Research on diabetes Prediction and Intelligent Diagnosis System Based on Machine Learning

Huang Lin *, Tengyi Wang

College of Artificial Intelligence and Big Data, Hefei University, Hefei, China

* Corresponding Author Email: 1436401334@qq.com

Abstract. This paper mainly focuses on the diagnosis and prediction of diabetes, and uses machine learning algorithm to study the prediction model of diabetes and intelligent consultation recommendation system. First, based on the public dataset of diabetes in UCL database, logistic regression model was used to screen variables, and four main characteristics were obtained: Pregnancies, Glucose, BMI and DPF. Secondly, four machine learning algorithms, K-nearest neighbor algorithm (KNN), naive Bayesian algorithm, decision tree and support vector machine, are used to establish the prediction model of diabetes. The model results show that the decision tree algorithm and K-nearest neighbor algorithm have high accuracy in specificity and negative case hit rate; However, the ROC curves of K-nearest neighbor algorithm and support vector machine algorithm are overlapped. Therefore, AUC value is further used to compare the prediction results of the two models. Finally, the prediction model based on support vector machine has certain advantages in prediction accuracy. Finally, taking diabetes patients as the object, on the basis of early prediction, using data mining technology, an intelligent diagnosis system for diabetes based on Java and fuzzy reasoning is constructed. The system functions include system user management, user information management, diabetes diagnosis and advice, system management and other functional modules, of which diabetes diagnosis and advice module is the core module of the system. The intelligent diagnosis system can provide basic medical diagnosis and suggestions, thus improving the diagnosis efficiency and becoming an important auxiliary tool for medical diagnosis.

Keywords: Diabetes diagnosis; Logistic regression; Machine learning; Fuzzy reasoning; Intelligent diagnosis system

1. Introduction

With the continuous improvement of the social level and the continuous enrichment of spiritual civilization, the residents' physical health and medical security system are receiving more and more attention from the people and society. At the same time, the Party Central Committee and the State Council attach great importance to the ability to build a healthy China's national health quality and medical security, which will be an important indicator to measure a country's development level in the future.

At present, the number of diabetes patients in China is nearly 100 million, and there are many high-risk groups of diabetes. The symptoms of diabetes are related to other diseases. For some regions with underdeveloped medical technology departments, the diagnosis of diabetes is also a problem. Therefore, with the blessing of various adverse factors, some high-risk people with diabetes are more likely to convert to diabetes. And considering that patients with diabetes have limited knowledge of diabetes after suffering from it, they cannot find it in time and go to the hospital for treatment. In this regard, this paper uses machine learning algorithms to predict different physiological indicators for the diagnosis of diabetes, so as to get the risk of diabetes of the target, and carries out fuzzy reasoning diagnosis combined with the prediction results,In order to better improve the prediction accuracy of the diagnosis system, this paper adopts four algorithms, K-nearest neighbor, naive Bayes, decision tree and support vector machine, to establish the prediction model. At the same time, fuzzy query is used in the diagnosis system to improve the fault tolerance rate of the diagnosis.

2. Journals reviewed

At present, there are relevant studies on diabetes models abroad, which can be generally divided into three categories: the first category is a prediction model built for the complications of diabetes patients, which can predict the probability of complications of patients in many years to come. As far as forecasting models are concerned, many researches have been done on such models and they are also widely used. For the model prediction of coronary heart disease in literature [1], compare the prediction performance of three models. The three prediction models based on logical regression, naive Bayes, and decision tree algorithms all have good prediction ability of coronary heart disease, and also have important reference value for the research of diabetes prediction algorithm in this paper. For example, the KNN algorithm in this paper also uses the mode of single factor analysis to analyze and solve the model, First, all collected physiological indicators are analyzed by correlation coefficient, and then feature screening is conducted by analyzing single physiological indicators.

According to the literature [2], we can understand the etiology and pathogenesis of diabetes, and preliminarily select the physiological indicators needed for diagnosis. Based on the understanding of different algorithms in literature [3] for prediction of diagnosis results, and the reference to the basic principle of fuzzy diagnosis mechanism in literature [4], including the construction of rule base and the construction of intelligent diagnosis system after fuzzy reasoning. Through the establishment of the membership function model, the fuzzy processing of data can improve the accuracy of prediction at the expense of a certain precision, and the construction of ontology knowledge base for the implementation of Jess fuzzy inference algorithm[5], so as to have a preliminary solution to the disease prediction and diagnosis mechanism.

3. Data acquisition and processing

3.1. Data source and analysis

In order to analyze the correlation of various indicators, 748 pieces of data were selected in the UCL public data set. Due to the existence of abnormal values, the data were preprocessed, and finally 729 pieces of data with high availability were obtained. Thus, the correlation matrix of eight independent variables is shown in Table 1.

Table 1. Variable correlation matrix

Corrcoef	Pns	Gce	BP	ST	Isn	BMI	DPF	Age
Pns	1.000	0.129	0.141	-0.082	-0.074	0.018	-0.034	0.544
Gce	0.129	1.000	0.153	0.057	0.331	0.221	0.137	0.264
BP	0.141	0.153	1.000	0.207	0.089	0.282	0.041	0.240
ST	-0.082	0.057	0.207	1.000	0.437	0.393	0.184	-0.114
Isn	-0.074	0.331	0.089	0.437	1.000	0.198	0.185	-0.042
BMI	0.018	0.221	0.282	0.393	0.198	1.000	0.141	0.036
DPF	-0.034	0.137	0.041	0.184	0.185	0.141	1.000	0.034
Age	0.544	0.264	0.240	-0.114	-0.042	0.036	0.034	1.000

Pns represents Pregnancies, Gce represents Glucose, BD represents Blood Pressure, ST represents Skin Thickness, and DPF represents Diabetes Pedigree Function.

3.2. Independent variable screening method

Logistic regression model is usually used to screen multiple variables of dichotomous results. Let the eight independent variables Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age be$\boldsymbol{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8}$respectively, and let the dependent variable Outcome be Y, and the expression is:

$$\boldsymbol{Y = \begin{cases} 1, \text{ With diabetes} \\ 0, \text{ No diabetes} \end{cases}} \tag{1}$$

You can select logical functions (Sigmaid functions): $\mathbf{g(z)} = \frac{\mathbf{1}}{\mathbf{1+e^{-z}}}$, Its function image is shown in Fig 1.

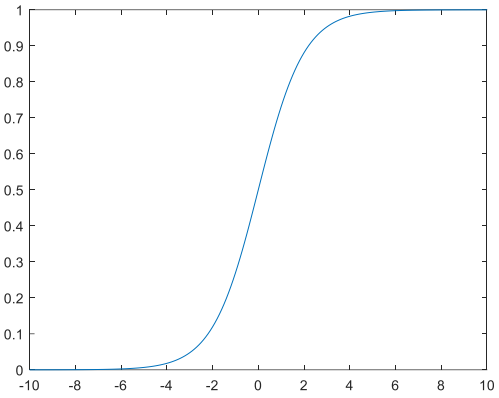

Fig 1. Sigmaid function image

Therefore, it can be determined that the logical regression function is

$$\boldsymbol{g(y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\theta_0+\theta x_1+\cdots\theta_n x_n)}} = \frac{1}{1+e^{-\theta^T x}}} \tag{2}$$

The probability of illness under the action of independent variable is recorded as P,$\mathbf{P = g(y)}$, According to the above formula, we can get the ratio of the probability of the occurrence of disease to the probability of the occurrence of no disease, and we can get the formula of the logit model, which is

$$\boldsymbol{logit(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_8 x} \tag{3}$$

In order to determine the size of parameter $\boldsymbol{\beta}$, maximum likelihood estimation is used to obtain the logarithmic maximum likelihood function as follows:

$$\mathbf{1(\theta) = ln\big(L(\theta)\big) = \sum_{i=1}^{m} \left(y_i ln\left(g_\theta(x_i) + (1-y_i) ln\left(1 - g_\theta(x_i)\right)\right)\right)} \tag{4}$$

Since the result value is independent of $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_0}$ is regarded as an invalid parameter. Next, construct the loss function. For the entire dataset, average the log likelihood loss to determine the loss function as

$$\boldsymbol{J(\beta) = -\frac{1}{m} ln\,(L(\beta))} \tag{5}$$

Stop the iteration when the loss function iteration is less than a threshold, and select the maximum number of iterations as 100. For the selection of variables, this paper selects the likelihood ratio statistics to test, and compares the value of the logarithmic likelihood function under two different assumptions. First prepare a logical model that does not contain the factors to be tested, obtain a function value $\mathbf{I_1}$, and then add the test factors to obtain $\mathbf{I_2}$

$$\boldsymbol{G = 2(I_2 - I_1)} \tag{6}$$

When the sample size is large enough, the G statistic can be seen as the χ^2 distribution of the number of samples in $\mathbf{n} = \mathbf{I_2}$ minus the number of samples in $\mathbf{I_1}$, and then the samples are filtered according to the determined confidence interval.

3.3. Data processing

In this paper, eight indicators in the dataset, namely Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age, are used as independent variables, and whether diabetes is a dependent variable. Logistic regression is used to screen the variables and study whether each indicator has a significant impact on whether diabetes is a dependent variable. Then the selected indicators are used as dependent **variables** [6]. First, all eight indicators are included for model establishment and factor screening, and the disease is classified with 0.5 as the threshold. When it is less than 0.5, it is considered that there is no diabetes, and when it is greater than 0.5, there is diabetes. Take the significance level of 0.05 for test, and obtain the variable analysis table as shown in Table 2 through model test.

Table 2. Variable Analysis Table

	B	Standard error	Significance	ranking
Pregnancies	0.115	0.033	0.001	2
Glucose	0.034	0.004	0.000	1
BD	-0.010	0.009	0.245	4
ST	0.001	0.007	0.915	6
Insulin	-0.001	0.001	0.262	5
BMI	0.096	0.017	0.000	1
DPF	0.98	0.306	0.001	2
Age	0.018	0.010	0.066	3

In Table 2, they are ranked by significance, among which Glucose and BMI rank first, and Pregnancies and DPF rank second. It can be clearly seen from Table 3 that the significance of the four indicators, namely, Pregnancies, Glucose, BMI and DPF, is less than that of the other four indicators. It can also be concluded that the four indicators, namely, Pregnancies, Glucose, BMI and DPF, are finally selected as the indicators used in the subsequent analysis. In order to intuitively display the size of correlation, a correlation matrix thermodynamic diagram is drawn as shown in Fig 2.

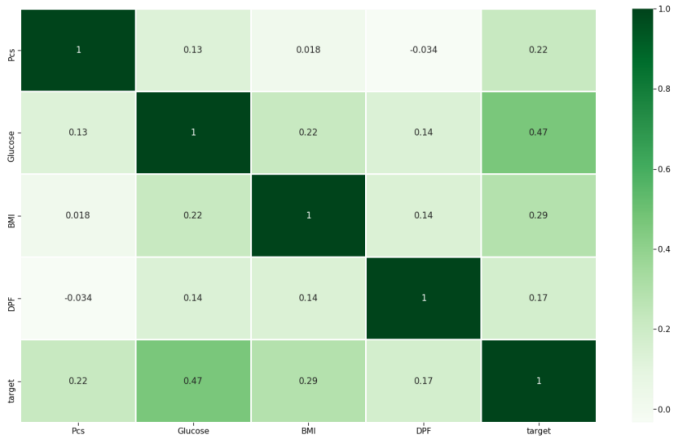

Fig. 2 Thermodynamic diagram of variable correlation matrix

4. Overview of research methods

In this paper, Logistic regression is used to screen variables, and four machine learning methods, KNN algorithm, decision tree, naive Bayesian algorithm and support vector machine, are used to establish diabetes prediction models, and further compare the prediction performance of each model. On the basis of the prediction model, an intelligent diagnosis system for diabetes based on Java and data mining is developed [7].

4.1. KNN algorithm

For an undetermined sample point, the known sample points around are usually observed. If the nearest K sample points around are the first type, the sample is divided into the first type. This is KNN algorithm. For more intuitive representation, a model diagram is shown in Fig 3.

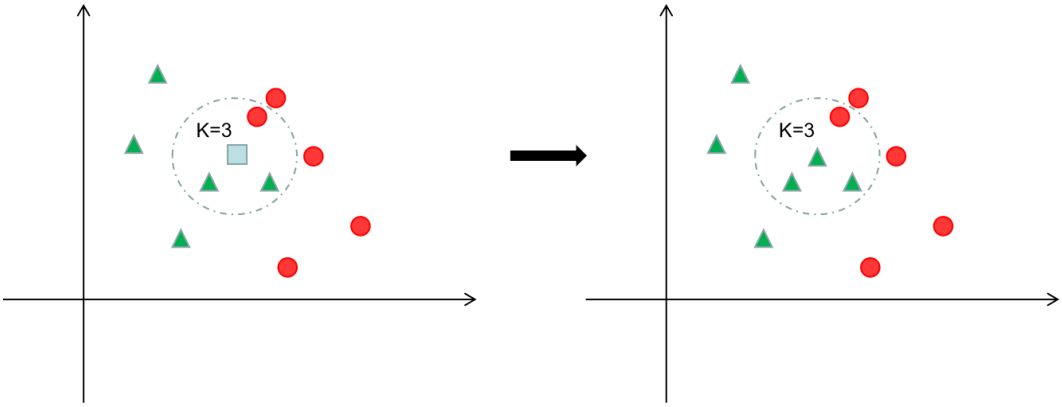

Fig. 3 Model diagram when K=3

When K=3, assume that the green points in the graph are the points to be predicted. Then, KNN algorithm will automatically find the three points closest to the prediction point, and calculate the sample type with the largest variety. For example, the blue triangle shown in Fig 4 accounts for the largest proportion among the three points closest to the prediction point, then the green point is determined as a blue triangle

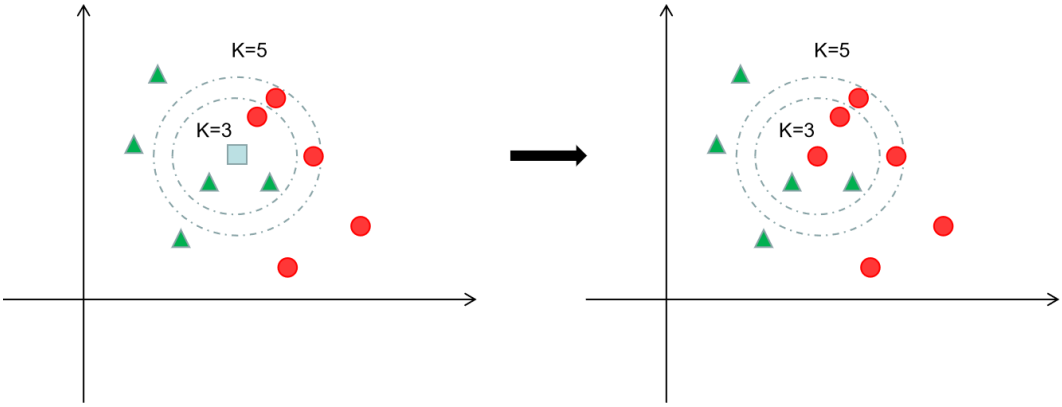

Fig. 4 Model diagram when K=5

When K=5, there are more red dot instances, and new instances are classified as red dot instances. From this example, it can be seen that the selection of K value has a great impact on the final effect of the model.

Set the n feature vector sets $\left(f_1^i, f_2^i, \cdots, f_n^i\right)$ of the ith training instance and the related target vector $\left(\mathbf{t_1}, \mathbf{t_2^i}, \cdots, \mathbf{t_m^i}\right)$ with m attribute values [8]. Let the similarity measure be Euclid distance, and the distance between the new instance and the ith training instance is calculated as follows:

$$\sqrt{\sum_{x=1}^{n}\left(f_x^i - q_x\right)^2} \tag{7}$$

Take the target vector of K nearest neighbors as $(\mathbf{t^1}, \mathbf{t^2}, \cdots, \mathbf{t^k})$, calculate its mean value and take the mean value as the target of the new instance:

$$\sum_{i=1}^{k} \frac{t^i}{k} \tag{8}$$

4.2. Decision tree

The algorithm divides attributes according to information gain. Similar to a tree structure, each internal node represents an attribute judgment, and each branch represents the output result of a judgment. The node with the maximum information gain is selected as the split node to form a decision tree.

Set S as the sample set, $s_i \in S$ $(i = 1, 2, 3, \ldots, s)$, and the category attribute as C_i $(i = 1, 2, 3, \ldots, m)$. Assuming s_i is the number of samples in category C_i, the information entropy contained in this set S is

$$\text{Entropy } A(S) = \sum_{i=1}^{k} \frac{|s_i|}{|S|} \text{Entropy } (s_i) \tag{9}$$

$p_i = s_i/S$ is the probability that any data object belongs to C_i. Suppose attribute A is used to divide the samples in set S, and there are k different values in attribute A, then the conditional entropy of attribute A to the sample set S is:

$$\text{Entropy } A(S) = \sum_{i=1}^{k} \frac{|s_i|}{|S|} \text{Entropy } (s_i) \tag{10}$$

$|s_i|$ and $|S|$ are the number of samples contained in and respectively. If attribute A is used to divide sample set S, the information gain Gain (A) is:

$$\text{Gain } (A) = \text{Entropy } (S) - \text{ Entropy } A(S) \tag{11}$$

In this paper, we use the above formula to calculate the information gain of clinical manifestations of Pregnancies, Glucose, BMI and DPF.

4.3. Simple Bayes

Naive Bayesian model is a "standard" probability statistical model based on probability theory and mathematical statistics [9]. The estimation, confidence interval and hypothesis test of relevant parameters can be obtained by using the methods and theories of classical statistics. However, Bayesian thought goes beyond this. It tries to express the uncertainty of parameters by using the probability distribution function θ of $p(\theta)$. Introducing Bayesian formula:

$$P(Y = c_k \mid X = x) = \frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \tag{12}$$

With Bayesian formula, there are still many things that cannot be calculated. Next, we introduce simple assumptions:

$$\begin{aligned} P(X = x \mid Y = c_k) &= P\left(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} \mid Y = c_k\right. \\ &= \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \mid Y = c_k\right) \end{aligned} \tag{13}$$

It can be known that the calculation formula of naive Bayesian model is:

$$P(Y = c_k \mid X = x) = \frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)},\ k = 1, 2, \cdots, K \tag{14}$$

As naive Bayes is used to classify tasks

$$y = f(x) = \arg\max_{c_k} \frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)} \tag{15}$$

Therefore, the naive Bayesian formula is:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (16)$$

4.4. Support Vector Machines

The support vector machine (SVM) algorithm is an algorithm that seeks to minimize the structural risk. It can effectively overcome the complex computing problems caused by high dimensions by using optimization methods to solve problems. At the same time, the optimization process of support vector machine is realized by solving a quadratic programming problem, which can ensure to find the global optimal solution.

The result value of the sample can be expressed as a data set of 0-1 variables. The separation vector can be obtained by calculating the position of the sample in space to correctly distinguish the categories of samples in the prediction set. Generally, support vector machine models can be divided into four categories: linear separable, linear nonseparable, nonlinear separable, and nonlinear nonseparable[10]. When the training data is linearly indivisible, the support vector machine can map the nonlinear feature generated by the kernel function to the input vector, map it to a higher dimensional feature space, and then establish the best classification hyperplane from the new space. The classification decision function can be obtained as

$$y = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \quad (17)$$

Where α_i is the Lagrange multiplier.

5. Comparative analysis of diabetes prediction models based on machine learning algorithm

5.1. KNN algorithm

For the four selected variables, Glucose has the highest correlation with the target results, with a correlation coefficient of 0.47, followed by the BMI index, but also has a correlation of 0.31. Because there are many factors disturbing the Glucose index, it is impossible to confirm whether all the Glucose data in the data set are measured on an empty stomach, so here we use BMI for univariate analysis. Import the BMI data into Jupyter, and the BMI distribution is shown in Fig 5.

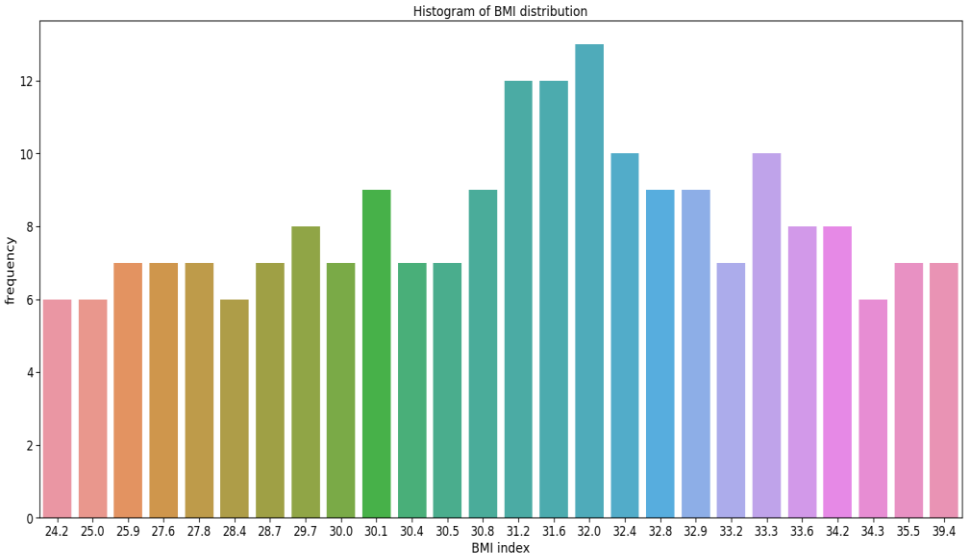

Fig 5. BMI Distribution Statistics

It can be clearly seen here that the number of people whose BMI index is around 32 is the largest. Then the statistical data can be obtained in Table 3.

Table 3. Description Statistics

slimmest	fattest	average BMI
18.2	47.1	32.45

It is known from the data in Table 4 that the overall BMI of the sample data test population is on the high side. It can be seen that the overall health level of the test population is on the low side. Then, the samples are divided into three categories according to the BMI index: lean (BMI ≤ 20), normal (20 ≤ BMI ≤ 25), and overweight (25 ≤ BMI). The results are shown in Fig 6.

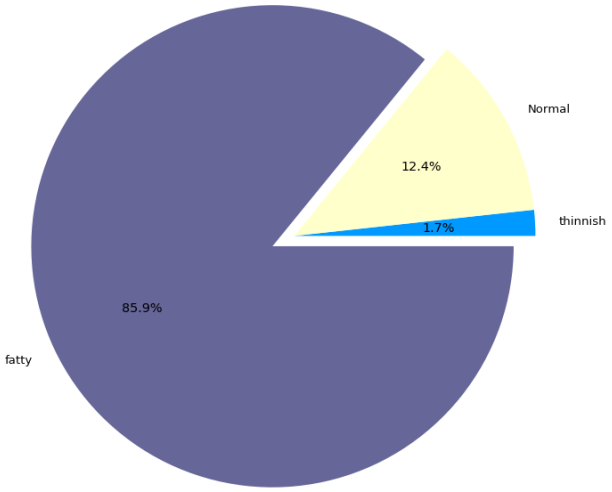

Fig 6. Proportion of patients under different BMI

It can be seen that the proportion of overweight people suffering from diabetes has reached 85.9%. It can be seen that the risk of obesity suffering from diabetes is the highest. The risk of diabetes in lean people is the smallest, only 1.7%. It is preliminarily inferred here that having a slim body can reduce the risk of diabetes.

For the test of prediction effect of KNN model, first of all, it is assumed that these four characteristics are equally important. To determine whether the relationship between model complexity and accuracy can be confirmed, the prediction effect is shown in Fig 7.

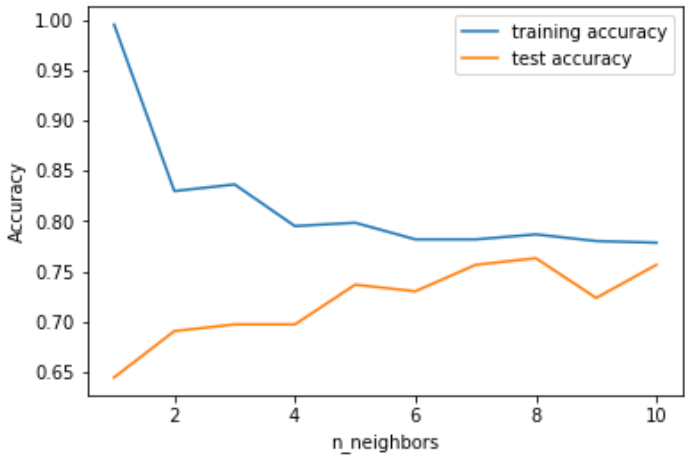

Fig. 7 Accurate Curve of Training Set and Test Set

Fig 7 uses 70% of the data as the training set and the remaining 30% as the test set. The figure shows that the precision of training and test set on y-axis is relative to n on x-axis_ Setting of neighbors. Considering that if a nearest neighbor is selected, the prediction on the training set is perfect. However, when more neighbors are considered, the training accuracy decreases, indicating that the use of a single nearest neighbor will cause the model to be too complex. The best performance is near 8 neighbors.

Finally, the KNN model was trained and tested, and the results are shown in Table 4:

Table 4. Result Analysis Table

Test set accuracy	Training set accuracy	Best parameters	Best estimator
0.7546	0.7912	8	KNeighborsClassifier(n_neighbors=8)

The accuracy of training set is 0.7546, and the best K value is 8.

6. Conclusion

The naive Bayesian model in this paper originates from the classical mathematical theory, has a theoretical basis, and the classification efficiency is stable. However, when using large-scale training sets, the number of features of each item is relatively small, and the algorithm is not sensitive to missing data, so the algorithm can obtain ideal results when the sample size is small. The decision tree algorithm does not rely on complex parameter optimization and specific training set segmentation, and has good compatibility, robustness and portability. However, when there are many decisions, the time complexity is high, and the results obtained are over fitting. Support vector machine has high accuracy, and the data is linearly indivisible in the original feature space. As long as an appropriate kernel function is given, the results will be ideal, which is convenient for solving the machine learning problem under small sample size. In this paper, four methods are used for prediction analysis. The KNN model performs well in the prediction. The naive Bayesian model trains and processes data quickly, and has a solid theoretical foundation. As the data in this paper has been preprocessed, the coupling degree between the data has been greatly reduced. The prediction effect of naive Bayesian algorithm and support vector machine is relatively ideal, while the prediction of decision tree model is not satisfactory.

References

[1] Lu Haoxuan, Xu Jinyan, and Cheng Cute Construction and comparison of heart disease prediction models based on multi factor regression analysis and machine learning algorithm [J] Journal of Ningbo University, 2022.5, 35 (3)

[2] Chen Yaming Research on disease prediction and prescription recommendation algorithm based on medical data analysis [D] Henan University of Science and Technology: 2019.5

[3] Liu Hongwei Research on risk prediction and online calculation of gestational diabetes based on machine learning algorithm [D] Tianjin Medical University: 2020.5

[4] Gan Ziqi, Pang Chen, Wang Wenli, Yan Wei Research on intelligent medical diagnosis system based on Jess fuzzy inference [J] Computer Technology and Development two thousand and nineteen

[5] Hong Ye Research on diabetes prediction model based on machine learning algorithm [D] Harbin University of Technology: June 2016

[6] Zhang Yan The construction of aging cardiovascular disease inpatient decline prediction model based on decision tree [D] Shantou University: Medical College, June 2021

[7] Lei Fei Research on Text Classification and Its Application Based on Neural Network and Decision Tree [D] University of Electronic Science and Technology: August, 2018

[8] Sun Huiran A decision tree based machine learning method and inference engine design [D] Chang'an University of Technology: April, 2014

[9] Liu Yuxiu, Jia Hong, Li Ailing Common methods and standards for screening and diagnosis of diabetes [J] Medical Review, 2005, 11 (12)

[10] GouldMK, HuangBZ, TammemagiMC, etal. Machinelearning for early lung cancer identification using routine