical distribution of influencing factors that cause disease or not. For example, from the analysis of age factors, the age fluctuation of the population without disease was mostly distributed in the mid-20s, but the population with disease was mostly distributed in the older age range. According to the factor analysis of the number of pregnancies, the number of pregnancies in the population without the disease was mostly less than 5 times, and the number of pregnancies in the population with the disease was more than 5 times, and even the 97.5% sub-point of the violin diagram reached 14 times. In terms of blood glucose factors, the distribution of people without the disease was mostly between 90 and 110, while that of people with the disease was mostly between 110 and 180. According to the comparison of the differences in the distribution patterns of various violin charts, factors comprising but not restricted to pregnancies,

blood glucose, body mass index, age and other factors had a certain profound impact on diabetes. These factors needed to be further analyzed.

3.2. Inferential Statistics

3.2.1. Random Forest

Table 3. Summary of basic information of Random Forest classification

Name	Option	Frequency	Percentage
Target	0	500	65.10%
	1	268	34.90%
	Total	768	100.00%
Summarize	Effectivity	768	100.00%
	Deficiency	0	0.00%
	Total	768	100.00%

Eight predictive variables are included in the random forest model as independent variables and the dependent variable used in prediction research is the target variable. Table 3 provides evidence for that samples without diabetes accounted for 65.10% of the total and the proportion of samples with diabetes accounted for 34.90%. 768 samples in all are used in the analysis, and the participation rate of all samples is 100%.

Table 4. Feature weight values (Random Forest)

Item	Weighted value
pregnancies	0.09
blood glucose	0.248
blood pressure	0.088
triceps skin thickness	0.075
serum insulin	0.077
body mass index	0.161
diabetes pedigree function	0.122
age	0.139

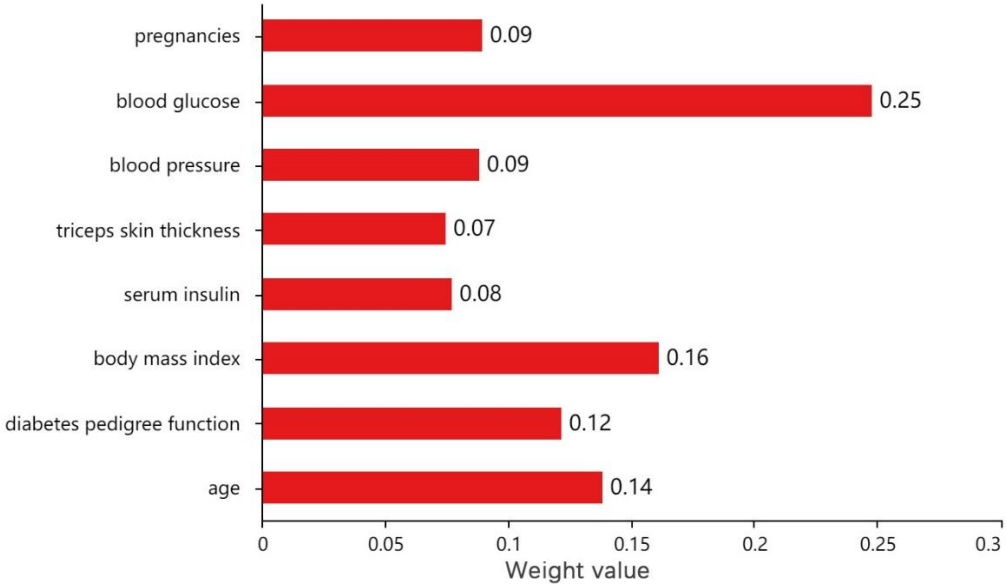


Figure 5. Random Forest feature weight diagram (Photo credit: Original)

The feature weight of each variable illustrates how significant each variable is to the model, and the overall value is 1. From Table 4 and Fig. 5, it can be seen that blood glucose accounts for 24.83%, it gains the greatest weight and is crucial to the creation of models. The proportion of body mass index, age, diabetes pedigree function and pregnancies are 16.13%, 13.86%, 12.16% and 8.99%

respectively. The proportion of the above 5 characteristics accounted for 75.97%; The proportion of blood pressure, serum insulin and triceps skin thickness of the remaining 3 items are 8.82%, 7.74% and 7.47% respectively. This effectively indicates that blood glucose, body mass index, age and diabetes pedigree function have a strong influence on the random forest model, therefore the most important attribute variables are blood glucose, body mass index, age and diabetes pedigree function.

Table 5. Results of the training set model's evaluation (Random Forest)

Item	Ratio of precision	Rate of recall	F1-score	Sample capacity
0	1	1	1	401
1	1	1	1	213
Rate of accuracy			1	614
Value of average	1	1	1	614
Value of average (synthesis)	1	1	1	614

oobScore=1(out-of-pocket data test)

Table 6. Results of the test set model's evaluation (Random Forest)

Item	Ratio of precision	Rate of recall	F1-score	Sample capacity
0	0.83	0.86	0.84	99
1	0.73	0.67	0.7	55
Rate of accuracy			0.79	154
Value of average	0.78	0.77	0.77	154
Value of average (synthesis)	0.79	0.79	0.79	154

Table 5 and Table 6 represent the RF model's performance on the training set and the test set respectively. The percentage of correct samples in the overall samples of the projected results is known as rate of accuracy; The percentage of positive samples in the projected findings is known as the ratio of precision; The percentage of positive samples within the positive samples is referred to as rate of recall; F1-score is the comprehensive evaluation index of the precision ratio and recall rate, and is also the harmonic average of precision ratio and recall rate. When precision ratio and recall rate are higher, better results will be achieved. But these two rates consistently contradict each other, F1-score is thus frequently utilized to comprehensively assess the classifier's impact. Its value is in the 0 to 1 range, and the impact improves with proximity to 1. Table 6 shows that the value of F1-score is as high as about 80%, which proves that the model has a good effect.

Table 7. Table of model overview (Random Forest)

Define	Name of parameter	Values of parameter
Model parameter setting	Data preparation	None
	Ratio of training sets	0.8
	Decision tree quantity	100
	Node splitting criterion	Gini
	Minimum sample size for dividing nodes	2
	Minimum sample size for leaf nodes	1
	Largest tree depth	No restriction
	Limit the maximum quantity of features	auto
	Whether there is put back sampling	Yes
	Whether to perform out-of-pocket data testing	Yes
Model evaluation effect	Rate of accuracy	79.22%
	Ratio of precision (synthesis)	78.96%
	Rate of recall (synthesis)	79.22%
	F1-score(synthesis)	79.00%

The largest tree depth is unrestricted, and Gini is the node dividing criterion. There are 100 decision trees, with a training set proportion of 0.8. As can be seen from Table 7, with regard to the test set's final model, the rate of accuracy is 79.22%, the ratio of precision (synthesis) is 78.96%, the

rate of recall (synthesis) is 79.22%, and the F1-score(synthesis) is 79.00%. The model's rate of accuracy, precision, recall and f1-score are as high as about 80%, indicating that it has beneficial results from a statistical standpoint (Table 7).

3.2.2. Decision Tree

Table 8. Summary of basic information on Decision Tree classification

Name	Option	Frequency	Percentage
Target	0	500	65.10%
	1	268	34.90%
	Total	768	100.00%
Summarize	Effectivity	768	100.00%
	Deficiency	0	0.00%
	Total	768	100.00%

Eight predictor variables are taken as independent variables and Target is the dependent variable for DT modeling. Table 8 indicates that all of 768 samples are used in the test, and sample participation rate is consistent with the RF model.

Table 9. Feature weight value (Decision Tree)

Item	Weighted value
pregnancies	0.052
blood glucose	0.3
blood pressure	0.102
triceps skin thickness	0.049
serum insulin	0.043
body mass index	0.185
diabetes pedigree function	0.121
age	0.148

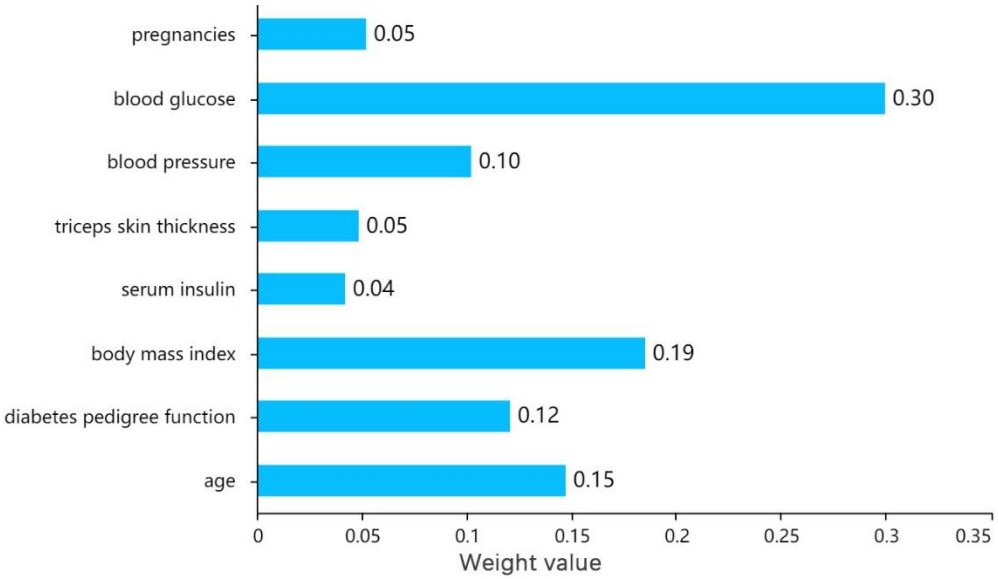


Figure 6. Decision Tree feature weight diagram (Photo credit: Original)

The feature weight obtained by the decision tree method demonstrates how important each title is to the model, and its total value is 1. Table 9 and Fig. 6 represent that the proportion of blood glucose is 30.02%, it gains the greatest weight and is crucial to the creation of models. The proportion of body mass index, age and diabetes pedigree function are 18.51%, 14.77% and 12.11% respectively. The proportion of the above 4 characteristics accounted for 75.41%; The proportions of blood pressure, pregnancies, triceps skin thickness and serum insulin of the remaining 4 items were 10.24%, 5.20%, 4.89% and 4.26% respectively. This effectively indicates that blood glucose, body mass index, age

and diabetes pedigree function have a strong influence on the DT model. It is concluded that the most important attribute variables were blood glucose, body mass index, age and diabetes pedigree function. Table 10 and Table 11 show that the performance of the decision tree model on the training set and the test set, the ratio of F1-score from the test set is about 72%, indicating that the model has a good effect.

Table 10. Results of the training set model's evaluation (Decision Tree)

Item	Ratio of precision	Rate of recall	F1-score	Sample capacity
0	1	1	1	401
1	1	1	1	213
Rate of accuracy			1	614
Value of average	1	1	1	614
Value of average (synthesis)	1	1	1	614

Table 11. Results of the test set model's evaluation (Decision Tree)

Item	Ratio of precision	Rate of recall	F1-score	Sample capacity
0	0.77	0.82	0.79	99
1	0.63	0.56	0.6	55
Rate of accuracy			0.73	154
Value of average	0.7	0.69	0.7	154
Value of average (synthesis)	0.72	0.73	0.72	154

Table 12. Table of model overview (decision tree)

Define	Name of parameter	values of parameter
Model parameter setting	Data preparation	None
	Ratio of training sets	0.8
	Decision tree quantity	Gini
	Node splitting criterion	best
	Minimum sample size for dividing nodes	2
	Minimum sample size for leaf nodes	1
	Largest tree depth	No restriction
Model evaluation effect	Rate of accuracy	72.73%
	Ratio of precision (synthesis)	72.19%
	Rate of recall (synthesis)	72.73%
	F1-score(synthesis)	72.00%

The largest tree depth is unrestricted, and Gini is the node dividing criterion. There are 100 decision trees, with a training set proportion of 0.8. Tables 10-12 provide evidence for that the final model's test set rate of accuracy is 72.73%, rate of accuracy(synthesis) is 72.19%, the rate of recall(synthesis) is 72.73%, and the F1-score(synthesis) is 72.00%. From the perspective of various indicators, the index rate of the decision tree model is about 72%, which is slightly worse than the model effect of the RF model. But the overall view still has some reference value.

4. Conclusion

In this paper, the PID data set was selected as this article research data. A total of eight influencing factors were diabetes, including pregnancies, blood glucose, blood pressure, triceps skin thickness, serum insulin, body mass index, diabetes pedigree function and age. Target variable was whether diabetes was diagnosed or not. The RF and DT algorithms were used to predict the data set, and the rate of accuracy, ratio of precision, rate of recall and F1-score ratio were utilized to analyze the model. The results showed that the RF model was more suitable for predicting the data set, and this model had a stronger prediction ability than the DT model. According to the analysis of influencing factors of diabetes in this data set by RF model, the weight of four features such as blood glucose, body mass

index, age and diabetes pedigree function accounted for the largest proportion, which proved that the degree of influence was stronger than other factors. Therefore, the situation of elevated blood glucose should be focused on, timely medical examination and prevention as soon as possible. The increase in body mass index and age makes the risk of diabetes gradually rise, obese people and elderly people should be vigilant, strengthen the awareness of diabetes prevention, pay attention to the balance of dietary structure and maintain good eating habits. In general, this paper aims to provide support for the prevention and research of diabetes, and the several predictive variables with the top weight ratio of the above characteristics can become the focus of diabetes research and the primary prevention object of diabetes prevention.

References

[1] Ma Y H. Regression and classification prediction model of diabetes mellitus based on ensemble learning [D]. Shanghai normal university, 2023.

[2] Chinese Society of Diabetology. Chinese type II diabetes prevention guide [J]. Chinese endocrine metabolism, 2014, 30 (008): 893 - 942.

[3] Smith J W, Everhart J E, Dickson W C, et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus [J]. Johns Hopkins Apl Technical Digest, 1988, 10, 261 - 265.

[4] Aiswarya I, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques [J]. International Journal of Data Mining & Knowledge Management Process, 2015, 5 (1): 1.

[5] Anuja K A, Chitra R. Classification of diabetes disease using support vector machine [J]. international journal of engineering research and applications, 2013, 3 (2): 1797.

[6] Vigneswari D, Kumar N K, Raj V G, et al. Machine learning tree classifiers in predicting diabetes mellitus [C]//2019 5th international conference on advanced computing & communication systems, IEEE, 2019: 84 - 87.

[7] Qin W, Gao M, Shen Y, et al. Prediction of 3-month blood glucose in patients with type 2 diabetes based on machine learning algorithm[J]. Chinese Journal of Disease Control, 2019, 23 (11): 1313 - 1317.

[8] Zhang C F. Application of machine learning algorithm based on parameter optimization in diabetes prediction [D]. Southwest University of Science and Technology, 2020.

[9] Lin X, Li J, Liu L, et al. Comparison of risk prediction models for type 2 diabetic nephropathy [J]. Chinese Journal of Medical Library and Information, 2019, 28 (04): 41 - 45.

[10] Zhang Q, Zhou L Q, Zou Q, et al. Research on the forecasting methods of diabetes based on Stacking [J]. Intelligent Computers and Applications, 2020, 10 (02): 107 - 110.

[11] Chen S J, Luo S L, Pan L M, et al. Quantitative analysis of the influence of risk factors of type II diabetes mellitus on blood glucose [J]. Journal of Beijing Institute of Technology, 2014, 34 (2): 201 - 206.