Proceedings of the 3rd International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/24/20231151

Identification of pivotal nutritional factors to reduce risk of prediabetes by machine learning

Ke Ni

Lake Forest Academy, Lake Forest, The United States

ke.ni@students.lfanet.org

Abstract. Type 2 Diabetes is a major concern in healthcare worldwide with increasingly public health burden. Prediabetes has been considered as a critical stage to reduce the risk of Type 2 Diabetes since it's highly correlated with lifestyle changes. The present study explored the risk factor for prediabetes and built a machine learning model to prioritize the important factors for effective prevention. In this study, data was extracted from National Health and Nutrition Examination Survey (NHANES), including over 36,000 people classified to diabetes, prediabetes and healthy. Six machine learning models, namely, random forest, k-Nearest Neighbor (KNN), support vector machine (SVM), gradient boosting, LDA, and Xgboosting were trained for prediabetes risk factor prediction. Random forest eventually outperformed the other four models with an area under curve(AUC) score of 0.71. Also, explorations on the top features that are highly correlated with prediabetes were made. Multiple nutritional factors were identified, including thiamin, folic acid, and caffeine, which shows significant difference between healthy control and prediabetes and the lifestyles regarding these nutrition are recommended to change accordingly.

Keywords: Type 2 Diabetes, Prediabetes, Machine Learning, Random Forest.

1. Introduction

Type 2 Diabetes is a prominent chronic disease worldwide [1], characterized by abnormally high blood glucose that continues to appear because of the process of making or utilizing insulin [2]. In the present day, more than 35 million people in the United States have type 2 diabetes, and a growing trend of type 2 diabetes among adolescents is being observed in all racial and ethnic minority groups [2]. This number is continuously increasing as shown by projected global prevalence of 8.6%, equivalent to 548 million people, by the end of 2045 [4]. The diabetes crumbles the society both physiologically and economically, and the best way to mitigate the losses is to prevent diabetes or even prediabetes from happening. Prediabetes is defined as fasting glucose between 100 mg/dL and 125 mg/dL [5]. Without diagnosis and treatment, prediabetes can easily convert to type 2 diabetes. While only 5% people with normal glucose level go on to have type 2 diabetes, 33–65% of people with prediabetes will have type 2 diabetes within 6 years [6].

DOI: 10.54254/2753-8818/24/20231151

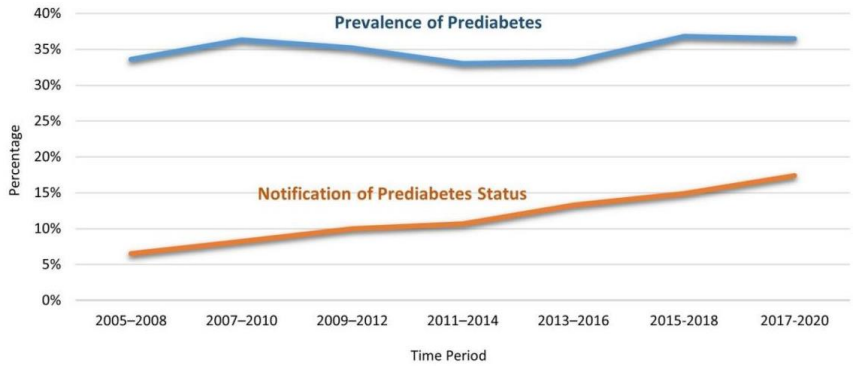

Figure 1. National Trends in Prevalence and Notification of Prediabetes Among US Adults Aged 18 Years or Older, 2005–2008 to 2017–2020.

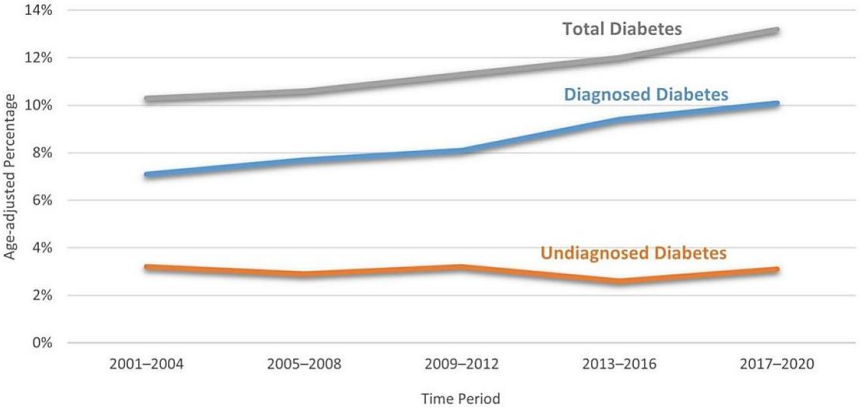

Figure 2. National Trends in Prevalence and Notification of Diabetes Among US Adults Aged 18 Years or Older, 2001–2004 to 2017–2020.

Nevertheless, the prediction and treatment at present are not satisfying, and an increasing trend of prediabetes is shown. Recently there has been a 9% to 23% increase in the prevalence of prediabetes among US adolescents aged 12 to 19 years from 2001–2004 to 2017–2020. An estimated 96 million adults aged 18 years or older had prediabetes in 2019 [7]. As shown in Figure 1, though the percentage of the US population who are notified of prediabetes diagnosis has increased from 6.5% to 17.4%, it is still less than 50% of the percentage of the population who are prediabetic [5]. Compared to Figure 2, Figure 1 suggested that the detection of prediabetes is not as effective as that of diabetes [8]. Without treatment, there is a strong possibility of progression from prediabetes to T2DM (type 2 diabetes) within 7 years [9]. More than 80% of people whose glucose level fits in the standard of prediabetes aren't aware of the fact that they have it [5].

Therefore, a precise model for people to make predictions on their risk of prediabetes is imperative, and more analysis of the associations between the risk of having prediabetes and different risk factors should be conducted. Constructing a precise model for prediabetes prediction is a challenging task. Researchers have adopted various methods to build accurate model using demographic and clinical data in recent years. Jiahua Wu et al. conducted study that they adopted Cox proportional-hazards model to analyze the associations between demographic and anthropometric parameters and risk of prediabetes with an AUC equals to 0.702 [10]. Igbe Tobore et al. proposed tapping electrocardiogram (ECG) rhythm and electroencephalogram (EEG) to build a prediction model for prediabetes with an ensemble learning classifier consisting of six individual classification models [11].

A variety of publicly accessible databases pave the way to train an accurate machine learning model for prediabetes risk prediction, such as Korean National Health and Nutrition Examination Survey (KNHANES) [12], NTT Medical Center Tokyo [13], and Qatar bio bank [14]. Various machine learning models have been adopted and optimized, including Artificial Neural Network (ANN),

DOI: 10.54254/2753-8818/24/20231151

Support Vector Machine (SVM), Xgboosting (XGB), Random forest (RF). Prior studies have identified key risk factors that may contribute to prediabetes, including demographic risk factors such as age, gender, marital status, level of education, a family history of diabetes, wealth status, region of residence-anthropometric measurements, weight, height, body mass index (BMI), lifestyle risk factors, cigarette smoking, alcohol drinking, tea and coffee drinking, consumption of beef, pork, mutton, fish, vegetables, and fruits, preference for sweet and salty food in daily life, work stress, physical activity, sleep duration, and examination risk factors, abdominal obesity, dyslipidemia, hypertension, maternal history of gestational diabetes [16].

However, the importance ranking of the risk factors for different people is needed to enhance the explainability of the machine learning models. Furthermore, bridging the behavioral risk factor with inherent biomedical mechanisms is critical to validify the reliability of the machine learning models. In the present study, an explainable machine learning pipeline to identify the key factors that are correlated with prediabetes was constructed. This research built a highly accurate random forest model for prediction of prediabetes. Then, from the nutritional perspective, the biochemical mechanisms that are highly associated with the factors such as protein intake and carbohydrate intake was explored. The study should contribute a more comprehensive and in-depth insight for prediction of prediabetes in a personalized manner.

2. Methods

2.1. Dataset

The dataset for identifying the risk factors of prediabetes is selected as the 2009-to-2018 National Health and Nutrition Examination Survey (NHANES) record. NHANES is a program that assesses the health and nutritional status of people all over the United States. Being such a health program, the NHANES survey questions are categorized as demographic, socioeconomic, dietary, and health-related. Other than the survey component, the examination component under NHANES's studies includes medical, dental, and physiological measurements, also laboratory tests administered by highly trained medical personnel [15]. According to the definition of prediabetes from CDC, people whose fasting glucose is between 100 mg/dL and 125 mg/dL are considered to be prediabetic [16]. This study added up the number of people who have been told by doctors that they have prediabetes but not yet diabetes (as shown by DIQ160 "Ever told you have prediabetes" and DIQ010 "Doctors told you have diabetes" in NHANES questionnaires) and the number of people who have never been told that they are prediabetics but fit in the 100 mg/dL to 125 mg/dL fasting glucose level according to their record in laboratory fasting glucose data. Then healthy individuals are labeled as 0, prediabetics are labeled as 1, and diabetics are labeled as 2.

2.2. Data preprocessing

The data preprocessing includes outlier rejection, missing values imputation, and standardization. Before performing data preprocessing, conversion of values of categorical variables to numerical ones were made, including gender, race. Regarding gender, male is labeled as 1 and female is labeled as 2. Regarding race/Hispanic origin, Mexican American is labeled as 1, Other Hispanic is labeled as 2, Non-Hispanic White is labeled as 3, Non-Hispanic Black is labeled as 4, Other Race - Including Multi-Racial is labeled as 5. Missing values were filled with the median values since this kind of imputation ensures the continuity of the entire dataset. Third, standardization and normalization were performed on the dataset.

2.3. Training dataset preparation

The technique of Stratified K Fold Cross Validation can sustain the target class frequencies in each subset. This technique requires dividing the entire dataset into k-independent subsets. Only one of the small subsets is used to train the classifier, and the rest of the subsets are used to evaluate the generalization error [17]. Different from K Fold Cross Validation, Stratified K Fold Cross Validation

DOI: 10.54254/2753-8818/24/20231151

makes sure that all the training and test subsets would have almost the same portion of the target class with regard to the original dataset. In addition, to prevent repeating k times of the training algorithm from the start, which implies k times long time taken, a Stratified K Fold approach was applied [18]. The preprocessed dataset is split into training and testing dataset with proportion of 80% and 20%. Stratified five-fold cross validation was applied on the training dataset. The six machine learning models were optimized with parameter tuning by Sklearn Grid Search. The machine learning model with best performance was selected for feature importance ranking.

2.4. Correlation analysis

Then the study employed a correlation analysis with the visualization of a heatmap. Spearman's rank correlation coefficient was used to measure the strength of association between two ranked variables. It is employed since qualitative features could be properly assessed by this arithmetic [19]. The rank correlation coefficient is denoted by ρ or r_R and is given by

$$\rho = r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

ρ= the strength of the rank correlation between variables, d_i=the difference between the x rank and the y rank for each pair of data, $\sum d_i^2$ = sum of the squared differences between x and y variable ranks, n=sample size.

2.5. Machine learning models

Multiple machine learning models are trained and tested, including random forest, k-Nearest Neighbor (KNN), support vector machine (SVM), gradient boosting, LDA, and Xgboosting. The random forest model makes predictions by combining decisions from a sequence of decision trees as the base learning models. Every decision tree has high variance, but when all of them were combined together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In random forests, each of the decision trees is constructed independently using a different subsample of the data [20]. Support Vector Machine (SVM) aims to find an optimal hyperplane to separate the classes in an N-dimensional space, where N is the number of features/attributes in the dataset. Extreme Gradient Boosting (XgBoosting) is an implementation of Gradient Boosted decision trees, which is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers, each predictor correcting its predecessor's error. The k-nearest neighbors algorithm (k-NN) averages the values of the k nearest neighbors to generate the final output. It one of those instance-based learning, or lazy learning models because the function is only approximated locally and all computation is deferred until classification.

2.6. Evaluation metrics

Several quantitative metrics such as accuracy, specificity, sensitivity were adopted in order to measure and validate the performance of the classifiers. Further, ROC (receiver operating characteristics) was used to show the trade-off between sensitivity and specificity, and AUC (area under curve) score was adopted to assess the classification performance. There are four possible states: true positive (TP), true negative (TN),false positive (FP), false negative (FN). Accuracy: It represents the overall performance of the classifier, showing the ability to correctly predict given samples. It is calculated by:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Sensitivity: It shows the classifier's ability to identify positive results and is calculated by:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

DOI: 10.54254/2753-8818/24/20231151

Specificity: It is on contrary to sensitivity, telling the classifier's ability to identify negative results. This metric is calculated by:

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

On the ROC graph, the vertical axis represents sensitivity, while the horizontal axis represents one minus specificity. AUC represents the area under the ROC curve. Since the ideal result is both sensitivity and specificity equal to 1, a model with the biggest AUC has the highest accuracy.

2.7. Feature importance ranking

Different methods of feature importance ranking are applied to the six machine learning models, which are random forest, k-Nearest Neighbor (KNN), support vector machine (SVM), gradient boosting, LDA, and Xgboosting. For random forest, the ranking of feature importance is determined by the number of times a feature is used to split the data. Despite this method, Recursive Feature Elimination (RFE) and Permutation Importance are employed. In Recursive Feature Elimination, the performance of the model is recorded when each feature is removed from the dataset. In Permutation Importance, the values of features are randomly permuted and then the corresponding model's accuracy is measured.

3. Results and discussion

3.1. Descriptions of study design and statistics of study populations

Figure 3 shows the pipeline that uses machine learning models to investigate the risk factor of prediabetes among the U.S. population. The entire study design consists of four steps—data extraction, data preprocessing, machine learning model training, and output analysis. Look into the NHANES's 2009-2018 database, 4626 individuals were considered as diabetes, 6638 individuals were considered as prediabetics, and 24891 individuals were considered as healthy. 104 features (either discrete or continuous) of individual records of geographic, dietary, examination and laboratory data were selected and extracted for deeper analysis. Then the extracted dataset was standardized and explored tentatively. The trained and modified machine learning model with best performance score produced feature importance ranking as the final output. Based on the feature importance ranking and the distribution of features among diabetes, prediabetes, and healthy individuals, recommendations were also provided as a final result.

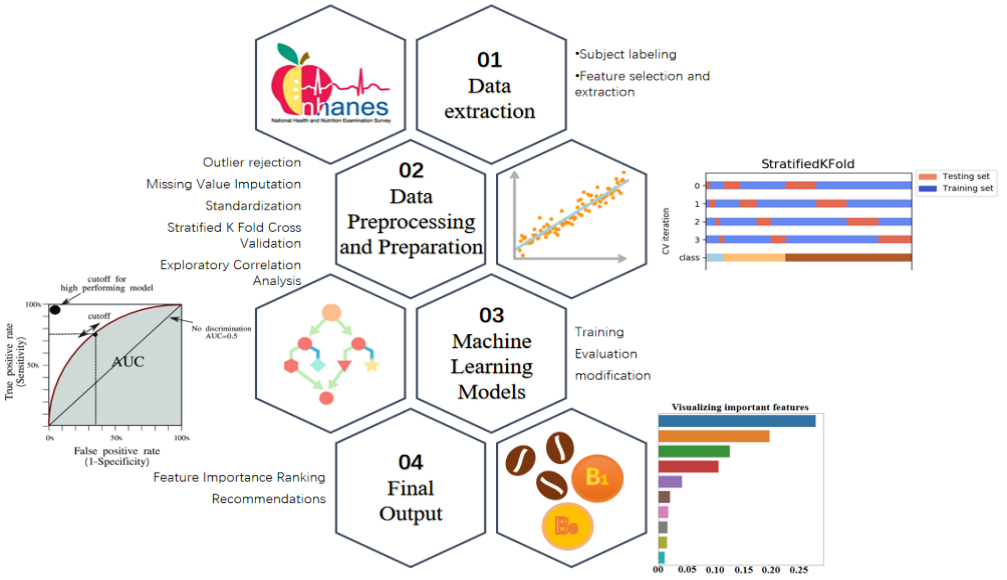

Figure 3. Workflow of machine learning-based prediabetes prediction.

DOI: 10.54254/2753-8818/24/20231151

3.2. Exploratory data analysis

The descriptive statistics regarding diabetes, prediabetes, and healthy individuals are summarized in Table 1. As shown in Table 1, gender, age, weight, BMI, and waist circumference are all positively related to diabetic risk. The proportion of males is larger in the diabetes and prediabetes group than that in the healthy group. The average age, weight, BMI, and waist circumference are the highest in diabetes, and then in prediabetes and healthy individuals. However, while total energy, carbohydrate, and sugar intake tend to be lower in diabetic and prediabetic populations, cholesterol intake exhibits the opposite. This abnormality implies that nutritional factors may play critical roles in modulating the risk of diabetes, which is further investigated in this study.

Table 1. Descriptive statistics about three groups in the study.

Feature	**Diabetes(mean/std)**	**Prediabetes(mean/std)**	**Health(mean/std)**	**p-value**
Male*	2447(0.529)	3256(0.536)	11391(0.470)	<0.001
Age	60.827(14.562)	48.989(19.242)	38.380(20.140)	<0.001
Energy	1897.639(898.475)	2124.937(996.467)	2135.509(1013.678)	0.481
Weight	89.297(24.342)	84.042(22.434)	75.359(21.075)	<0.001
BMI	32.250(7.707)	30.031(7.351)	27.091(6.685)	<0.001
Waist circumference	108.882(16.495)	101.256(16.731)	92.409(16.642)	<0.001
Carbohydrate	223.629(110.671)	257.013(130.391)	262.177(129.304)	0.008
Sugar	90.276(65.976)	113.101(79.331)	116.858(79.305)	0.0016
Cholesterol intake	298.502(243.460)	296.887(239.974)	284.315(238.461)	<0.001
Blood cholesterol	181.508(46.075)	191.708(41.794)	183.981(40.954)	<0.001
Energy	1778.385(855.710)	1965.497(896.497)	1957.662(935.509)	0.589

*Male is labeled as 1 and female is labeled as 2. The statistics shown represent the total sum.

DOI: 10.54254/2753-8818/24/20231151

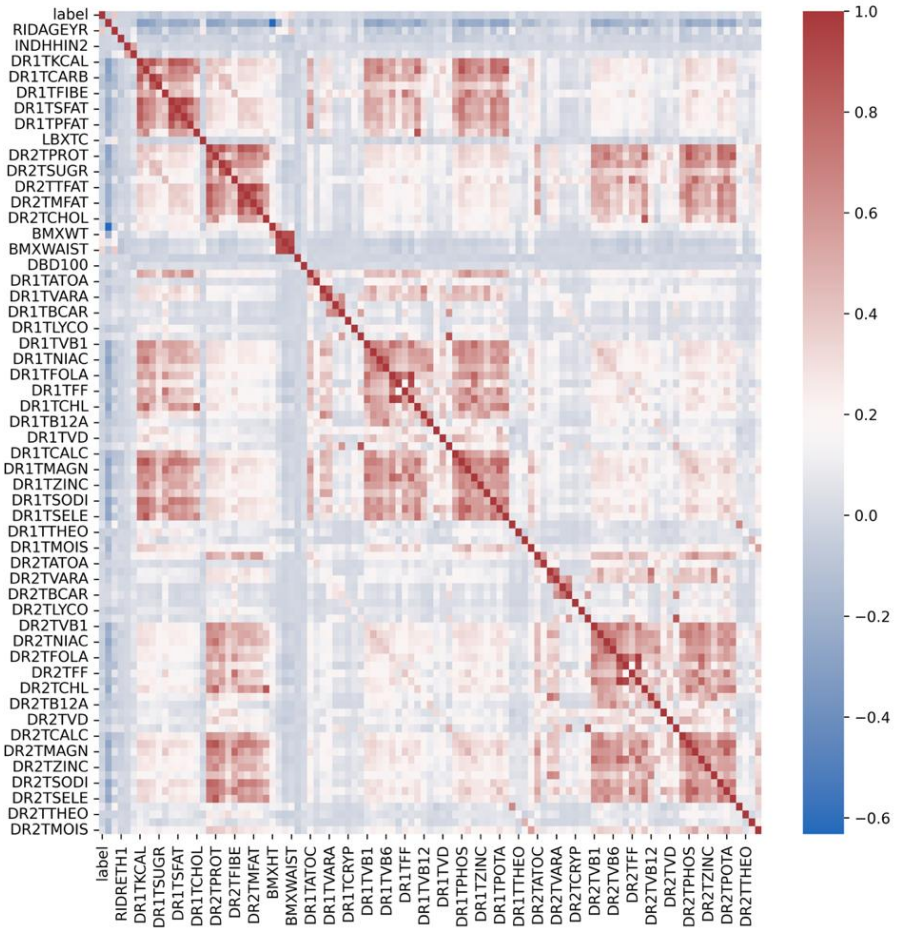

Figure 4. Heatmap of exploratory correlation analysis.

Correlations among variables are analyzed to seek important variables that contribute to diabetes. As shown in Figure 4, the correlation coefficients range from -0.6 to 1. Among all features, weight (kg), Body Mass Index (kg/m**2), and waist Circumference (cm) show the highest correlations with label ranges from 0.2 to 0.4 among all variables. In addition, a strong correlation that nearly reaches 1.0 also exists between dietary factors such as total calorie intake and sugar intake and, also, among weight (kg), Body Mass Index (kg/m**2), and waist Circumference (cm). The remaining factors show no observable correlation with the label.

3.3. Machine learning modeling prediction and Feature importance ranking

To select the best ML model for optimal prediabetes prediction, evaluation and comparison of six ML models were done, including Random Forest, KNN, SVM, gradient boosting, LDA, and Xgboosting. ROC AUC was used as a metric to compare the ML models. The six models have performance scores of 0.71, 0.6, 0.45, 0.63, 0.68, and 0.65 respectively as shown in Figure 5. RF achieved the highest performance a score of 0.71 and SVM has the lowest score of 0.67. Therefore, Random Forest was selected for ranking the importance of features. Figure 5 also shows 20 top variables in the feature importance ranking model. The importance score ranges from 0.2227 to 0.0053.

DOI: 10.54254/2753-8818/24/20231151

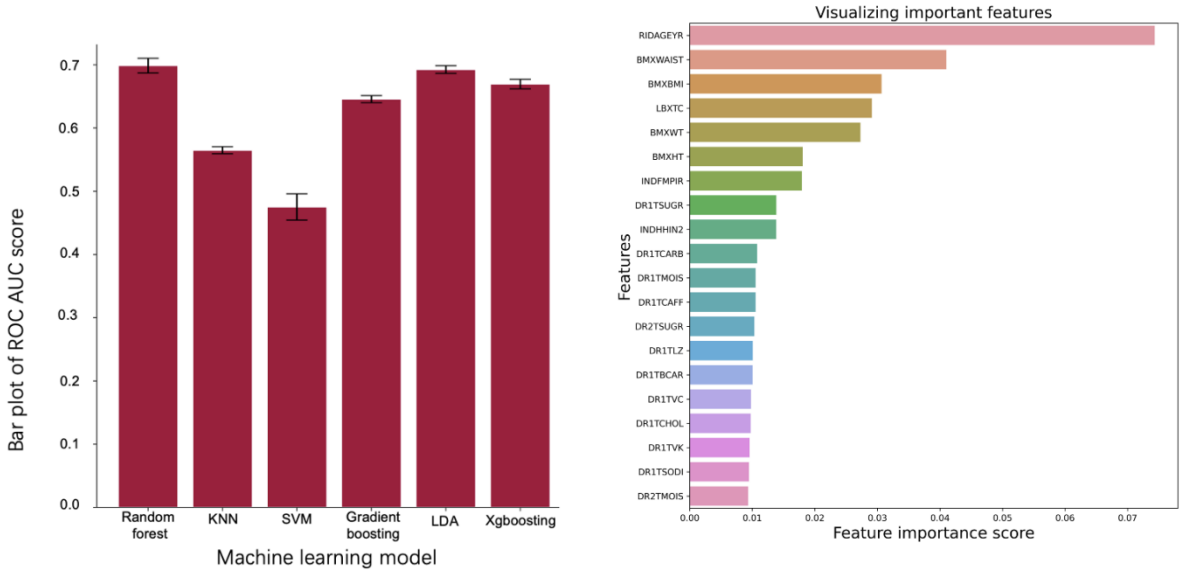

Figure 5. ROC AUC score of machine learning models and feature importance ranking.

3.4. Insights regarding nutritional factors associated with prediabetes

Based on the feature importance ranking, I further investigated the nutritional factors that are highly related with risk of prediabetes shown in Table 2. Daily cholesterol consumption has a positive relationship with prediabetes and diabetes, and it is supported by past meta-analyses for 5 studies [21]. Daily caffeine intake has a positive relationship with prediabetes and diabetes. It is hard to say whether the consumption of caffeine rises diabetic risk or not because studies on both sides of the story exist. A number of studies suggest that caffeine helps to reduce diabetic risks. Published in 2009, a study of 40,000 participants indicated that daily consumption of three cups of coffee or tea reduces the risk of type 2 diabetes by 42% [22]. A 2014 study also shows an 11% risk reduction associated with an increase in coffee consumption by more than 1 cup/day over a 4-year period compared with those who didn't change their consumption. (Participants who reduced their coffee consumption by more than 1 cup/day (median change eaquals 2 cups/day reduction) had a 17% (95% CI 8%, 26%) higher type 2 diabetes risk.) [23]. However, some hold the idea that caffeine intake could cause abnormalities in glucose levels [24]. According to James, caffeine could cause transient insulin resistance thus increasing prediabetic risk. A 2012 study also found that abstinence from caffeine improves control of chronic glucose in its diabetic research subjects [25]. A randomized controlled trials (RCTs) study investigated the effect of caffeine on insulin sensitivity in healthy humans without diabetes and the result suggests a short-term shift from glycemic homeostasis toward hyperglycemia caused by caffeine consumption [26]. A study also revealed that caffeine consumption boosts glucose levels in diabetic subjects by 16-28%, insulin concentration by 19-48%, and reduces insulin sensitivity by 14-37% [27]. Three possible mechanisms of how caffeine changes insulin production is offered. Caffeine raises levels of certain stress hormones, like epinephrine (also called adrenaline), which can prevent cells from processing as much sugar, keeping your body from making as much insulin. It also blocks a chemical called adenosine, which affects insulin production. Further, it is possible that excess caffeine consumption lowers sleep quality and thus lowers insulin production. Combined with the fact that caffeine is frequently used as a flavor enhancer—in the US, over 60% of soft drinks sold contain caffeine—it is hard to trace the real cause of the trend observed in this study. Nevertheless, it is for sure that caffeine intake should be something to take care within everyone's dietary plan. Further investigation into caffeine intake and diabetic risk is definitely necessary.

DOI: 10.54254/2753-8818/24/20231151

Daily thiamin, or Vitamin B1, consumption has a negative relationship with prediabetes and diabetes risk. Research reports have long been indicating natural products that are rich in polyphenols and vitamins can help to manage diabetes. Vitamin B1 is especially effective as a supplement for type 2 diabetes. Its antiglycation effect on glucose-induced glycation has been tested in the laboratory with various parameters [28]. Further, another paper suggests that external supplementation of vitamin 1 helps amend clinical symptoms of diabetes [30]. This nutrient does not only help people who are already diabetic but also reduces diabetic risk in non-diabetic populations. Women with a 2-fold increase in individual nutrient intake—including vitamin B1—show a lower prevalence of type 2 diabetes and other diseases [31]. Folic acid consumption, also, shows a negative relationship between prediabetes and diabetes risk as the supplementation of it shows its potency of glycemic control in adults. Experiment subjects received lowered fasting blood glucose and fasting insulin levels [32].

Table 2. Key nutritional factors associated with diabetes and prediabetes.

Feature	Diabetes(mean/std)	Prediabetes(mean/std)	Health(mean/std)	p-value
Carbohydrate	223.629(110.671)	257.013(130.391)	262.177(129.304)	0.008160
Sugar	90.276(65.976)	113.101(79.331)	116.858(79.305)	0.001591
Cholesterol intake	298.502(243.460)	296.887(239.974)	284.315(238.461)	0.000470
Blood cholesterol	181.508(46.075)	191.708(41.794)	183.981(40.954)	<0.001
Energy	1778.385(855.710)	1965.497(896.497)	1957.662(935.509)	0.589117
Thiamin (Vitamin B1)	1.498(0.805)	1.595(0.902)	1.603(0.941)	0.561783
Folic acid	159.791(160.703)	175.366(176.974)	189.795(194.263)	<0.001
Caffeine	140.527(184.936)	148.831(212.813)	117.589(184.237)	<0.001
Alcohol	5.438(21.541)	8.378(23.413)	8.805(27.705)	0.242952

4. Conclusions

This study used the NHANES data to identify the key nutritional factors associated with prediabetes and bulit effective machine learning model for prediabetes risk prediction. Random forest was selected the optimal model with an AUC score of 0.71. The key nutritional factors were identified, including thiamin, folic acid, and caffeine, which shows significant difference between healthy control and prediabetes. It is recommended to take caution when taking nutrients with these gradients. Since NHANES is a multi-center cross-sectional study, long-term studies to observe the outcome about the lifestyle changes are needed. The current study provides an efficient method and provide insights to reduce the risk of prediabetes by nutritional facors.

Acknowledgment

It's not uncommon to be prediabetic in our family. My father and my grandparents were all diagnosed with prediabetes and are controlling their diet since I was in middle school. However, I wasn't aware of the importance of preventing prediabetes from developing into type 2 diabetes until I learned about the stunning amount of people who cannot afford treatments at the stage of type 2 diabetes when I was doing research for my debate tournaments. It is critical for people to put more attention on the issue of prediabetes as it is more reversible and affordable.

I would like to first thank my parents, my sister, and my friends for providing me with consistent support. They encouraged me several times when I feel lost or confused for the purpose and meaning of my project. It is impossible for me to finish this project without them.

I would also like to extend my thanks to my advisor, Mr. Hagen., for guiding me and giving me practical suggestions throughout the process.

DOI: 10.54254/2753-8818/24/20231151

References

[1] Cho, Nam H., et al. "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045." Diabetes research and clinical practice 138 (2018): 271-281.

[2] Type, C. D. C. "Diabetes. Centers for Disease Control and Prevention, 2022." (2022).

[3] Lawrence, Jean M., et al. "Trends in prevalence of type 1 and type 2 diabetes in children and adolescents in the US, 2001-2017." Jama 326.8 (2021): 717-727.

[4] Saeedi, Pouya, et al. "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas." Diabetes research and clinical practice 157 (2019): 107843.

[5] Type, C. D. C. "Diabetes. Centers for Disease Control and Prevention. Published May 30, 2019." (2).

[6] de Vegt, Femmie, et al. "Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: The Hoorn Study." Jama 285.16 (2001): 2109-2113.

[7] May, Ashleigh L., Elena V. Kuklina, and Paula W. Yoon. "Prevalence of cardiovascular disease risk factors among US adolescents, 1999− 2008." Pediatrics 129.6 (2012): 1035-1041.

[8] National and State Diabetes Trends CDC. https://www.cdc.gov/diabetes/library/reports/reportcard/national-state-diabetes-trends.html (2022).

[9] Nichols, Gregory A., Teresa A. Hillier, and Jonathan B. Brown. "Progression from newly acquired impaired fasting glusose to type 2 diabetes." Diabetes care 30.2 (2007): 228-233.

[10] Wu, Jiahua, et al. "A prediction model for prediabetes risk in middle-Aged and elderly populations: a prospective cohort study in China." International Journal of Endocrinology 2021 (2021).

[11] Tobore, Igbe, et al. "Towards adequate prediction of prediabetes using spatiotemporal ECG and EEG feature analysis and weight-based multi-model approach." Knowledge-Based Systems 209 (2020): 106464.

[12] Choi, Soo Beom, et al. "Screening for prediabetes using machine learning models." Computational and mathematical methods in medicine 2014 (2014).

[13] Maeta, Katsutoshi, et al. "Prediction of glucose metabolism disorder risk using a machine learning algorithm: pilot study." JMIR diabetes 3.4 (2018): e10212.

[14] Abbas, Mostafa, et al. "Simple risk score to screen for prediabetes: A cross-sectional study from the Qatar Biobank cohort." Journal of Diabetes Investigation 12.6 (2021): 988-997.

[15] Centers for Disease Control and Prevention. "National Health and Nutrition Examination Survey: Overview." Hyattsville: National Center for Health Statistics (2016).

[16] Prevalence of Prediabetes Among Adults | Diabetes | CDC. https://www.cdc.gov/diabetes/data/statistics-report/prevalence-of-prediabetes.html (2022).

[17] Anguita, Davide, et al. "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines." DMIN. 2009.

[18] Prusty, Sashikanta, Srikanta Patnaik, and Sujit Kumar Dash. "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer." Frontiers in Nanotechnology 4 (2022): 972421.

[19] Spearman's Rank Correlation Coefficient: Definition, Interpretation, Formula, Examples. Embibe Exams https://www.embibe.com/exams/spearmans-rank-correlation-coefficient/ (2022).

[20] Liaw, Andy. "Wiener M." Classification and regression by randomforest R News 2.3 (2002): 18.

[21] Tajima, Ryoko, et al. "High cholesterol intake is associated with elevated risk of type 2 diabetes mellitus–a meta-analysis." Clinical nutrition 33.6 (2014): 946-950.

[22] Van Dieren, S., et al. "Coffee and tea consumption and risk of type 2 diabetes." Diabetologia 52.12 (2009): 2561-2569.

[23] Bhupathiraju, Shilpa N., et al. "Changes in coffee intake and subsequent risk of type 2 diabetes: three large cohorts of US men and women." Diabetologia 57 (2014): 1346-1354.

[24] Lane, James D. "Caffeine, glucose metabolism, and type 2 diabetes." Journal of caffeine research 1.1 (2011): 23-28.
[25] Lane, James D., et al. "Pilot study of caffeine abstinence for control of chronic glucose in type 2 diabetes." Journal of caffeine research 2.1 (2012): 45-47.
[26] Shi, Xiuqin, et al. "Acute caffeine ingestion reduces insulin sensitivity in healthy subjects: a systematic review and meta-analysis." Nutrition journal 15.1 (2016): 1-8.
[27] Whitehead, N., and H. White. "Systematic review of randomised controlled trials of the effects of caffeine or caffeinated drinks on blood glucose concentrations and insulin sensitivity in people with diabetes mellitus." Journal of Human Nutrition and Dietetics 26.2 (2013): 111-125.
[28] Abdullah, K. M., et al. "Insight into the In vitro antiglycation and in vivo antidiabetic effects of thiamine: Implications of vitamin B1 in controlling diabetes." ACS omega 6.19 (2021): 12605-12614.
[29] Deshmukh, Shreeya V., Bala Prabhakar, and Yogesh A. Kulkarni. "Water soluble vitamins and their role in diabetes and its complications." Current Diabetes Reviews 16.7 (2020): 649-656.
[30] Olsen, Birthe S., et al. "Thiamine-responsive megaloblastic anaemia: a cause of syndromic diabetes in childhood." Pediatric Diabetes 8.4 (2007): 239-241.
[31] Nguyen, Hai Duc, Hojin Oh, and Min-Sun Kim. "Higher intakes of nutrients are linked with a lower risk of cardiovascular diseases, type 2 diabetes mellitus, arthritis, and depression among Korean adults." Nutrition Research 100 (2022): 19-32.
[32] Asbaghi, Omid, et al. "Folic acid supplementation improves glycemic control for diabetes prevention and management: a systematic review and dose-response meta-analysis of randomized controlled trials." Nutrients 13.7 (2021): 2355.