Multifactor Analysis Revealing Key Factors of Chronic NCDS Based on Random Forest Models

Jinzhu Cai [1, #], Ziying Zhao [2, #], Xinzhe Zheng [1, #, *], Qiulin Feng [3], Jie Deng [4], Yu Zhang [1]

[1] School of Materials Science and Engineering, Jiamusi University, Jiamusi City, Heilongjiang Province 154007

[2] School of College of Pharmacy, Jiamusi University, Jiamusi City, Heilongjiang Province, 15407

[3] School of Academy of music, Jiamusi University, Jiamusi City, Heilongjiang Province, 154007

[4] School of Computer Science and Technology, Zhuhai University of Science and Technology, Zhuhai City Guangdong Province, 519090

* Corresponding Author Email: 2199915298z@sina.com

[#]These authors contributed equally

Abstract. This article investigates chronic non-communicable diseases with a focus on lifestyle and dietary habits. By establishing separate models, eliminating the influence of highly correlated variables, and addressing data imbalance using the SMOTE algorithm, seven independent variables were identified, including basic information, lifestyle, and dietary habits. Random forest algorithm analysis revealed that smoking and alcohol consumption significantly impact the occurrence of chronic diseases, with different dietary factors associated with specific diseases. Occupational type, work intensity, and stress also have a notable influence on the risk of chronic diseases. Furthermore, increasing daily physical activity is associated with a lower disease risk. These findings contribute to a better understanding of the occurrence and management of chronic diseases, providing valuable information for prevention and treatment.

Keywords: Chronic Disease, Smote Algorithm, Random Forest.

1. Introduction

Since the 21st century, China has undergone rapid economic growth, achieving significant milestones, including being the world's second-largest economy by GDP and having a global influence through "Made in China."[1] This development has led to improvements in material and cultural living standards, advancements in technology and healthcare, and an increase in the average life expectancy of residents [2]. However, accompanying economic growth are formidable health challenges, primarily characterized by the threats posed by chronic non-communicable diseases and injuries to residents' health [3].

Chronic diseases such as cardiovascular diseases, diabetes, and malignancies have become major health concerns, closely linked to unhealthy lifestyles, poor dietary habits, insufficient physical activity, tobacco and alcohol abuse, and changes in dietary patterns. The health status of urban residents is influenced by various factors, including age, working conditions, and lifestyle. [4]

Therefore, there is a need for more comprehensive and in-depth research to understand the health of Chinese residents and the factors influencing it. [5-8]Consequently, we aim to investigate the correlation between common chronic diseases and lifestyle habits, work nature, exercise routines, and other factors. We seek to clarify the extent of influence between these variables. Using secondary data from the Palestinian Central Bureau of Statistics, H.F. Abukhdeir and his team found significant differences in the prevalence of diabetes, hypertension, cardiovascular disease and cancer among Palestinians across different demographic variables [9]

However, we found that their study did not take into account the internal correlation of variables, so we further explored this basis. Initially, we explore the correlation between common chronic diseases and lifestyle habits, work nature, exercise routines, and other factors to assist in selecting

relevant variables. We then analyze the prevalence of common chronic diseases, resulting in two sets of binary variables. Given the imbalanced numerical proportions in the samples, we opt to employ the SMOTE algorithm for oversampling. Subsequently, we establish random forest models to analyze the importance of various indicators and examine the degrees of correlation between common chronic diseases and each independent variable.

2. Methods

2.1. Correlation study

We have found that there is a certain correlation between lifestyle and dietary habits. Therefore, independently analyzing their impact on chronic diseases may lead to some errors. Hence, we need to remove some highly correlated lifestyle and dietary habits.

To explore these relationships, we appropriately model lifestyle and dietary habits and perform correlation analysis using an appropriate correlation coefficient based on a normal distribution test.

Pearson Correlation Coefficient: The Pearson correlation coefficient is a statistical measure calculated from sample data that quantifies the strength of a linear relationship between two variables. Sometimes, it is also referred to as the product-moment correlation or moment-product correlation. Assuming there is a matrix$X = (x_{ij})_{m\times n}$, the Pearson correlation coefficient between two variables can be calculated using the following formula:

$$\rho(a,b) = \frac{\sum_{i=1}^{m}(x_{ai} - \overline{x_a})(x_{bi} - \overline{x_b})}{\sum_{i=1}^{m}(x_{ai} - \overline{x_a})^2 \sum_{i=1}^{m}(x_{bi} - \overline{x_b})^2} \tag{1}$$

Where m is the length of each column, and the coefficient's result falls within the range [-1, 1].

Spearman Correlation Coefficient: The Spearman correlation coefficient is defined as$rho(a,b)$, where d represents the difference in ranks for each column, and m represents the length of each column:

$$rho(a,b) = \frac{6\sum d^2}{m(m^2-1)} \tag{2}$$

Lifestyle Habits: We primarily consider smoking habits, and the questionnaire provides information about smoking duration, average weekly consumption, and frequency per smoking session, as well as data on passive smoking. Based on this, the description of smoking habits is categorized in several directions: smoking or not, types of smokers, etc. The provided data include the birth year of the samples, which provides limited insight. According to the provided questionnaire, it is derived from the "Epidemiological Survey of Chronic Non-Communicable Diseases and Related Risk Factors in Shenzhen" conducted by the Shenzhen Health Bureau in June 2009. Therefore, the age of each sample can be calculated based on their birth year, using 2009 as the reference year [7-10]:

$$age_i = 2009 - born_i \tag{3}$$

Where age_i represents the age of the ith sample, and $born_i$ represents their birth year.

Smoking Status: Define it as a numerical variable related to whether the respondent currently smokes, where 1 represents non-smokers, 2 represents former smokers, and 3 represents current smokers. The smoking status can be treated as a continuous variable since it shows a gradual increase in smoking intensity.

Types of Smokers: For smokers, we have defined several types, considering the smoking behavior definitions recommended by UNICEF and combining them with smoking indices as follows:

$$\lambda_i = day_{si} \times year_{si} \tag{4}$$

Where λ_i represents the smoking index of the individual in the ith sample,day_{si} represents the average number of cigarettes smoked per day, and $year_{si}$ represents the smoking duration, and the parameters for types of smokers are shown in the Table.1.

Table 1: Parameters for Types of Smokers.

Smoker Type	Index Parameters
Light Smoker	Smoking Index < 300
Moderate Smoke	300 < Smoking Index < 500
Heavy Smoker	Index > 500

Passive Smoking: Passive smoking refers to inhaling secondhand smoke, primarily related to the surrounding environment. Passive smoking may be correlated with certain factors to some extent. We define the passive smoking coefficient as the daily frequency of passive smoking, which is the ratio of the number of days of passive smoking per week to the total number of days:

$$PS_i = \frac{ps_i}{7} \times 100\% \tag{5}$$

Where ps_i represents the weekly frequency of passive smoking for the ith sample, and PS_i represents the daily frequency.

Physical Activity: Further exploring the relationship between physical activity and various factors, we define a physical activity index based on questionnaire data, which is the product of exercise intensity and exercise duration. Since exercise intensity is categorized as either moderate or high, with moderate being 1 and high being 2, the physical activity index is as follows:

$$Sport_i = degree_i \times time_i \tag{6}$$

Where $degree_i$ and $time_i$ represent the exercise intensity and exercise duration of the ith sample, respectively.

Dietary Habits: Dietary habits are primarily based on the model for Question 1. We start by exploring the relationships between dietary habits and age, gender, ethnicity, education level, marital status, and occupation, focusing on the coefficients for eating patterns, the intake of various foods, and the consumption of oil, salt, alcohol, and sugar.

Similar to Question 1, the definition of the eating pattern coefficient is as follows:

$$\mu_i = \frac{b_i + l_i + d_i}{3 \times 7} \tag{7}$$

Where μ_i represents the eating pattern coefficient of the ith sample, and$b_i + l_i + d_i$ represents the number of regular meals consumed within a week.

The intake of different foods is mainly related to intake rates and the quantity consumed each time. The intake rate is defined as follows:

$$\eta_{ij} = \begin{cases} 1 \times 100\% \text{ the daily intake is not none} \\ \frac{a_{ij}}{7} \times 100\% \text{ the daily intake is none, the proportion it occupies within a week is } a_{ij} \\ \frac{b_{ij}}{30} \times 100\% \text{ the daily intake is none, the proportion it occupies within a month is } b_{ij} \end{cases} \tag{8}$$

Where η_{ij} represents the intake ratio, which represents the intake frequency of the jth type of food by the ith questionnaire target within a unit time. When the daily intake frequency is non-zero, it indicates daily consumption of that food, and the ratio is 100%. When the daily intake frequency is zero but the weekly intake frequency is non-zero, the corresponding ratio a_{ij} is the proportion it occupies within a week (7 days). When the daily intake frequency is zero but the monthly intake frequency is non-zero, the corresponding ratio b_{ij} is the proportion it occupies within a month, set as its proportion within 30 days.

The intake quantity is related to the quantity consumed each time and unit conversion. Assuming 1 liang = 50g, 1 cup = 350g, 1 spoon = 5g, and 1 egg = 50g, the intake quantity is calculated as follows:

$$I_{ij} = \eta_{ij} \times \gamma \times e_{ij} \tag{9}$$

Where I_{ij} represents the total intake of the jth type of food by the ith questionnaire target, η_{ij} represents the intake ratio, e_{ij} represents the average intake quantity (inconsistent units), and γ represents the food unit conversion coefficient, which is the number of grams per unit of the jth food.

Table 2. Table of the relationship between living habits and influencing factors.

Factor	**Smoker Type (Pearson)**	**Smoker Type (Spearman)**	**Smoking Index (Pearson)**	**Smoking Index (Spearman)**
Age	-0.049	-0.027	-0.019	-0.056
Gender	-0.579	-0.568	-0.500	-0.426
Ethnicity	-0.003	-0.001	-0.002	-0.015
Education Level	0.006	0.017	0.004	-0.027
Marital Status	0.005	-0.015	-0.008	-0.004
Occupation	-0.161	-0.161	-0.142	-0.117
Factor	Passive Smoking (Pearson)	Passive Smoking (Spearman)	Exercise Coefficient (Pearson)	Exercise Coefficient (Spearman)
Age	0.004	0.011	-0.121	-0.172
Gender	-0.212	-0.193	-0.062	-0.017
Ethnicity	-0.016	-0.022	-0.009	-0.013
Education Level	-0.021	-0.036	0.127	0.068
Marital Status	-0.011	-0.019	-0.042	0.029
Occupation	-0.106	-0.097	0.085	0.074

For lifestyle habits, from Table 2, it can be observed that among factors related to smoking, gender has the most significant impact. The likelihood of smoking is higher in males compared to females. Age and occupation follow, with both correlation coefficients indicating a negative correlation between age and occupation with smoker types. In other words, as age increases, the probability of being a smoker decreases. Conversely, the negative impact coefficient of ethnicity is relatively small. Lastly, education level and marital status show a weak positive correlation.

Regarding the smoking index, gender is also the most significant factor, showing a negative correlation. Male smokers tend to have higher smoking indices than female smokers. Next in importance are age and occupation, both having a substantial negative impact on the smoking index? Ethnicity and marital status have weaker negative impacts, while education level, being ordinal data, is assessed using the Spearman correlation coefficient, and it shows a negative correlation. As education level increases, the smoking index tends to decrease.

Passive smoking is weakly positively correlated with age and negatively correlated with other factors. Gender has the most significant impact, followed by occupation, ethnicity, education level, and marital status, with the weakest correlation.

The exercise coefficient is mainly positively correlated with occupation and ethnicity. A higher occupational category number implies more leisure time, leading to a higher exercise coefficient. Non-Han ethnicities tend to have better exercise habits than Han ethnicity. Conversely, it is negatively correlated with age, gender, education level, and marital status.

For dietary habits, the eating pattern coefficient shows a positive correlation with gender, marital status, occupation, and education level, and the correlations are relatively strong. However, it exhibits a negative correlation with age, and its impact on ethnicity is relatively weak. In other words, the eating pattern coefficient is more closely related to the lifestyle factors of the samples, specifically their family and employment situations.

2.2. Variables determination

Based on our previous analysis, we are now beginning to formulate indicators while considering the correlations between factors. This paper aims to establish a multifactorial model, and we first design and define the dependent and independent variables.

Dependent Variables:

First, we process the chronic disease parameters in the data. For hypertension, we use the international measurement standards, where systolic blood pressure greater than 139 mmHg or diastolic blood pressure greater than 89 mmHg is considered hypertension. Since the questionnaire provides data on both systolic and diastolic blood pressure, we distinguish whether an individual has hypertension as follows:

$$H_i = \begin{cases} 1 & Ht_i = 1 \text{or } H_{si} > 139 \text{ or } H_{di} > 89 \\ 0 & \text{etc.} \end{cases} \tag{10}$$

Where H_i represents whether an individual has hypertension (a binary variable), H_{si} represents the systolic blood pressure of the ith sample, and H_{di} represents the diastolic blood pressure of the ith sample. Additionally, for individuals who have already been diagnosed with hypertension, their history of hypertension is considered. Ht_i This variable indicates whether they have ever been diagnosed with hypertension, where 1 represents having the condition, and 2 represents not having it.

Similarly, for fasting blood sugar data, we have an evaluation indicator. When fasting blood sugar is greater than 7 mmol/L, it is defined as diabetes as follows:

$$D_i = \begin{cases} 1 & Dt_i = 1 or \ D_{bi} > 7 \\ 0 & etc. \end{cases} \tag{11}$$

Where D_i represents whether an individual has diabetes (a binary variable), and D_{bi} represents the fasting blood sugar level of the ith sample. Just like with hypertension, we also consider the history of diabetes. Dt_i Represent the history of hypertension, where 1 represents having the condition, and 2 represents not having it.

Independent Variables:

Smoking: The data related to smoking is defined as the smoking index used in the paper:

$$\lambda_i = day_{si} \times year_{si} \tag{12}$$

Where λ_i represents the smoking index of the ith sample, day_{si} represents the average number of cigarettes smoked per day, and $year_{si}$ represents the smoking duration.

Alcohol Consumption: The data related to alcohol consumption is defined as the daily average alcohol intake, as mentioned in the paper:

$$A_i = \sum_{j=1}^{6} a_{ij} \times \varepsilon_j \times \overline{\eta_{ija}} \tag{13}$$

Where A_i represents the daily average alcohol intake of the ith sample, a_{ij} represents the average amount of each type of alcohol consumed by the ith questionnaire target, ε_j represents the alcohol content coefficient for the jth type of alcohol, and $\overline{\eta_{ija}}$ represents the selection coefficient for the jth type of alcohol by the ith sample.

Diet: Different dietary factors are related to various diseases. Given the involvement of diabetes and hypertension, we consider the intake of sugar and salt. In the questionnaire data, added sugar is only related to beverages, so we introduce a beverage sugar content coefficient ϱ to represent the proportion of sugar in a unit beverage. The daily intake of sugar is then calculated as follows:

$$S_i = s_i \times \varrho \times \overline{\eta_{is}} \tag{14}$$

Where S_i represents the daily average added sugar intake, s_i represents the average amount of each beverage consumed by the ith sample, ϱ represents the beverage sugar content coefficient, and $\overline{\eta_{is}}$ represents the beverage selection coefficient for the ith sample.

We define the average intake of dietary salt as a general representation for the sample, which is converted from the monthly family intake, considering the number of family members:

$$Y_i = \frac{y_i}{30 \times \overline{num}} \tag{15}$$

Where Y_i represents the daily average salt intake, y_i represents the average amount of salt consumed per meal by the ith sample, n represents the sample count, and $\overline{num}$ represents the average number of people per family.

Nature of Occupation: This primarily considers the level of job intensity, defined numerically as 1 for light, 2 for moderate, and 3 for heavy.

Physical Activity: This mainly considers the average daily exercise time. If there is no exercise habit, it is set to 0. Its parameter is defined as the product of exercise intensity and the average exercise time.

2.3. Optimized random forest model based on Smote

SMOTE Algorithm Establishment: The SMOTE algorithm is a method for synthetic minority class oversampling, making the sample set more balanced [10]. It achieves this by randomly generating a sample along the line connecting a sample with its k-nearest neighbors. This helps prevent overfitting. The main steps of the SMOTE algorithm are as follows:

Step 1: Select the nearest neighbor algorithm and calculate the k-nearest neighbors for each minority class sample.

Step 2: Randomly select N samples from the k-nearest neighbors and perform random linear interpolation.

Step 3: Create synthetic minority class samples.

Step 4: Combine the new samples with the original samples to create a new dataset.

Random Forest Establishment: Define the input variables as the X matrix and the output variables as the Y matrix, where the X matrix is a multidimensional matrix, and the Y matrix is a one-dimensional matrix. Define xxxxx as the value of the jth variable for the ith object, and xxxxx represents the output value for that variable, satisfying certain function conditions:

$$f\left(x_i^{(1)}, x_i^{(2)}, x_i^{(3)} \ldots x_i^{(j)}\right) = y_i \tag{16}$$

Assume that the output is divided into N classes:

$$O_1, O_2, O_3 \ldots O_N \tag{17}$$

Each category has a fixed output value defined asC_n. The regression tree is defined as:

$$f(x) = \sum_{i=1}^{N} C_n I \tag{18}$$

Meanwhile, for eachC_n, it is defined as the average output value within the region:

$$C_n = \frac{\sum y_i}{num\left(x_i^{(j)} | x_i^{(j)} \in O_N\right)} \tag{19}$$

Next, the optimal split variable and split point are selected, and the input values are divided. Assume that the pth variable is the split point and the position is k. The region of division is defined as:

$$O_1(p,k) = \{x|x^{(j)} \le k\}$$
$$O_2(p,k) = \{x|x^{(j)} > k\} \quad (20)$$

Each split is a binary split, and both the split point and split variable can be calculated:

$$\hat{C}_1 = ave(y_i|x_i \in O_1(p,k))$$
$$\hat{C}_2 = ave(y_i|x_i \in O_2(p,k)) \quad (21)$$
$$\hat{C}_n = ave(y_i|x_i \in O_n(p,k))$$

Finally, all variables and split points are defined, and the best decision tree is built. The input variables here are: Smoking Index, Eating Pattern Coefficient, Daily Alcohol Intake, Added Sugar, Dietary Salt, Nature of Occupation, Exercise Coefficient, and seven other variables. The output variables are two groups of binary variables (0 and 1) representing the presence or absence of hypertension and diabetes after being processed by the SMOTE algorithm.

3. Results

3.1. Parameter setting

We observed that there are more patients with hypertension than diabetes. In the sample, there are 1002 patients with hypertension and 233 patients with diabetes. The probability of having hypertension is much higher than that of having diabetes, with a patient proportion of 13%, which is significantly higher than the 3% proportion of diabetes patients. Next, we established random forest models to assess the correlation of hypertension and diabetes with seven sets of variables. We have defined some parameter settings in Table 3:

Table 3: Random Forest Parameter Settings Table.

Parameter	Parameter Setting
Training Set Ratio	0.8
Number of Decision Trees	100

3.2. The results of the Random Forest Model

Furthermore, the SMOTE algorithm was employed to oversample the training classification objects, and random forests were established based on the processed data. For the random forest models assessing the correlation of hypertension and diabetes with relevant factors, decision tree classification results were obtained as shown in Figure 1 and Figure 2:

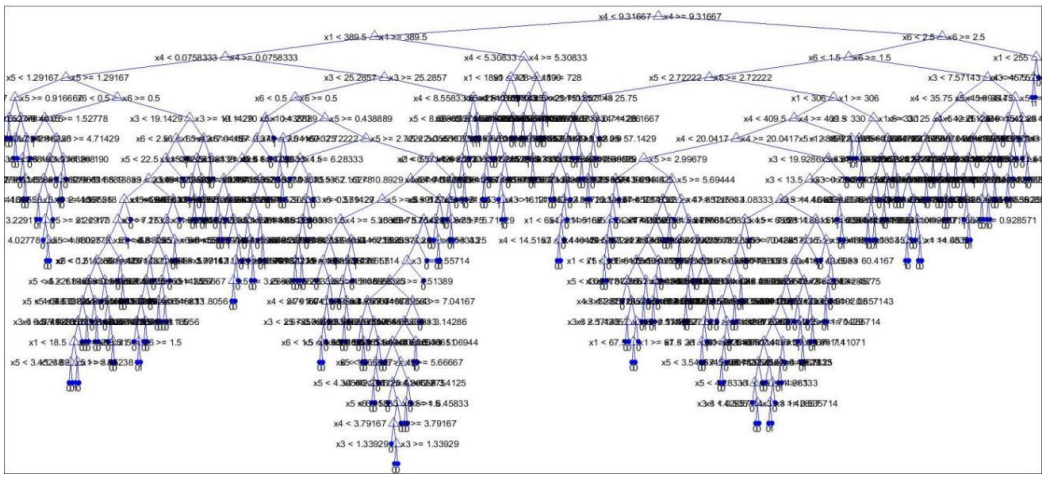

Figure 1: Decision Tree of the Hypertension Random Forest Model.

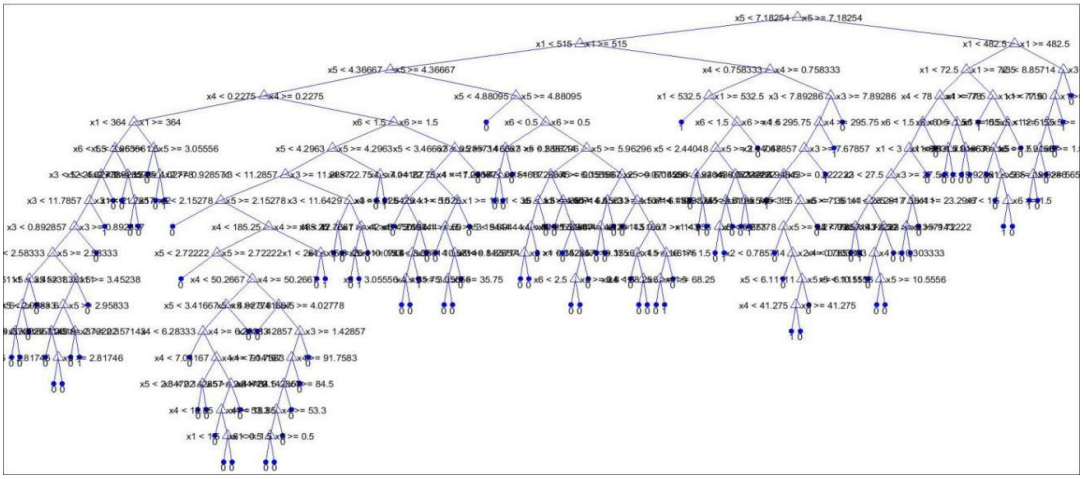

Figure 2: Decision Tree of the Diabetes Random Forest Model.

Furthermore, the importance levels of various relevant factors for both models were calculated. This was done by calculating the index importance using the function OOBPermuted Predictor Delta Error and comparing it with the numerical ratio of the training results and the actual results from the test set as a measure of model accuracy. The results are presented in Table 4 below:

Table 4. Random Forest Index Importance Table.

Independent Variable	**Index Importance**							**Model Accuracy**
Smoking Index	Daily Alcohol	Eating Rate	Added Sugar	Dietary Salt	Occupation Type	Exercise Hours	/	
Hypertension	1.52	0.82	-0.13	0.8	1.59	0.51	-0.34	87.30%
Diabetes	1.37	1.21	-0.11	1.35	1.01	0.37	-0.45	96.94%

Based on the index importance parameters, the following conclusions can be drawn:

Smoking and alcohol consumption have a greater impact on the occurrence of chronic diseases compared to daily dietary factors.

Various factors within eating habits are correlated with the type of disease. For hypertension, salt intake has a greater impact, while for diabetes, sugar intake has a greater impact. Additionally, having regular eating habits has a certain negative influence, indicating that individuals with regular eating habits are less likely to develop these diseases.

Occupation type also has a significant impact, with greater work intensity and work-related stress increasing the likelihood of developing chronic diseases.

The more regular physical activity in daily life, the lower the likelihood of developing these diseases.

4. Conclusions

We initially investigated the correlations between residents' lifestyle and dietary habits and factors such as age, gender, marital status, education level, and occupation to assist in selecting relevant variables. The team first established models for lifestyle and dietary habits separately. For lifestyle habits, the team introduced the smoking index and processed the data based on smoking status, smoker types, passive smoking exposure, and exercise habits. Then, in the context of dietary habits, the team considered variables like eating regularity, food intake, and the consumption of oil, salt, sugar, and alcohol. These variables were analyzed using both Pearson and Spearman correlation

coefficients in relation to influencing factors. Based on the analysis results, we systematically examined the positive or negative direction and the strength of the relationships between various variables and influencing factors.

We found strong correlations between some variables. Therefore, in selecting dependent variables, we decided to exclude variables with excessively high correlations. We then determined the dependent variables as the presence or absence of hypertension and diabetes. Since these variables were binary and had imbalanced proportions, we employed the SMOTE algorithm for oversampling. Next, we identified seven independent variables, encompassing the participants' basic information and their lifestyle and dietary habits. Finally, we used the random forest algorithm to analyze the importance of these variables. As a result, the model achieved accuracy rates of 87.30% for hypertension data and 96.94% for diabetes data, demonstrating a high level of accuracy.

Furthermore, we discovered that smoking and alcohol consumption had a more significant impact on the occurrence of chronic diseases compared to daily dietary factors. Various dietary factors were associated with the type of disease. For hypertension, salt intake had a greater influence, while for diabetes, sugar intake played a more significant role. Additionally, having regular eating habits had a certain negative impact, suggesting that individuals with consistent eating patterns were less likely to develop these diseases. Occupation type also had a substantial impact, with higher work intensity and job-related stress increasing the risk of chronic diseases. Finally, greater levels of daily physical activity were associated with a lower risk of disease.

References

[1] Naughton B J. The Chinese economy: Adaptation and growth [M]. Mit Press, 2018.

[2] Woolf S H, Schoomaker H. Life expectancy and mortality rates in the United States, 1959-2017[J]. Jama, 2019, 322(20): 1996-2016.

[3] Reddy K S, Shah B, Varghese C, et al. responding to the threat of chronic diseases in India [J]. The Lancet, 2005, 366(9498): 1744-1749.

[4] Schmidt H, Mah C L, Cook B, et al. Chronic disease prevention and health promotion[J]. Public health ethics: Cases spanning the globe, 2016: 137-176.

[5] Tian M, Wang H, Tong X, et al. Essential public health services' accessibility and its determinants among adults with chronic diseases in China[J]. PLoS One, 2015, 10(4): e0125262.

[6] Tian M, Chen Y, Zhao R, et al. Chronic disease knowledge and its determinants among chronically ill adults in rural areas of Shanxi Province in China: a cross-sectional study [J]. BMC public health, 2011, 11(1): 1-9.

[7] Kuhail M A, Ahmad B, Rottinghaus C. Smart resident: A personalized transportation guidance system[C]//2018 IEEE 5th International Congress on Information Science and Technology (CiSt). IEEE, 2018: 547-551.

[8] Sharon T. Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare [J]. Philosophy & Technology, 2017, 30(1): 93-121.

[9] Al-Hajje A, Awada S, Rachidi S, et al. Factors affecting medication adherence in Lebanese patients with chronic diseases [J]. Pharmacy practice, 2015, 13(3).

[10] Fernández A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of artificial intelligence research, 2018, 61: 863-905.