Research on Type 2 Diabetes Risk based on Lifestyle Factors

Weiyi Dong *

Chung-Ang University, Seoul, 06974, Korea

* Corresponding author: onlyovo@cau.ac.kr

Abstract. Diabetes is a global chronic disease, and the number of patients and medical expenses are increasing. It is predicted that by 2045, the number of adult diabetics in the world will reach 693 million, and the expenditure on health care will increase to 958 billion US dollars. Diabetes can be divided into two types: type 1 and type 2, of which type 2 diabetes accounts for 95% of all diabetes cases. In this study, principal component analysis and stochastic forest algorithm are used to evaluate the performance of the model by using large-scale survey data and ROC curve, which provides a new idea for diabetes prediction research. The results show that the stochastic forest algorithm performs well in dealing with high-dimensional data, while the principal component analysis method can reduce the dimension of high-dimensional data, reduce redundant features and improve the prediction ability of the model. Therefore, the method proposed in this paper can help medical researchers to make better use of diabetes data for research, and provide a new method for diabetes risk prediction.

Keywords: Diabetes, principal component analysis, random forest, machine learning.

1. Introduction

There are 415 million diabetics worldwide, andabout193 million undiagnosed [1]. Diabetes and its prediabetes account for 8% and 7% of the world's population, respectively [2]. Primary diabetes is divided into two types: type 1 and type 2 diabetes, with type 2 diabetes occurring more frequently and accounting for 95% of all cases of diabetes [3]. According to the forecast released by IDP in 2018, an estimated 451 million people aged 18 to 99 were living with diabetes worldwide in 2017, and by 2045, it is estimated that the number of diabetes patients will reach 693million [4]. Type 2 diabetes is also an extremely costly disease. Because type 2 diabetes can lead to microvascular and macrovascular complications, it is a huge challenge for patients and medical workers. It will not only harm the patient's body but also a serious burden on the mental health of healthcare workers. In other words, there is a certain negative impact on the health care system. Not only that, but diabetes affects individuals, economies and societies across the globe [2]. In 2017, total global healthcare expenditures for people aged 18-99 with diabetes were approximately $ 850 billion. Global healthcare spending for adults with diabetes will increase by 7% to $ 958 billion by 2045 [4]. Type 2 diabetes is an extremely serious chronic disease. According to the IDF Diabetes Atlas, diabetes will reach about 32 million people in 2021and 49 million by 2045 in South and Central America. Around the world, one person dies every five seconds from diabetes [5]. The prevalence of type 2 diabetes is rising rapidly, and many risk factors are associated with the pathogenicity of type 2 diabetes, such as the complex interaction between genes and the environment [6].

Type 2 diabetes results in the development of a variety of physiological abnormalities, most notably usually accompanied by varying degrees of insulin resistance. The most important factors currently known to cause type 2 diabetes are adipocyte insulin resistance and inflammation. More than 80 percent of patients with type 2 diabetes are overweight in the United States. The mechanism of insulin resistance caused by obesity is mainly due to improper diet or less exercise in patients, which leads to fat-derived hormones and related cytokines in the body, such as leptin, resistin, free fatty acids, etc. increased. Changes in these adipose cytokines not only affect fat storage and release, but also affect the sensitivity of human tissue to insulin, low-grade inflammatory response and abnormal blood coagulation [7]. The body mass index (BMI) is one of the most important indicators for predicting the risk of diabetes. Multiple data on American men and women show that the body mass index (BMI) and diabetes risk are strongly positively related, and the risk of diabetes increases

with increasing body mass index [8, 9]. Not only obesity will affect the prevalence of type 2 diabetes, but also other comprehensive metabolic components, such as hypertension and hyperlipidemia. According to research data from Italy and Greece, the incidence of type 2 diabetes in patients with high blood pressure is two to three times that of people with normal blood pressure [10]. A large amount of research evidence shows that having a healthy lifestyle is the most effective intervention to prevent and control diabetes. Such as maintaining a healthy weight, having a good diet, not smoking and drinking, and maintaining a certain amount of exercise every day [11]. According to some non-diabetic patients with fasting and elevated plasma glucose concentration after load, lifestyle changes are more effective in reducing the incidence of type 2 diabetes than related drug treatment [12].

This article analyzes the responses to health-risk behaviors, chronic health conditions and the use of prevention services collected annually by the behavioral risk factor surveillance system (BRFSS) for more than 400,000 Americans. The purpose is to focus on the risk factors affecting type 2 diabetes, predict and judge the important influencing factors that can be changed, and determine which factors can be prevented and intervened by improving lifestyle.

2. Method

2.1. Data Sources and Description

The main data of this study is the data set of diabetes health indicators from Kaggle official website, including 253,680 survey responses from cleared BRFSS 2015 + balanced dataset. The target variable diabetes binary has two classes. 0 means no diabetes and 1 means diabetes or pre-diabetes. This data set contains 21 characteristic variables that are balanced. This data set has 21 characteristic variables and is unbalanced.

2.2. Variable Selection

In the follow-up experiment and modeling, some variables are encoded, and the encoding variables and coding criteria are as Table 1 shows:

Table 1. Data information

Variable	Description	Type		
		0	1	2
Diabetes_012	-	Non-diabetes	Pre-diabetes	Diabetes
		No hypertension	Hypertension	
Cholesterol		No high cholesterol	High cholesterol	
Cholesterol test		No cholesterol test within 5 years 1	Cholesterol test within 5 years;	
Smoker	Smoked at least 100 cigarettes in your life?	No	yes	
Stroke	(once told) a stroke	No	Yes	
Heart attack	Coronary heart disease (CHD) or myocardial infarction (MI)	No	Yes	
Physical activity	Physical activity in the past 30 days-excluding work	No	Yes	
Fruit	Eat fruit once or more times a day	No	Yes	

2.3. Research Method

The main methods used in the following data processing and modeling in this study include principal component analysis and random forest, which will be introduced in this section. Among them, principal component analysis is used to extract components and select variables from selected variables. Random forest is used to select and predict the risk factors of type 2 diabetes, predict and

judge the important influencing factors that can be changed, and determine which factors can be prevented and intervened by improving lifestyle.

2.3.1. Principal component analysis (PCA)

Principal component analysis is a common method for dimension reduction, which extracts the main information from data by transforming the original data into a set of new, irrelevant and less dimensional variables. The main idea of PCA is to transform the related variables in the original data into a group of irrelevant linear variables. These linear variables are called principal components, and each principal component explains a part of the variance of the original data, so the original data can be replaced by fewer principal components, thus simplifying the data processing and modeling process.

In principal component analysis, data preprocessing is needed, including data cleaning, variable standardization and other steps. Then by calculating the covariance matrix or correlation coefficient matrix, the relationship matrix between the original variables is obtained. Then, through eigenvalue decomposition and eigenvector extraction, a group of principal components can be obtained, in which each principal component represents a comprehensive variable, which can explain the variance and covariance structure of the original variable. When selecting principal components, the first few principal components are generally selected, and the cumulative variance contribution rate of these principal components should reach a certain threshold, such as 70%, 80%, etc.

Principal component analysis is a common method of multivariate data analysis. Its main principle is to transform the original variables into a new set of linear comprehensive variables through linear transformation, so as to minimize the correlation between the new variables and retain most of the information in the original data.

2.3.2. Random forest (RF)

An ensemble learning system built on decision trees called random forest incorporates various decision trees for classification or regression tasks. Multiple decision tree models are trained using random selection of the data and characteristics, and the final prediction outcomes are then obtained through voting or averaging.

Random forest algorithm has high accuracy, robustness and generalization ability, and can avoid over-fitting problem. It is suitable for processing high-dimensional data, and can be used for classification and regression tasks. Random forest can also be used in feature selection and anomaly detection.

3. Results and Discussion

Diabetes is a significant chronic condition that impairs a person's capacity to control blood glucose levels and can shorten a person's life expectancy and quality of life. During digestion, many meals break down into sugar, which is then released into the blood. The pancreas receives a signal from this to release insulin. Insulin aids the body's cells in utilizing blood sugar for energy. In general, diabetes is characterized by insufficient insulin production or inefficient insulin use.

3.1. Descriptive Analysis

The Behavioral Risk Factor Surveillance System (BRFSS) annually gathers feedback data on health risk behaviors, chronic health issues, and utilization of preventative treatments from more than 400,000 Americans. It consists of 70,692 survey responses from CDC BRFSS 2015. This section will analyze and visualize the descriptive statistics of the data.

BMI: BMI in the data set is continuously distributed data, and the distribution of data is shown in the following Figure 1.

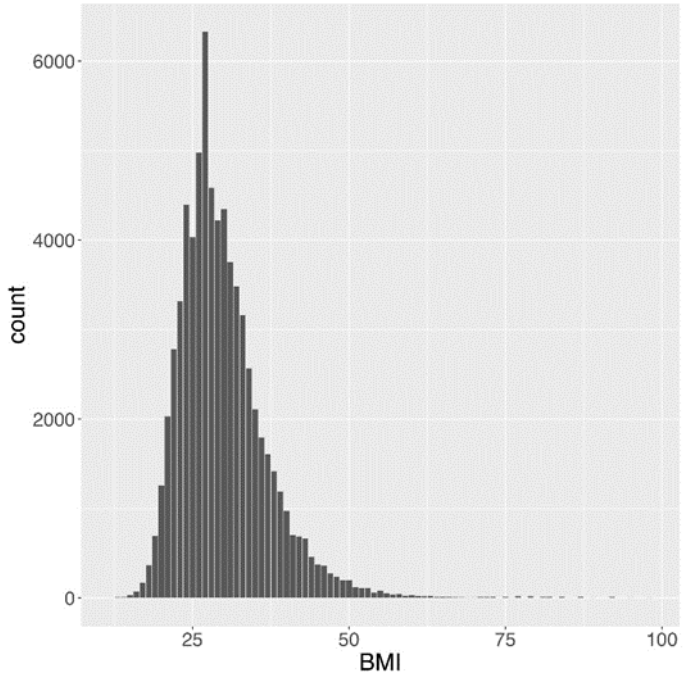

Figure 1. BMI density

It can be seen that the BMI distribution of the respondents basically presents a normal distribution. According to the statistical results of BMI distribution data, the minimum value is 12.00 and the maximum value is 98.00. The mean value is 29.86, and the quartiles are 25.00 and 33.00.

Age: Age in the dataset is discrete data, and the distribution of data is shown in the following Figure 2.

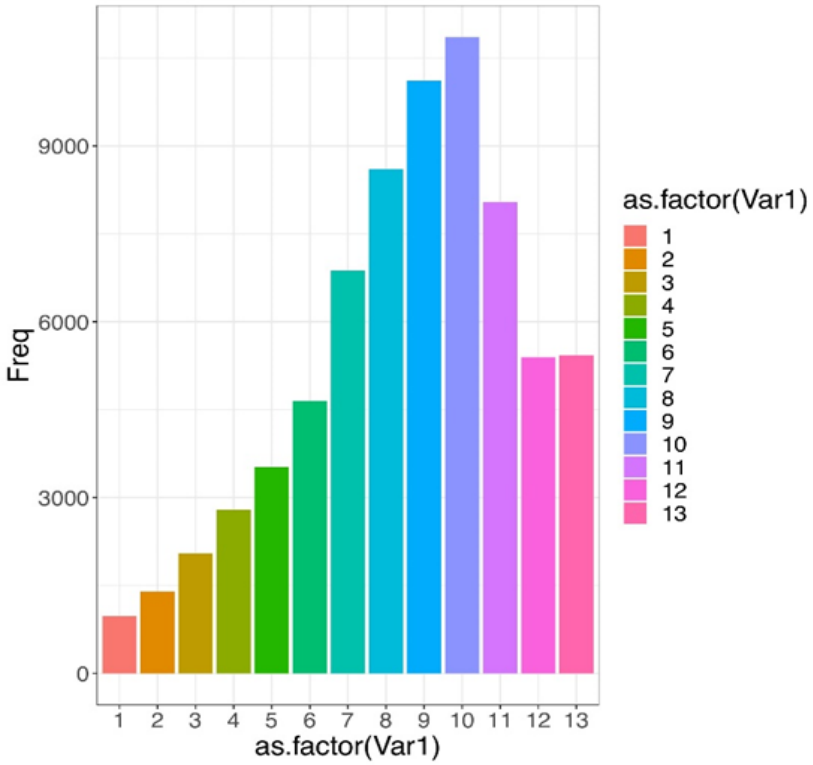

Figure 2. Age density

Dividing the age distribution into 13 intervals, it can be seen that the number of people distributed in interval 10 is the largest, followed by interval 9, interval 8 and interval 11. This reflects that the overall age distribution of the respondents is too large, and the middle-aged and elderly people are the majority samples of the respondents.

Education: The Education in the data set is discrete distributed data, and the distribution of data is shown in Figure 3.

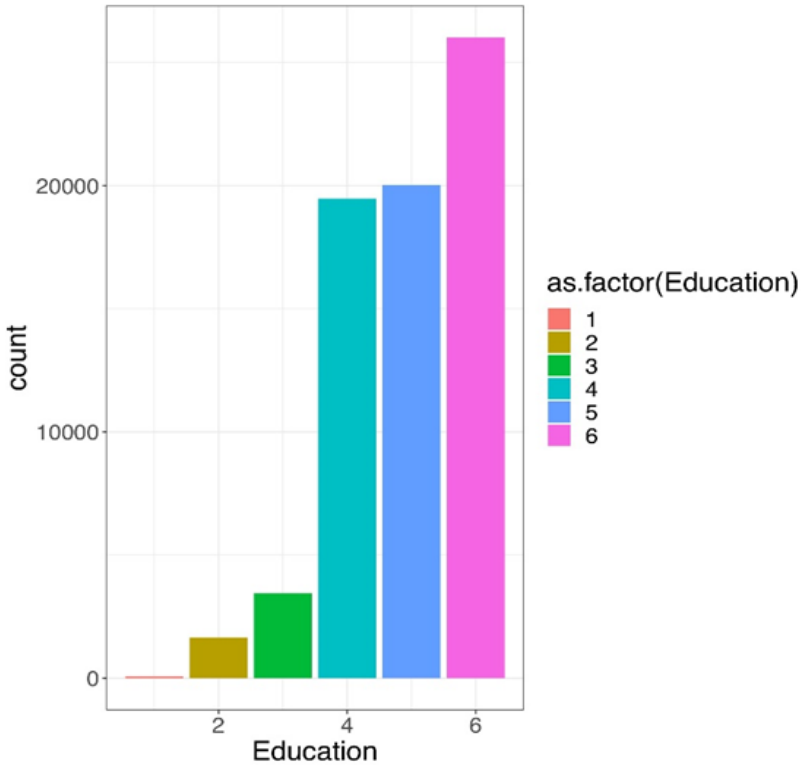

Figure 3. Education density

By dividing the educational level distribution of the respondents into six categories, we can find that the proportion of people with the highest educational level is the highest.

GenHlth: GenHlth in the data set is discrete data, and the distribution of data is shown in Figure 4.

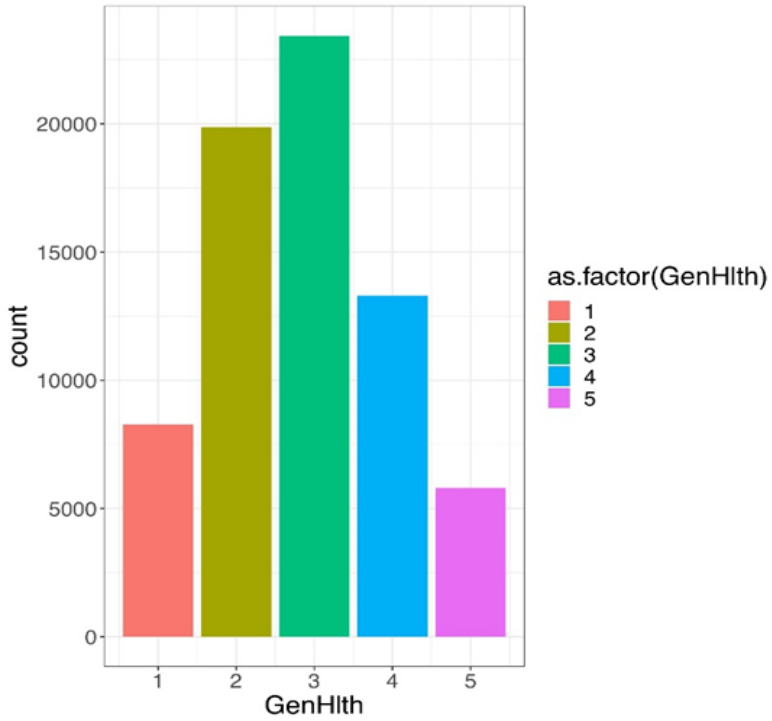

Figure 4. GenHlth density

By dividing GenHlth feature variables into five categories, we can find that the third category has the largest number, followed by the second category and the fourth category in turn.

Income: Income in the data set is discrete data, and the distribution of data is shown in Figure 5.

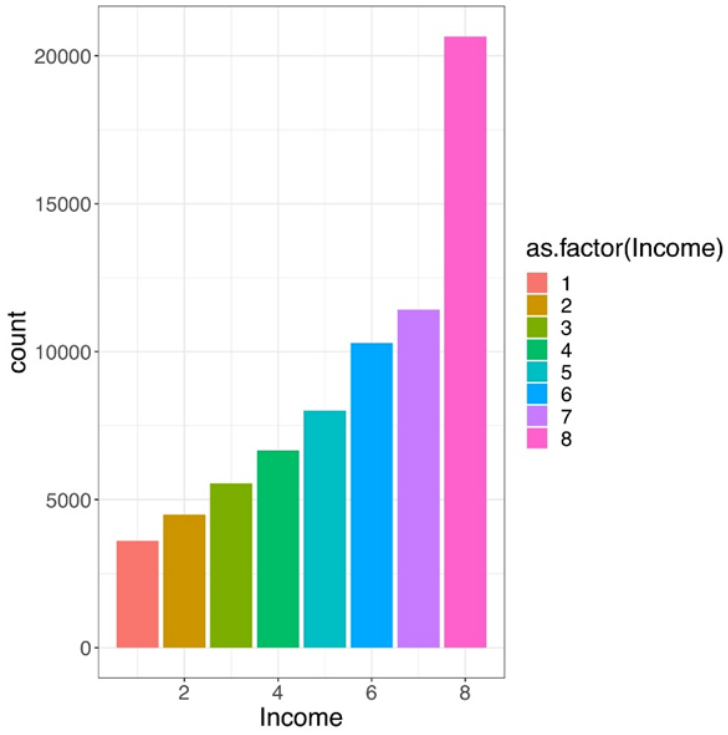

Figure 5. Income density

By dividing Income characteristic variables into 8 categories, we can find that the proportion of high-Income groups is higher.

MentHlth: MentHlth in the data set is discrete data, and the distribution of data is shown in figure 6.

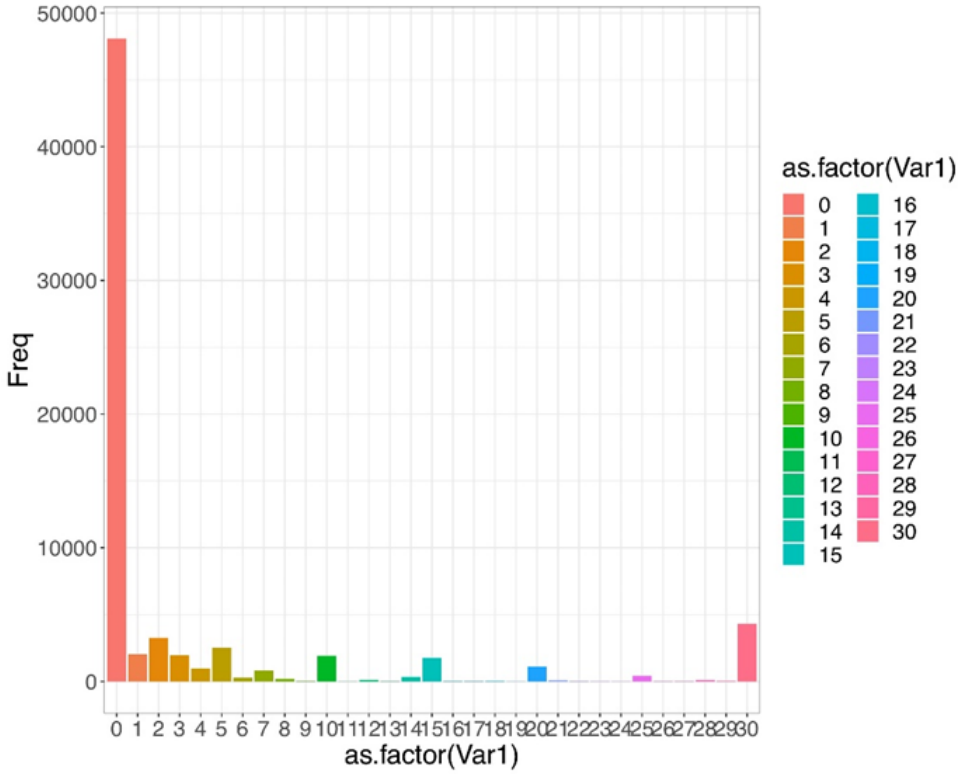

Figure 6. MenHlth density

By dividing MenHlth feature variables into 31 categories, we can find that the whole presents a two-stage distribution trend, and the first category has the highest proportion, far exceeding other categories.

PhysHlth: PhysHlth in the data set is discrete data, and the distribution of data is shown in figure 7.

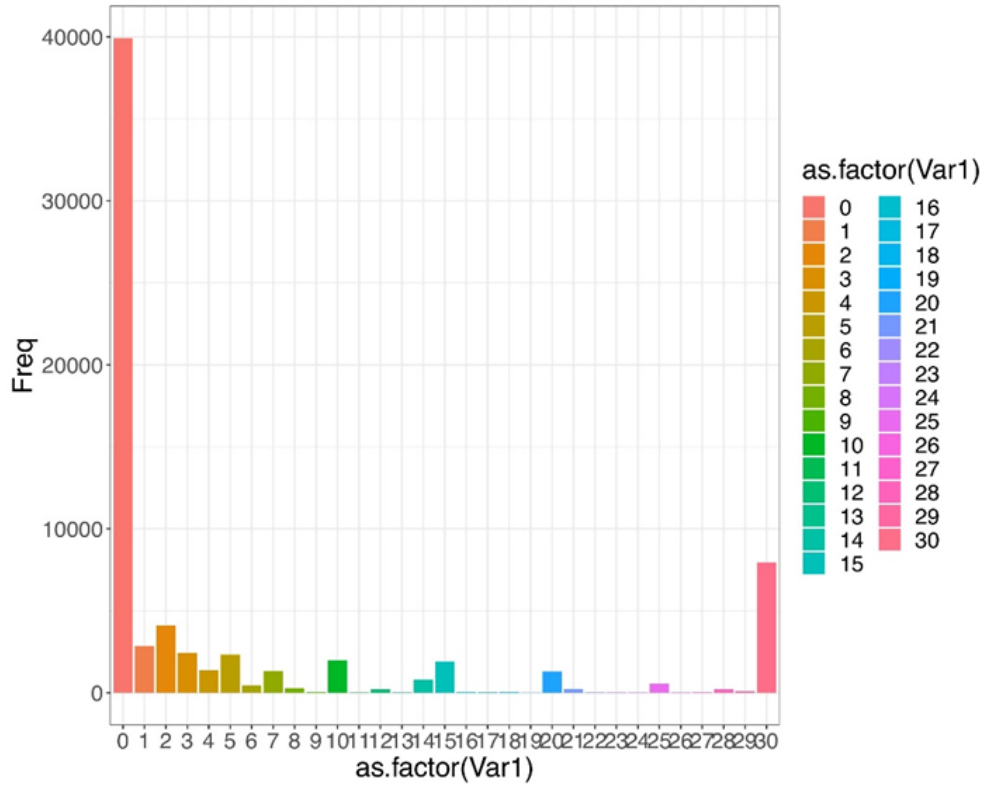

Figure 7. PhysHlth density

By dividing PhysHlth feature variables into 31 categories, we can find that the whole presents a two-stage distribution trend, and the first category has the highest proportion, far exceeding other categories.

Among all the respondents, those without diabetes and those with pre-diabetes or diabetes were equally divided. The proportion of each characteristic variable in the data set corresponding to the prevalence of respondents is shown in Table 2.

Table 2. Distribution of tags responding to feature variables

Variables	Label: 0	Label: 1
Diabetes_binary	35346	35346
HighBP	30860	39832
HighChol	33529	37163
CholCheck	1749	68943
Smoker	37094	33598
Stroke	66297	4395
HeartDiseaseorAttack	60243	10449
PhysActivity	20993	49699
Fruits	27443	43249
Veggies	14932	55760
HvyAlcoholConsumer	67672	3020
AnyHealthcare	3184	67508
NoDocbcCost	64053	6639
DiffWalk	52826	17866
Sex	38386	32306

As shown in the table 2, the target variable Diabetes_binary has two classes. 0 means non-diabetes, and 1 means pre-diabetes or diabetes. The data set has 21 characteristic variables and is balanced.

3.2. PCA Results

The purpose of PCA is to reduce the dimension and noise of data set through dimension reduction and variable selection, and improve the interpretability and modeling effect of data set. Principal component analysis can transform a large number of correlated variables into a group of irrelevant principal components, so as to better describe the structure and relationship of data. In this study, the researchers performed principal component analysis on 21 variables in BRFSS data set, and selected the indicators related to the risk factors of type 2 diabetes for subsequent analysis. In this study, PCA dimension reduction analysis was done according to Diabetes_binary. According to the variance of each eigenvalue in the data and the contribution rate of variance, the gravel diagram is drawn as follows:

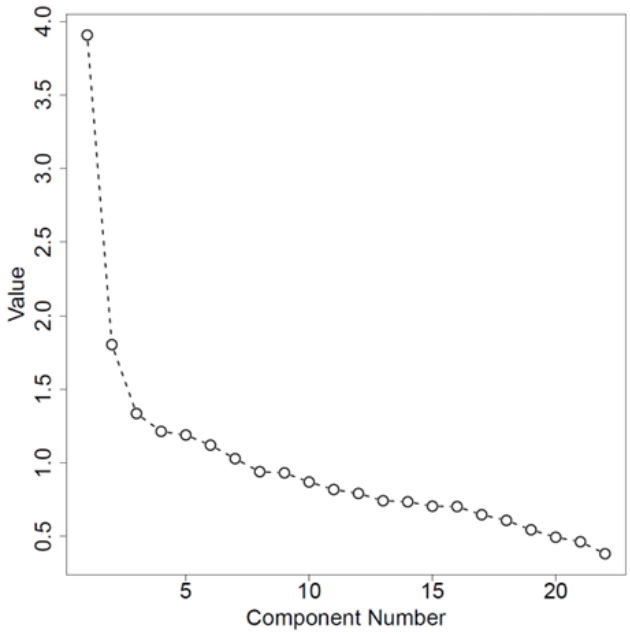

Figure 8. Gravel diagram

As illustrated in Figure 8, the Cattell gravel test is employed here to represent the relationship between eigenvalues and the number of primary components. According to the general principal, the number of maintained principal components should have an eigenvalue more than 1 and greater than the eigenvalue of parallel analysis. As can be seen from the above figure, the eigenvalue of Component Number 7 is greater than 1 and greater than the eigenvalue of parallel analysis. However, the proportion of the eigenvalue with Component Number 2 is much larger than that of the last five eigenvalues, so choosing two principal components can retain most of the information of the data set. This result shows that in PCA dimensionality reduction analysis of diabetes data set, choosing two principal components can retain most of the information of the data set. Among them, the first principal component has the highest degree of interpretation to the data set, which shows that the first principal component can explain most of the variance in the data set. However, the second principal component has a low degree of explanation, but it can also provide certain information. At the same time, through drawing the gravel map, it can be seen that with the increase of the number of principal components, the variance that each principal component can explain gradually decreases, but the proportion of the total variance retained gradually increases, which shows that each principal component provides different information, and the selected principal component number should comprehensively consider the balance between interpretation ability and retained information. The next step is to select the corresponding principal components.

Figure 9. PCA diagram

As shown in Figure 9, PC1 (17.76%) and PC2 (8.2%) can explain 25.96 of these 21 variables. This result shows that PC1 and PC2 can explain 25.96% of the variance of these 21 variables after PCA dimension reduction, that is, most of the information in the data set can be captured by using these two principal components. Among them, PC1 has the highest degree of interpretation of data sets, while PC2 has a relatively low degree of interpretation, but it also provides certain information.

It should be noted that although PC1 and PC2 can explain most of the variance, it does not necessarily mean that these two principal components are the best choice. In practical application, it is necessary to consider the specific data set and the application purpose, and comprehensively consider the balance between the number of principal components and the interpretation ability, and choose the most suitable number of principal components.

3.3. Random Forest Results

The second step is to build a model based on the data set to distinguish the two categories of Diabetes binary. Model optimization and screening using random forest algorithm is a commonly used machine learning algorithm. Its basic idea is to build multiple decision tree models and get the final prediction results by voting or averaging. Random forest algorithm can effectively avoid over-fitting problem, and has high prediction accuracy and stability.

In the process of model optimization and screening, we use 50-fold cross-validation method, which is a commonly used model evaluation method. Its main idea is to divide the data set into k subsets, select one subset as the validation set in turn, and the other k-1 subsets as the training set for model training and evaluation, repeat k times, and finally take the average value of k results as the performance index of the model. 50-fold cross-validation can effectively avoid over-fitting and under-fitting problems of the model, and improve the generalization ability of the model. Specifically, in this study, the data set is divided into training set (70% samples) and validation set (30% samples), and the random forest algorithm is used to optimize and filter the model. The ntree parameters of random forest are 500-1000 step as 100, and Mean Decrease Accuracy 0-100 step as 20. The model is optimized and screened by using 20 times 50-fold cross-validation method, and finally the model with the largest average acc of 50-fold cross-validation is selected for the final modeling.

In this study, indicators related to risk factors of type 2 diabetes were selected after PCA analysis, such as GenHIth (general health), BMI (body mass index), AgeHighBP (hypertensive age), HighChol (high cholesterol), CholCheck (cholesterol test), Income (Income), HeartDiseaseorAttack (heart

disease or heart attack), HvyAlcoholConsumer (excessive drinking), DifWalk (walking difficulty), PhysHith (physical health), Stroke (Stroke), Sex (gender) and Education (Education level). These indicators reflect the health status, lifestyle, socio-economic status and other information of the population, and have a high correlation with the risk factors of type 2 diabetes, which is an important indicator for follow-up analysis. The feature vectors included in the model and their importance ranking are as follows:

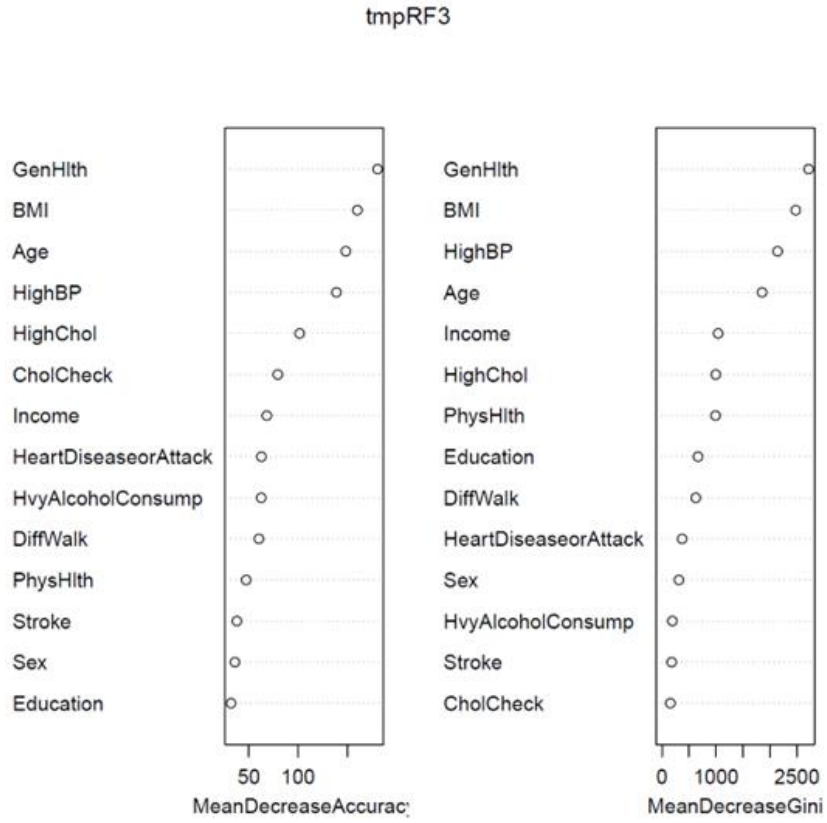

Figure 10. Random Forest diagram

As shown in Figure 10, different feature sorting can get different results. Feature ranking refers to ranking according to the importance of features in the algorithm. MeanDecreaseAccuracy and MeanDecreaseGini are two different methods to calculate the importance of features. The results show that these characteristics are important for the prediction of diabetes. BMI, GenHIth and AgeHighBP are the top three characteristics of the two methods, which is consistent with the common risk factors of diabetes. Other characteristics, such as HighChol, HeartDiseaseorAttack, DifWalk, PhysHith, Income, etc., are also considered as risk factors for or associated with diabetes. Characteristics such as gender, educational background and cholesterol test rank lower in different rankings, which may be due to their relatively weak relationship with diabetes. It should be noted that the feature ranking obtained here is only the importance ranking in random forest algorithm, and cannot be directly used as the basis for feature selection. Feature selection also needs to combine practical application scenarios and feature engineering methods.

The ntree parameter selection of the random forest of the final model is 800, and the Mean Decree Accuracy selection is 20, and all the training set samples are used for modeling. This result shows that the model established by random forest algorithm can well distinguish two kinds of samples in diabetes data set, and the accuracy rate (acc) of training set is 0.82, AUC is 0.91, and the accuracy rate (acc) of verification set is 0.75, AUC is 0.82. Among them, AUC is a common index to evaluate the performance of binary classification model. The larger the AUC value, the better the performance of the model. Therefore, the results of the model are good, and it can distinguish the two categories of Diabetes_binary well.

ROC curve is used to evaluate the performance of the model. ROC curve is a graphical display method that comprehensively considers the sensitivity and 1-specificity of the model. On the ROC curve, the abscissa is 1-specificity and the ordinate is sensitivity. The closer the ROC curve is to the upper left corner, the better the performance of the model. The ROC curve on the training set is shown in figure 11.

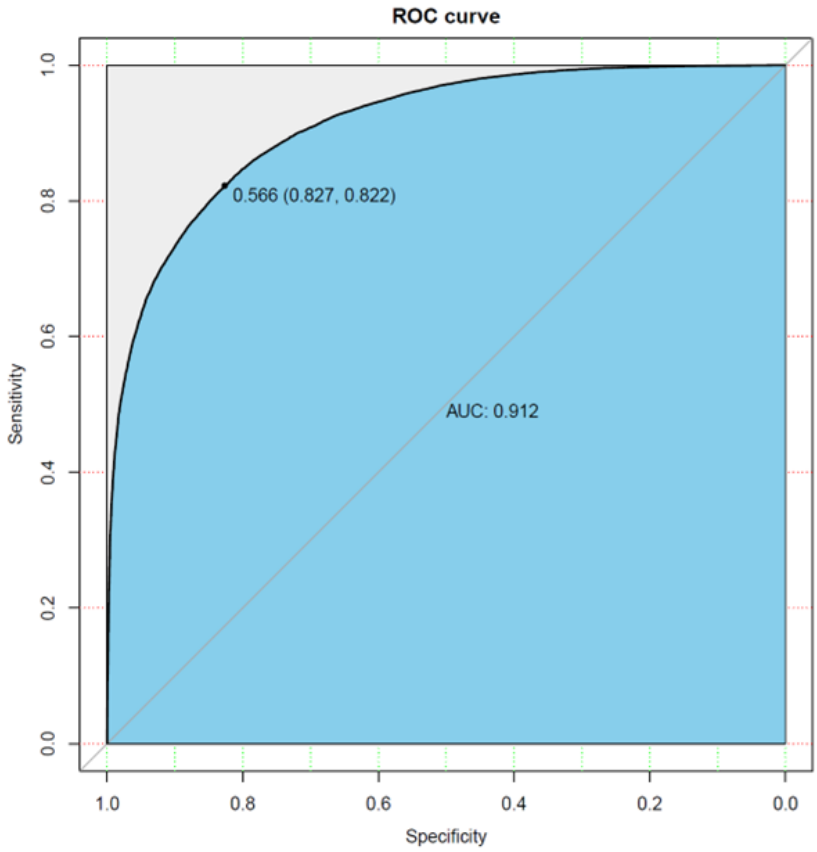

Figure 11. ROC curve of training set

As shown in Figure 11, there is a point 0.566 (0.827, 0822) on the ROC curve on the training set, which represents a specific position of the model on the ROC curve on the training set, and its corresponding coordinate is (0.827, 0.822). Among them, 0.827 is the True Positive Rate (TPR) of the model on the training set, also known as Sensitivity, which indicates the proportion of diabetic patients correctly predicted by the model. 0.822 is the False Positive Rate (FPR) of the model on the training set, also known as 1-Specific, which indicates the proportion of non-diabetic patients who were wrongly predicted as diabetic by the model. In this data, the point of 0.566 falls above the ROC curve, which shows that the performance of the model on the training set is not bad, but it is not optimal.

It should be noted that this point only represents the performance of the model on the training set, and cannot be simply used to evaluate the performance of the model, because the performance of the model on the verification set may be different. Therefore, it is necessary to comprehensively evaluate the performance of the model according to the performance of the verification set. The ROC curve on the validation set is shown in figure 12.

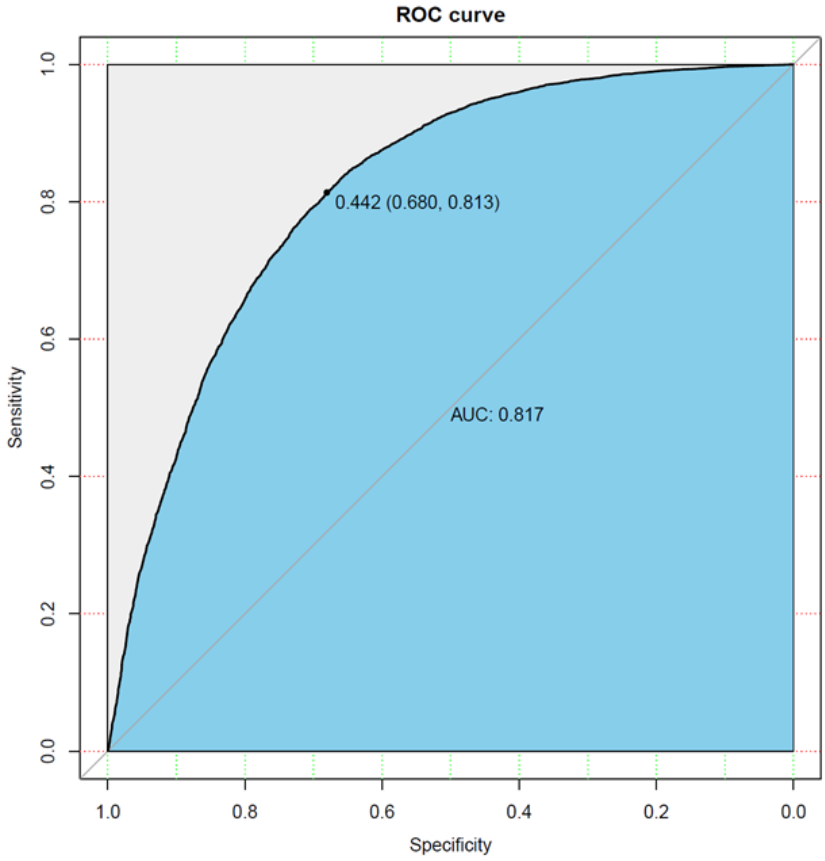

Figure 12. ROC curve of validation set

As shown in Figure 12, there is a point of 0.442 (0.680, 0.813) on the roc curve of type 2 diabetes on the validation set, and the data of this point (0.442, 0.680, 0.813) indicates that when the sensitivity rate is 0.442, the corresponding false positive rate is 0.680, and the AUC is 0.813. It can be found that the true positive rate at this point is relatively low, but the false positive rate is not particularly high. The overall AUC value is about 0.8, which shows that the classification ability of the model is more accurate, but misjudgment may occur in some cases.

Moreover, the selection of parameters in stochastic forest model in this study has undergone 50% cross-validation and the selection of average acc, which is scientific and reliable to some extent. At the same time, the model is stable in training set and verification set, which shows that the model has certain generalization ability.

4. Conclusion

In this study, the risk factors of type 2 diabetes were predicted and judged by collecting large-scale survey data of behavioral risk factor monitoring system (BRFSS). Specifically, this study selected the indicators with high influence on the risk factors of type 2 diabetes by principal component analysis, and established the prediction model by using random forest algorithm. The results of this study provide certain reference value for the prevention and intervention of type 2 diabetes.

This study uses modern statistical analysis techniques such as principal component analysis (PCA) and random forest algorithm to predict diabetes risk. The combination of PCA and random forest optimizes feature selection and modeling. PCA reduces over-fitting and improves the generalization performance of the model by removing redundant variables and noise. The method can help medical researchers better utilize diabetes data for risk prediction, and may also be applied to other diseases. PCA simplifies the data structure, making it easier to train various machine learning algorithms. The resulting model can improve the accuracy and reliability of diabetes diagnosis and treatment. And this study uses the random forest algorithm to predict the incidence of type 2 diabetes by establishing a stochastic forest model using 21 indexes from a diabetes data set. The model is trained on a training

set and evaluated on a verification set, with accuracy, AUC and ROC curves used as evaluation indexes. The results show that the model has good prediction performance and can distinguish the incidence of type 2 diabetes. However, the model's prediction effect may not be stable enough and requires further optimization. Using random forest algorithm can help doctors better understand the disease situation and formulate accurate treatment plans and preventive measures.

However, this study has limitations, including sampling error and self-report bias, which may affect the accuracy and reliability of the results. Moreover, some potential influencing factors such as genetic and environmental factors were not considered, which could impact the risk of type 2 diabetes. Future studies need to consider more factors and use various methods for analysis and evaluation, enlarge the data set, and increase the number of samples. Feature selection methods like Lasso regression and ridge regression could be explored to improve the reliability and generalization ability of the model.

References

[1] Chatterjee S, et al. Type 2 diabetes. The lancet, 2017, 389 (10085): 2239 - 2251.

[2] Galaviz K I, et al. Lifestyle and the prevention of type 2 diabetes: a status report. American journal of lifestyle medicine, 2018, 12 (1): 4 - 20.

[3] Wang Qiuyue, Liu Guoliang. Living habits and diabetes. Chinese Journal of Physician Training, 2000, 23 (9): 8 - 9.

[4] Edition I D. International Diabetes Federation. IDF Diabetes Atlas, 8th edn. Brussels, Belgium: International Diabetes Federation, 2017.

[5] Kyrou I, et al. Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: a narrative review with emphasis on data from Europe. BMC endocrine disorders, 2020, 20: 1 - 13.

[6] DeFronzo R A. Pathogenesis of Type 2 Diabetes mellitus. Medical clinics, 2004, 88 (4): 787 - 835.

[7] Chan J M, et al. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. Diabetes care, 1994, 17 (9): 961 - 969.

[8] Colditz G A, et al. Weight gain as a risk factor for clinical diabetes mellitus in women. Annals of internal medicine, 1995, 122 (7): 481 - 486.

[9] Melidonis A, et al. Increased prevalence of diabetes mellitus in a rural Greek population. Rural and Remote Health, 2006, 6 (1): 1 - 8.

[10] Zhang Y, et al. Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies. Diabetologia, 2020, 63 (1): 21 - 33.

[11] Knowler W C, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med, 2002, 346 (6): 393 – 403.