ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

Study on the factors influencing diabetes under interpretable machine learning model: A case study of Shenzhen residents' health survey data

Haiyuan Nong[1], Hailing Lin[1], Xiaomin Chen[1], Guorui Zhao[1,*]

[1]School of Computer Science and Engineering, Guangdong Ocean University, Yangjiang, 529500, China
[*]Corresponding author

Abstract: *Diabetes, a chronic metabolic disease, has long posed a significant challenge in the realm of global public health, seriously threatening the health of individuals worldwide. Utilizing epidemiological survey data on chronic noncommunicable diseases from Shenzhen, this study explored the relationship between lifestyle habits, dietary practices, and individual characteristics of diabetes. Through baseline analysis, 26 key variables, including age, gender, education, and hypertension, were identified. In comparing various statistical models and multiple machine learning algorithms, the XGBoost model was selected as the most effective diabetes prediction model due to its superior predictive performance. The SHAP model was subsequently employed to elucidate the XGBoost model's findings, revealing that age and hypertension emerged as significant positive factors, while social attributes and physical activity were identified as negative factors. Additionally, interactions between age and hypertension, as well as individual differences in dietary habits, were uncovered. The results of this study provide valuable insights into the prevention and control of diabetes.*

Keywords: *Diabetes mellitus, Influencing effects, Interaction, Machine learning, SHAP*

1. Introduction

Diabetes mellitus, a global chronic metabolic disease, severely affects multiple systems in the human body, particularly the nervous and vascular systems. With urbanization and lifestyle changes, diabetes incidence is rising worldwide. The World Health Organization (WHO)[1] reports approximately 537 million people aged 20 to 79 have diabetes, and this number is expected to increase to 640 million by 2030, accounting for 11.3% of the global population. In China, diabetes prevalence has increased yearly; by 2021 it reached 12.8%, accounting for one-quarter of the global total. In rapidly developing areas like Shenzhen, where economic growth and population mobility are high, diabetes poses significant public health challenges. Thus, understanding its pathogenesis and identifying key risk factors are crucial for effective prevention strategies and an important issue in global public health.

The research on the risk factors of diabetes has received extensive attention at home and abroad. In the international field, Kanko Kayo's study[2] showed that the simultaneous increase of GGT and ALT was significantly associated with the incidence of type 2 diabetes. Meghan O 'Hearn[3] global survey found that about 70% of diabetes cases are caused by poor diet. Vanderbilt University researcher Andrew S Perry[4] used a Cox regression model with time-dependent covariates to conclude that walking 10,000 steps a day is associated with a 44% reduction in diabetes risk. Naresh M Punjabi[5] et al. found a positive association between SDB and insulin resistance. Bajaj M[6], using smoking cessation interventions, found that increased nicotine in quittors reduced muscle glucose intake, producing insulin resistance and leading to diabetes. Research by Jeroen H P M van der Velde[7] has shown that exercising in the afternoon or evening is more helpful in preventing diabetes. In China: Lin[8], through a comparative study of different dietary habits and diabetes prevalence and analysis of clinical trials, found that areas with high fiber and high sugar diets had lower prevalence. Huang et al.[9] unconditional logistic regression was used to obtain the result that genetic factors are closely related to the onset of diabetes, and the 2-hour postprandial blood glucose has the most significant effect. Lin et al.[10]found that a fast-paced life is associated with an increased incidence of diabetes. Gong et al.[11]explored the independence of influencing factors through the Spearman correlation coefficient and univariate and multivariate logistic regression analysis. Lin et al.[12] recommended moderate-to-vigorous intensity exercise for more than 30 minutes, five times a week, through

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

Mendelian randomization. Zheng et al.[13] found that long-term exposure to outdoor artificial light at night was associated with a 28% increase in the risk of diabetes. Jin[14] used a case-control study to conduct an epidemiological analysis to understand the prevalence of diabetes and lifestyle, eating habits, and other factors.

The application of machine learning models in the analysis of factors affecting diabetes is becoming more and more widespread, with both domestic and foreign scholars studying different algorithms and models for diabetes risk factors. Hafiz Farooq Ahmad and Hamid Mukhtar[15] used five machine learning models, including Logistic, SVM, decision tree, random forest, and integrated majority voting, to explore the association between selected features and diabetes. T.veka and C. Chistopher[16] used the KNN classification method to determine whether diabetes was chronic or normal. In China: Zhang et al.[17] used a variety of machine learning models for a comparative study and found that both Logistic and LightGBM had a good prediction effect on the risk of diabetes. Zou Disha and Ye Yao et al.[18] used classification tree and logistic methods to select 6660 people for investigation and analysis, and found that the interaction of age, TG, and NAFLD was an important risk factor for type 2 diabetes.

Existing studies have identified various factors associated with the onset of diabetes, highlighting the potential of machine learning in analyzing diabetes risk. However, most studies focus on a single or a few factors, lacking systematic and comprehensive analysis, and lack of understanding of the interaction between factors and the joint mechanism of influence. Based on the above discussion, this study aims to use interpretable machine learning models to analyze the cross-sectional data of Shenzhen residents, construct an index system of diabetes-related factors, and deal with the complex nonlinear relationships in the data, so as to reveal the multi-factor impact path of diabetes more comprehensively and improve the prediction accuracy, and provide effective strategies and methods for the prevention and control of diabetes.

2. Methods

2.1. Research Framework

As a special economic zone in China, Shenzhen's rapid economic development and population mobility have significantly changed residents' living habits and dietary structure, which in turn have affected the incidence of diabetes. As shown in Figure 1, the aim of this study is to explore the relationship between socioeconomic, lifestyle, and genetic factors and the incidence of diabetes mellitus in Shenzhen. Interpretable machine learning models were used to analyze the influencing factors of diabetes and reveal the path relationship among the factors, so as to provide a scientific basis and strategic suggestions for the prevention and control of diabetes in Shenzhen, and theoretical support for the formulation of public health policies.

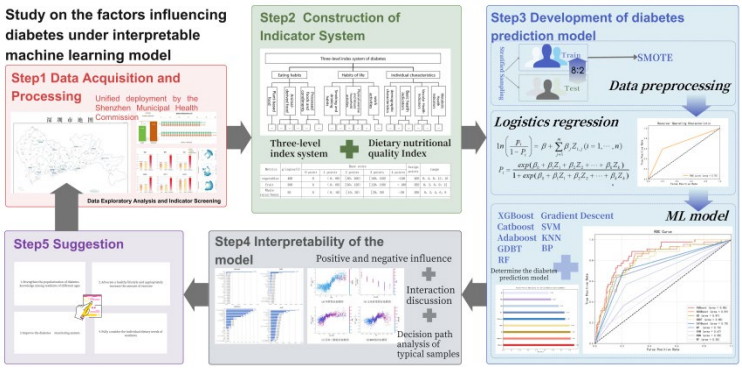

Figure 1: Research Framework

2.2. Data Acquisition and Processing

The data used in this study came from the public survey reports under the unified deployment of the Shenzhen Health Commission, and more than 700 staff from 10 district-level chronic disease prevention and treatment institutions and 100 community health service centers participated in the survey organized by Shenzhen Municipal Center for Chronic Disease Prevention and Control. The data were comprehensive and authoritative. Before the construction of the model, the questionnaire data

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

were strictly cleaned and denoised to ensure the accuracy and reliability of the results. Especially for the evaluation of diabetes, a complex and difficult-to-manage disease, this study not only constructed a comprehensive index system including dietary habits, lifestyle, and physiological indicators but also further refined the factors related to diabetes risk and established a detailed three-level index system (see Figure 2). This highlights key information such as the amount and variety of foods consumed, along with physical activity, alcohol, and tobacco habits.

Although diabetes is a systemic metabolic disease, the risk can be effectively reduced through scientific diet and lifestyle adjustment: In terms of dietary habits, this study focused on the intake and types of food, referring to the views of the American Diabetes Association, emphasizing the unique nutritional needs of each patient, and taking plant-based food, animal-derived food, processed foods and condiments as important secondary indicators. In terms of lifestyle, based on epidemiological studies, this study found that increasing physical activity was associated with a reduced risk of diabetes while smoking and drinking alcohol were associated with an increased risk. In addition, individual characteristics are also an important basis for assessing residents' behavioral habits and disease screening, which further reflects the comprehensiveness and scientific rigor of the index system construction in this study.

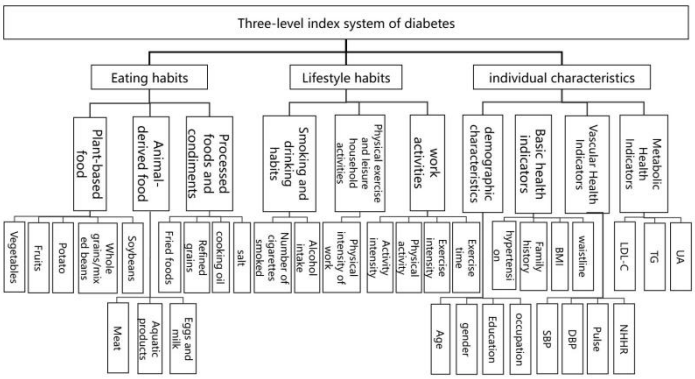

Figure 2: Chart of indicator system

2.3. Statistical Analysis

In this study, the datasets were first divided by stratified sampling, and SMOTE was only oversampled on the training set to avoid data overlap. Then, a wide range of models, from logistic regression to multiple advanced machine learning algorithms, were explored. These algorithms include Logistic, Gradient Descent, SVM, BP neural network, KNN, Random Forest, AdaBoost, GBDT, CatBoost, and XGBoost. Each of these algorithms has its own application scenarios[19-29]. Logistics are suitable for binary classification problems and have good interpretability, but it is difficult to deal with the problem[19] of data imbalance. GD uses the negative gradient direction as the optimal path, which is the basic method in unconstrained problems and can effectively approximate the minimum point[20]. SVM is good at dealing with high dimensionality and small sample learning problems and can deal with the interaction[21] of nonlinear features. The BP neural network can learn complex nonlinear mapping relationships and has a strong representation learning ability, making it suitable for large-scale datasets and multi-classification problems[22]. KNN is used for classification or regression based on nearest neighbor voting, which is suitable for non-parametric problems[23]. Random forest enhances the prediction accuracy and robustness of the model through the ensemble method, which can effectively deal with high-dimensional data and multi-classification problems and has strong robustness[24] to outliers and noise. AdaBoost is a high-precision classifier, which can use a variety of methods to build weak classifiers[25]. GBDT is an ensemble learning algorithm, which can deal with complex problems[26] such as high-dimension, sparse features, and nonlinear relationships. CatBoost sorts the ascending method against the noise of the training focus point, to avoid the deviation of gradient estimation and then solve the problem[27] of prediction deviation; XGBoost is for objective function optimization and introduces regularization technique is improved, such as in the paper the XGBoost: A Scalable Tree Boosting System "[28]has proven its model of low computation complexity, fast, high accuracy, etc[29]. It is necessary to comprehensively consider various indicators (ROC curve and AUC score) to find the most appropriate diabetes prediction model.

Although multiple algorithms provide high prediction accuracy, they generally have poor

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

interpretability. Therefore, an interpretable SHAP model was introduced to analyze the positive and negative effects, interactions, and typical sample decision paths of each indicator of disease. The SHAP value is mainly used to quantify the contribution of each feature to the model prediction, and then calculate the different marginal contributions of this feature in all feature sequences to reveal the positive and negative effects and interactions of each factor on the incidence of diabetes, thus providing deeper insight into the prediction of diabetes.

3. Results

3.1. Exploratory Analysis of Data

In this study, a sample of 221 diabetic residents and 7328 normal residents in Shenzhen were analyzed. The results are shown in Table 1, which presents the baseline statistics. The baseline table established the framework for the analysis of the relationship between the risk of diabetes and multiple factors (age, sex, lifestyle, and physiological indicators). Based on this framework, this study conducted data dimensionality reduction, excluded 14 indicators that failed the significance test, and retained the remaining 26 indicators, which provided strong data support for follow-up research. In the table, more than half (56.1%) of the affected group were women, and the mean age of patients at diagnosis was 61.1 (10.9) years. The average age of the healthy group was 47.6 (11.4). The mean WC, LDL-C, and systolic blood pressure of the disease group were 87.1 (8.8), 3.2 (0.7) and 122.7, respectively, which were significantly higher than those of the control group, at 80.1 (10.3), 2.9 (0.7) and 112.8 (13.6) respectively. The increase in these data is closely related to the prevalence of diabetes mellitus and the increased risk of cardiovascular disease. The proportion of illiteracy and primary school education in the diabetic group was higher than that in the healthy group, and the proportion of other education levels in the healthy group was higher than that in the diabetic group. In the occupational distribution, the proportion of the diabetic group in retirement was much higher than that of the healthy group, while the healthy group had a higher proportion in the categories of students, workers, and scientific and technical personnel than the diabetic group. The overall physical intensity of the residents was mild, but the proportion of the diabetes group was much higher than that of the healthy group in the work intensity of the separated and retired workers.

Table 1: Baseline statistics

Variable	Sum(*n*=7549)	Diabetes		Statistics	*P*value	SMD
		0 (*n*=7328)	1 (*n*=221)			
Age, Mean±SD	47.978±11.633	47.581±11.421	61.128±10.958	*t*=-17.394[1]	<0.001	1.212
gender, *n* (%)				χ^2=0.013[2]	0.908	0.008
1	3342 (44.271)	3245 (44.282)	97 (43.891)			
2	4207 (55.729)	4083 (55.718)	124 (56.109)			
Education, *n* (%)				*Z*=5.629[3]	<0.001	0.448
1	131 (1.735)	116 (1.583)	15 (6.787)			
2	730 (9.670)	686 (9.361)	44 (19.910)			
3	2093 (27.726)	2031 (27.716)	62 (28.054)			
4	2493 (33.024)	2437 (33.256)	56 (25.339)			
5	2026 (26.838)	1983 (27.061)	43 (19.457)			
6	76 (1.007)	75 (1.023)	1 (0.452)			
Number of cigarettes, M (Q_1, Q_3)	0.000(0.000, 1.750)	0.000(0.000, 1.750)	0.000(0.000, 1.750)	*Z*=0.043[3]	0.966	0.021
Fruits, M (Q_1, Q_3)	100.000(60.000, 150.000)	100.000(60.000, 150.000)	100.000(40.000, 150.000)	*Z*=2.246[3]	0.025	0.156
Whole grains, M (Q_1, Q_3)	6.667(1.667, 14.296)	6.667(1.667, 14.296)	10.000(1.667, 14.296)	*Z*=-1.672[3]	0.094	0.085
Soybeans, M (Q1, Q3)	60.000 (27.500, 95.000)	60.000 (28.333, 95.000)	53.333 (21.667, 93.333)	Z=1.9673	0.049	0.122
Physical intensity of work, n (%)				Z=4.7833	<0.001	0.614
0	636 (8.425)	573 (7.819)	63 (28.507)			
1	5025 (66.565)	4920 (67.140)	105 (47.511)			
2	1778 (23.553)	1735 (23.676)	43 (19.457)			

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

Variable	Sum(*n*=7549)	Diabetes		Statistics	*P*value	SMD
		0 (*n*=7328)	1 (*n*=221)			
3	110 (1.457)	100 (1.365)	10 (4.525)			
Exercise intensity, M (Q1, Q3)	1.000 (0.000, 2.000)	1.000 (0.000, 2.000)	2.000 (0.000, 2.000)	Z=-4.4673	<0.001	0.313
Exercise time, M (Q1, Q3)	20.000 (0.000, 42.000)	15.000 (0.000, 40.000)	30.000 (0.000, 60.000)	Z=-3.6363	<0.001	0.208
hypertension, n (%)				χ2=365.1802	<0.001	0.908
1	620 (8.213)	525 (7.164)	95 (42.986)			
2	6929 (91.787)	6803 (92.836)	126 (57.014)			
BMI, Mean±SD	23.006±3.220	22.955±3.219	24.723±2.740	t=-9.3975	<0.001	0.592
WHR, Mean±SD	0.851±0.069	0.849±0.069	0.900±0.060	t=-12.3145	<0.001	0.785
waistline, Mean±SD	80.293±10.346	80.089±10.320	87.075±8.804	t=-11.5595	<0.001	0.729
SBP, Mean±SD	113.100±13.782	112.811±13.661	122.677±14.359	t=-10.5611	<0.001	0.705
DBP, Mean±SD	74.098±8.847	73.984±8.825	77.908±8.771	t=-6.5141	<0.001	0.447
HDL-C, Mean±SD	1.215±0.251	1.218±0.250	1.142±0.248	t=4.4471	<0.001	0.305
NHHR, M (Q1, Q3)	2.961 (2.275, 3.757)	2.937 (2.264, 3.738)	3.591 (2.905, 4.191)	Z=-8.1323	<0.001	0.559
LDL-C, Mean±SD	2.977±0.697	2.969±0.697	3.241±0.676	t=-5.7221	<0.001	0.397
TG, M (Q1, Q3)	1.180 (0.830, 1.580)	1.170 (0.820, 1.561)	1.561 (1.200, 2.070)	Z=-8.6613	<0.001	0.581
UA, Mean±SD	297.609±76.633	297.151±76.774	312.817±70.287	t=-3.2555	0.001	0.213

Notes: 1. Independent sample t test; 2. Pearson χ^2 test; 3. Mann Whitney U test; 4. Fisher exact probability method based on Monte Carlo estimation; 5. Independent sample t test with variance correction.

3.2. Statistical Model and ML Model Results

Statistical models and nine machine learning algorithms were used in this study. The dataset was divided into a training set and a validation set by 8:2 stratified sampling. SMOTE technology was applied in the training set to expand the minority samples and improve the recognition ability of the model, while the original distribution was maintained in the validation set to ensure the accuracy of the evaluation. The model was trained on the SMOTE-processed training set, and the same test set was used for validation. The results of the 10 models were compared horizontally.

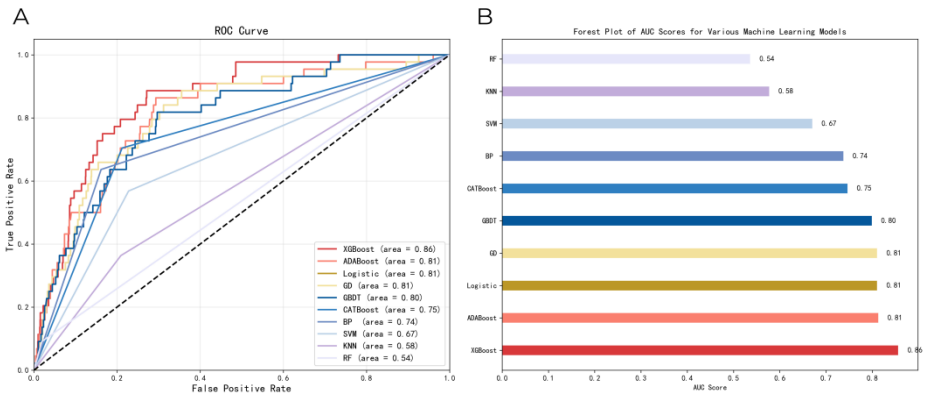

Figure 3: Comparison of 10 algorithms

As shown in Figure 3, XGBoost, AdaBoost, GD, GBDT, and CatBoost are all variants of ensemble learning algorithms. All the models showed high AUC values on the ROC curve (XGBoost=0.86, AdaBoost=0.81, GD=0.81, GBDT=0.80, CatBoost=0.75). Logistic (0.81), as a statistical model with strong explanatory power, can output the influence of each independent variable on the outcome, but may not fully capture the complex nonlinear relationship in the data. The AUC values of the BP neural network and SVM on the ROC curve were at a moderate level (BP=0.74, SVM=0.67). The AUC values of the RF and KNN on the ROC curve were low (KNN=0.58, RF=0.55), which were close to the level of random guess (AUC=0.5). XGBoost showed the highest stability and accuracy in diabetes prediction. Thus, XGBoost was chosen as the diabetes prediction model, and SHAP was employed to explain and

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

analyze the factors influencing diabetes, enhancing the model's interpretability.

3.3. SHAP Results

Figure 4A is a summary plot of SHAP, demonstrating the ranking of importance of factors affecting diabetes prevalence: hypertension, age, education, and TG had larger absolute values and significantly affected the model. Among them, hypertension is the most important factor, and it is positively correlated with the risk of diabetes. Age also showed a linear positive correlation, with the risk increasing beyond a certain threshold. Higher education level was associated with lower risk. Higher intensity of physical exercise decreased the risk. In addition, men are more likely to develop the disease than women.

Figure 4B selected the four characteristics of hypertension, age, education, and TG that had a significant influence on the model, drew the SHAP feature dependence diagram, and found that there was a significant interaction between age and hypertension on the incidence of diabetes. With the increase of BMI and TG values, SHAP values showed a trend of first increasing and then decreasing, indicating that these two factors had complex effects on the risk of diabetes.

SHAP can deeply analyze the risk factors of diabetes in individuals. In one patient predicted to have diabetes (Figure 4C), the red factors (hypertension, age 75 years, and high TG level 2.28mmol/L) had more significant effects than the blue factors (female gender, normal salt intake, and high education level), so the patient was diagnosed as having diabetes; In contrast, among the characteristics predicted to be non-diabetic (Figure 4D), the blue factors (no hypertension, younger age of 47 years, and higher education) had a significantly greater impact than the red factors (high uric acid of 200mmol/L) and were therefore defined as non-diabetic.

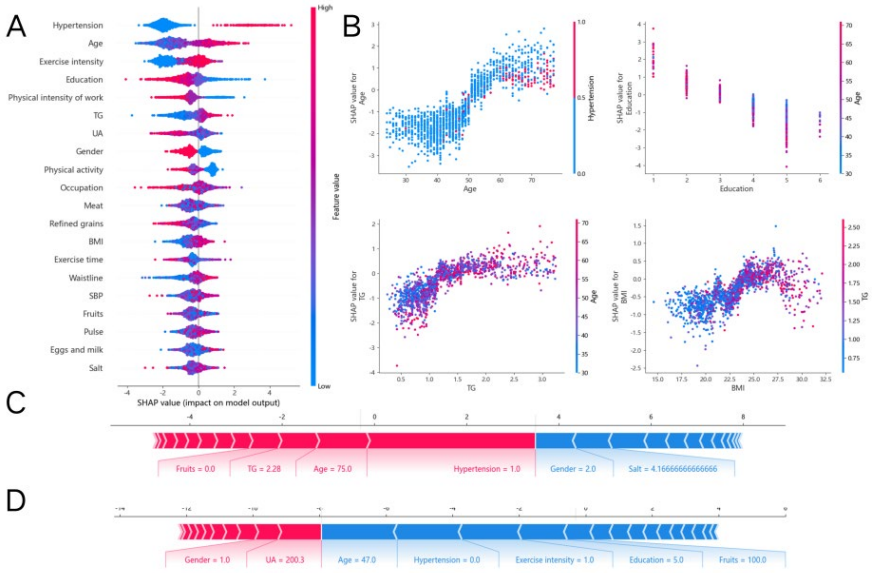

Figure 4: Interpretation of the model

4. Discussion

The dataset used in this study comes from the publicly available unfiltered large-scale dataset in Shenzhen. Statistical models and nine machine learning models were used and the results were compared to obtain the most appropriate XGBoost model (auc score=0.86). SHAP was used to enhance the model's interpretability, and the final model showed good performance in diabetes prediction. In the future, this model can be applied to actual clinical diagnosis and health management to provide more accurate and personalized intervention for diabetic patients. However, the data only came from the residents of Shenzhen, which has regional limitations. Only the indicators of the questionnaire were considered, and the possible influence of other diseases on diabetes in the results was not further explored. In the future, we will further explore the relationship between type 2 diabetes and diseases such as tumors, coronary heart disease, or some organ lesions, so as to provide a more comprehensive basis for the comprehensive prevention and treatment of diabetes. In the future, more physical

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

examination data from different regions should be obtained to evaluate the impact of related problems more comprehensively and fully.

5. Conclusions and Recommendations

The prevalence of diabetes and its associated risk factors often precedes the disease itself; therefore, prevention and reduction of these risk factors become key to diabetes control. This study focused on Shenzhen residents, aiming to provide scientific evidence for policy-making and intervention through the analysis of diabetes and its related risk factors. Studies have found that diabetic patients have significantly higher levels of physiological indicators such as LDL-C and systolic blood pressure than non-diabetic patients, and the abnormal elevation of these indicators is closely related to diabetes and its risk of cardiovascular diseases. SMOTE technology was used to deal with data imbalance. Based on the performance comparison of ten models, the XGBoost model with the highest accuracy (AUC=0.86) was finally selected. Based on this, SHAP analysis revealed that age, hypertension, exercise intensity, and education level were the main influencing factors. The interaction between age and hypertension increased with age, while education level showed an inverse relationship with diabetes. The effects of dietary habits on diabetes vary among individuals. The results showed that age and hypertension were the main positive factors, while social attributes and physical activity were the negative factors. The effect of dietary habits on diabetes varies among individuals. The present study proposes the following policy recommendations and interventions: (1) strengthen diabetes knowledge among residents of different age groups: for different age groups, especially the middle-aged and elderly, improve diabetes knowledge and promote healthy lifestyles through community, media, and lectures; (2) Advocating a healthy lifestyle and increasing exercise: it is recommended that residents reduce foods high in sugar, fat and salt, increase fiber intake, reduce sugar-sweetened beverages, and encourage regular physical exercise to reduce the risk of diabetes; (3) Improve the diabetes monitoring system: communities and hospitals should establish a diabetes monitoring system, regularly screen high-risk groups, pay attention to blood glucose, blood pressure, blood lipids, body weight, and other indicators, and regularly follow up diagnosed patients; (4) The individualized dietary needs of residents should be fully considered. Given the impact of dietary habits on the risk of diabetes, diabetic patients are recommended to work with dietitians or doctors to develop personalized dietary plans to control blood glucose. Healthy residents should also focus on a balanced diet to prevent diabetes.

Author Statement

Haiyuan Nong, Hailing Lin, and Xiaomin Chen are co-first authors of the article, completing the conceptualization, methodology, data curation, writing- original draft preparation, and visualization.

References

[1] Magliano, D. J., Boyko, E. J., & IDF Diabetes Atlas 10th edition scientific committee. IDF DIABETES ATLAS. (10th ed.). International Diabetes Federation, 2021.

[2] Kaneko, K., Yatsuya, H., Li, Y., & Aoyama, A. Association of gamma-glutamyl transferase and alanine aminotransferase with type 2 diabetes mellitus incidence in middle-aged Japanese men: 12-year follow up. Journal of diabetes investigation, 2019, 10(3), 837–845.

[3] O'Hearn, M., Lara-Castor, L., Cudhea, F., & Global Dietary Database. Incident type 2 diabetes attributable to suboptimal diet in 184 countries. Nature medicine, 2023, 29(4), 982–995.

[4] Perry, A. S., Annis, J. S., Master, H., & Brittain, E. L. Association of Longitudinal Activity Measures and Diabetes Risk: An Analysis From the National Institutes of Health All of Us Research Program. The Journal of clinical endocrinology and metabolism, 2023, 108(5), 1101–1109.

[5] Punjabi, N. M., Shahar, E., Redline, S., Gottlieb, D. J., Givelber, R., Resnick, H. E., & Sleep Heart Health Study Investigators. Sleep-disordered breathing, glucose intolerance, and insulin resistance: the Sleep Heart Health Study. American journal of epidemiology, 2004, 160(6), 521–530.

[6] Bajaj M. Nicotine and insulin resistance: when the smoke clears. Diabetes, 2012, 61(12), 3078–3080.

[7] van der Velde, J. H. P. M., Boone, S. C., Winters-van Eekelen, E., Rosendaal, F. R., & de Mutsert, R. Timing of physical activity in relation to liver fat content and insulin resistance. Diabetologia, 2023, 66(3), 461–471.

[8] Lin huoshui. The significance of dietary factors in the onset and prevention of diabetes. Medical

ISSN 2616-5775 Vol. 7, Issue 11: 34-41, DOI: 10.25236/AJCIS.2024.071105

Philosophy, 1991, (05), 39-40.
[9] Huang gaoming, Liang qiuping. Application of non-conditional logistic regression in the analysis of risk factors for the onset of diabetes. Guangxi Preventive Medicine, 2000, (04), 205-207.
[10] Lin baowang, Huang xiaoke, Wei jing... & Zhao tiejian. The relationship between the factors of onset of type 2 diabetes and the pace of life. Contemporary Medicine, 2010, (04), 156-157+27.
[11] Gong, D., Chen, X., Yang, L., Zhang, Y., Zhong, Q., Liu, J., Yan, & Wang, J. From normal population to prediabetes and diabetes: study of influencing factors and prediction models. Frontiers in endocrinology, 2023, 14, 1225696.
[12] Lin, Y., Sun, Y., Zhang, Z., Wang, Z., Wu, T., Wu, F., Li, Z., Meng, F., & Fu, M.T. A cross-sectional study of optimal exercise combinations for type 2 diabetes. Journal of Public Health, 2023, 1-11.
[13] Zheng, R., Xin, Z., T., & Xu, Y. Outdoor light at night in relation to glucose homoeostasis and diabetes in Chinese adults: a national and cross-sectional study of 98,658 participants from 162 study sites. Diabetologia, 2023, 66(2), 336–345.
[14] Jin yuelong, Chen yan, Kang yaowen,et al. A case-control study on influencing factors of type 2 diabetes mellitus [J]. Journal of Wannan Medical College, 2012, 31(1):55-59.
[15] Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. Applied Sciences, 2021, 11(3), 1173.
[16] Viveka, T., Columbus, C.C., Velmurugan, N.S. To control diabetes using machine learning algorithm and calorie measurement technique. Intelligent Automation & Soft Computing, 2022, 33(1), 535-547.
[17] Zhang hongmei, Zhang ning, Sun yujiao & Zhang zhou. Analysis of factors affecting mild cognitive impairment in type 2 diabetes mellitus based on machine learning and logistic regression analysis model. The Chinese journal of disease control, 2024, (03): 269-276.
[18] Zou, D., Ye, Y., Zou, N., & Yu, J. Analysis of risk factors and their interactions in type 2 diabetes mellitus: A cross-sectional survey in Guilin, China. Journal of diabetes investigation, 2017, 8(2), 188–194.
[19] Li Changshan. Construction of Enterprise financial risk early warning Model based on Logistic regression method [J]. Statistics and decision, 2018, (6): 185-188.
[20] Yuan Yaxiang, Optimization Theory and Method [M]. Beijing: Science Press, 1997:108-121.
[21] Zhang Xuegong. On statistical learning theory and support vector machine [J]. Journal of automation, 2000. (01): 36-46.
[22] Huang Li. Research on Algorithm Improvement and Application of BP Neural Network [D]. Chongqing Normal University, 2008.
[23] SANG Ying-bin,LIU Qiong-sun. Improved k-nearest neighbor classification algorithm[J]. Computer Engineering and Applications, 2009, 45(11): 145-146.
[24] Fang Kuangnan, Wu Jian-Bin, Zhu Jianping, A review of random forest method [J]. Statistics and Information Forum, 2011, 26(03):32-38.
[25] CAO Y, MIao Q G, Liu J C, Research Progress and Prospect of AdaBoost Algorithm [J]. Acta Automatica Sinica, 2013, 39(06):745-758.
[26] Cao Ying-chao. Improved decision tree classification GDBT iterative algorithm and its application [J]. Journal of horizon of science and technology, 2017, (12): 105 + 149.
[27] Miao Fengshun, Diabetes prediction method based on CatBoost algorithm [J]. Computer system application, 2019, 28 (9): 215-218.
[28] Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[29] Wang Yan, Guo Yuankai. Application of Improved XGBoost Model in Stock Prediction [J]. Computer Engineering and Applications,2019,55(20):202-207.