Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230214

Predictions of diabetes through machine learning models based on the health indicators dataset

Xinyi Ren

Lancaster University, Lancashire, LA1 4YW, The United Kingdom

1625201499@qq.com

Abstract. Diabetes is a chronic disease that is widespread in the United States. Patients with diabetes will lose the ability to effectively regulate blood glucose levels and the disease can lead to increased economic burden for patients and generate enormous public health impact. The main purpose of this paper is to find out the indicators that are highly associated with diabetes and build a model to predict diabetes. The original dataset is from BRFSS (the Behavioral Risk Factor Surveillance System). For this project, a cleaned dataset on Kaggle for the year 2015 was used, which has 253,680 survey responses to CDC (Centers for Disease Control and Prevention)'s BRFSS with the target variable diabetes and 21 feature variables. The Chi-square test was applied to investigate the association between feature indicators and diabetes and built several machine learning models for predicting the disease. The selected model is Cat Boost Classifier with 86.6% accuracy for the testing set. According to the Permutation Feature Importance based on the Cat Boost Classifier, the most important 5 features were General Health (GenHlth), BMI (Body Mass Index), Age, high blood pressure (HighBP), and high cholesterol (HighChol) variables.

Keywords: diabetes prediction, machine learning, health indicators, classification.

1. Introduction

Diabetes is one of the most common and widespread chronic diseases in the US [1]. A diabetes patient does not produce enough insulin to regulate sugar in the body and can have many complications. In 2018, among the US population, 34.3 million people of all ages had diabetes, taking up 10.5% of the population [2,3], which increases the risk of complications of diabetes such as heart disease, cardiovascular events, microvascular disease, and even premature death. It can be seen that diabetes has a great impact in the US and poses a threat to personal health. The Behavioral Risk Factor Surveillance System (BRFSS) is the largest health-related telephone survey system in the United States, completing more than 400,000 adult interviews each year [4]. The primary purpose of this system is to collect data on health-related risk behaviours, chronic conditions such as diabetes, and the use of preventive services among US residents.

In this project, the 2015 BRFSS dataset, with 21 health indicators that may be associated with diabetes, was analysed. The goal is to identify the indicators highly correlated with diabetes and develop predictive models for diabetes, which could help facilitate early diagnosis and intervention. There are 23580 records from the 2015 BRFSS dataset. Diabetes generally includes two main types: type 1 and type 2, accounting for 5% and 95% respectively [2]. The dataset in this project does not separate type 1

DOI: 10.54254/2755-2721/32/20230214

and type 2 diabetes. Due to the extremely high proportion of type 2, the characteristics analysed in this paper are more likely to be applicable to type 2 diabetes. The features that are highly correlated with diabetes were found by using correlation analysis and the Chi-square test. Through these selected feature variables, several models of supervised machine learning algorithms were built for predicting diabetes, including Gaussian Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, Linear Discriminant, and Cat Boost classifiers. These classifiers were evaluated by training and testing accuracy and Mean Squared Error to find the best model, thus providing help for the early diagnosis of diabetes, which is important to the prevention of the onset of complications.

2. Diabetes health indicators dataset

The US Diabetes Health Indicators Dataset, originally collected by BRFSS, is cleaned and published by Alex Teboul on Kaggle [5]. The dataset is a collection of 23580 records of adult respondents in the United States. It has 22 variables in total, including 21 feature variables and 1 label variable indicating diabetes or not [5]. The binary target variable takes value "1", indicating a positive result for diabetic or prediabetic, while "0" indicates a negative result for non-diabetic. Figure 1 shows the unbalanced distribution of diabetes cases in the dataset (86.1% for non-diabetic, 13.9% for diabetic and prediabetic). The 21 indicators are indicated by numeric and categorical variables respectively. The only numeric variable is the BMI index which has an interpretation of BMI status. As seen in Figure 2, BMI below 18.5 means underweight, 18.5–24.9 means healthy weight, 25.0–29.9 means overweight, and 30.0 and above means obesity.

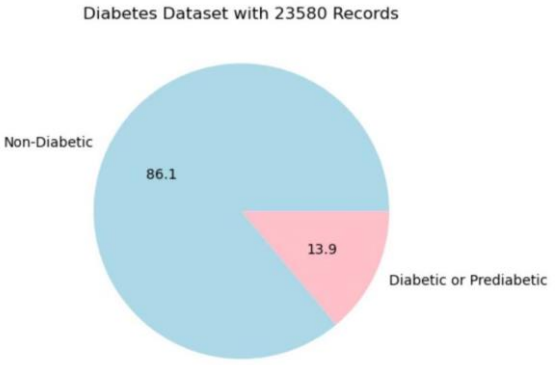

Figure 1. Distribution of Diabetes Dataset.

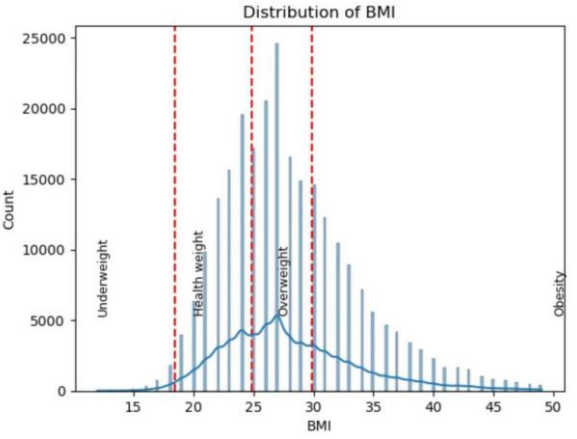

Figure 2. Distribution of BMI in Dataset.

The other 20 categorical features include 4 aspects. The features related to personal information are age, gender, education level, income, having health care service or not, and having financial difficulties to see a doctor or not. The binary features of physical disease denote high blood pressure, high cholesterol, cholesterol check, stroke, heart disease attack, and whether the patients have difficulty walking or not. The features that show the respondents' self-assessment of their health status are general health, mental health, and physical health. The binary features of personal habits are physical activity, smoking, fruits, veggies, and heavy alcohol consumption.

3. Method

3.1. Feature selection

Feature selection is a crucial step for removing irrelevant features and picking a subset of highly discriminant features for the target variable from the original dataset. The advantages of feature selection are a reduction in the execution time of the classifier and an improvement in model accuracy.

DOI: 10.54254/2755-2721/32/20230214

3.1.1. Feature correlation. Before selecting the feature, the correlation diagram (Figure 3) was plotted to check the correlation between the target variable "Diabetic" and 21 feature variables. The features of health issues and diseases, such as general health, high BP, walking difficulty, high BMI, high Cholesterol, and heart disease or attack, are highly and positively correlated with diabetes. The features denoted that personal habits have much less correlation with having diabetes. Among the features of personal information, education level and income are more correlated with diabetes in a negative way. Overall, the correlation diagram illustrates some indicators of influence in predicting diabetes, which is helpful feature selection.

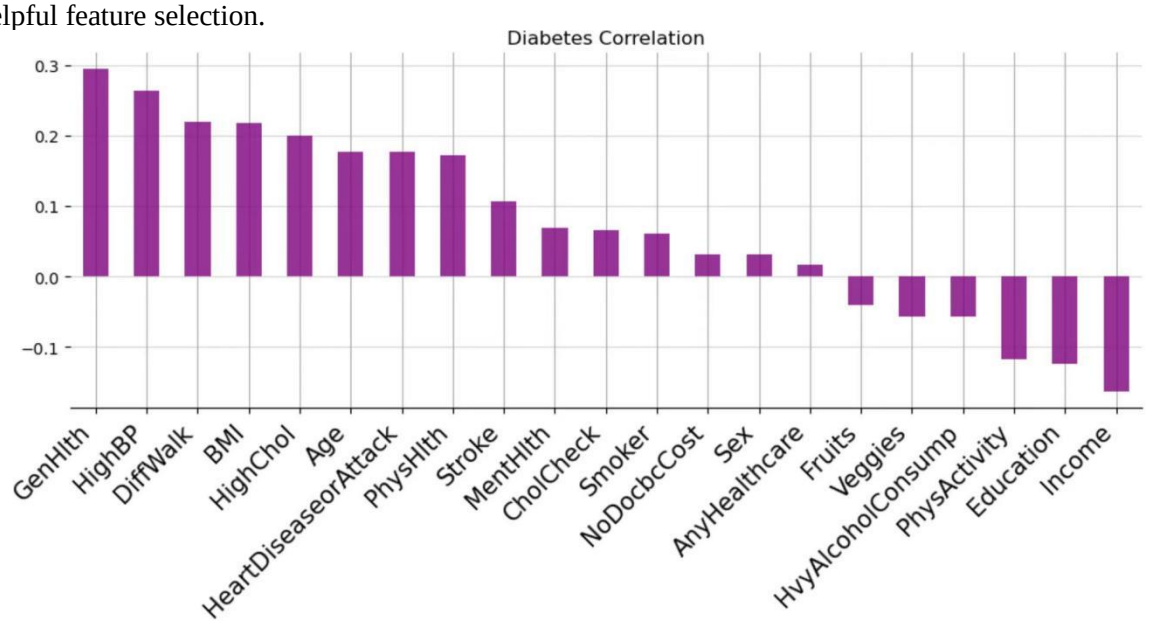

Figure 3. Correlation Diagram.

3.1.2. Chi-square test. It is not enough to select features only based on the correlation diagram. For further feature analysis, the Chi-square test was applied to this project, which helps to solve the problem in feature selection by testing the relationship between the feature variable and the target variable [6]. A Chi-square test is used in this project to test the independence of two features.

The formula for calculating a Chi-square statistic is [7]:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where O denotes the observed values, and E denotes the expected values.

By calculating the Chi-square value, the relationship between the independent category feature and the dependent category feature can be determined. In feature selection, the goal is to select the features with a high Chi-square value, which are highly dependent on the target "Diabetic". The Chi-square score of 21 feature variables is calculated, and Table 1 displays the best 15 features to be selected and their Chi-square scores. The other six features will be removed from the dataset due to their low Chi-square scores; they will not participate in the training of the models. The dropped features also show a lower correlation in the previous correlation diagram, which is expected.

DOI: 10.54254/2755-2721/32/20230214

Table 1. Feature with Chi-square Score.

Feature	Chi-square Score
PhysHlth	133424.41
MentHlth	21029.63
BMI	18355.17
DiffWalk	10059.51
HighBP	10029.01
GenHlth	9938.51
Age	9276.14
HeartDiseaseAttack	7221.98
HighChol	5859.71
Income	4829.82
Stroke	2725.23
PhysActivity	861.89
HvyAlcoholConsump	779.42
Education	756.04
Smoker	521.98

3.2. Data preprocessing

3.2.1. Data splitting. Before starting to train the model, the data set needs to be preprocessed. The dataset is cleaned without any missing values. And randomly shuffling the dataset ensures the random distribution of the data. In machine learning, if the data set is not shuffled, the "bias" of the model might occur in the training process, which reduces its generalisation ability, thereby reducing the training accuracy. For example, if a classification model was built in which the first 80% of the initial data is the first class and the last 20% is the second class, the accuracy of the model will be extremely low. The low correlation feature variables in Section 2 dropped, and the best 15 features remained as the final feature variables in the model training and testing. The data was divided into target variable data as 'X' and 15 feature variable data as "y". The portion of the dataset to allocate to the test set is 20%, and the counterpart of the train set is 80%.

3.2.2. SMOTE algorithm. Recall Section 2, the number of non-diabetic records is much greater than that of diabetic records. The imbalance of the training dataset might lead to bias and the poor performance of model training. Therefore, before building the model, the imbalance of the dataset needs to be resolved. The imbalance can be handled by undersampling and oversampling strategies. The undersampling is to select data randomly from the majority class until two classes have the same number of records [8]. However, the reduction of the majority class might lose useful information. Thus the SMOTE (Synthetic Minority Oversampling Technique) preprocessing algorithm was used to balance the training data. The basis of SMOTE was to generate synthetic samples for the minority class until the data is balanced [8]. Moreover, it can assist the classifiers to improve their generalisation capacity. Figure 4 displays the rebalance of diabetes distribution before and after oversampling.

DOI: 10.54254/2755-2721/32/20230214

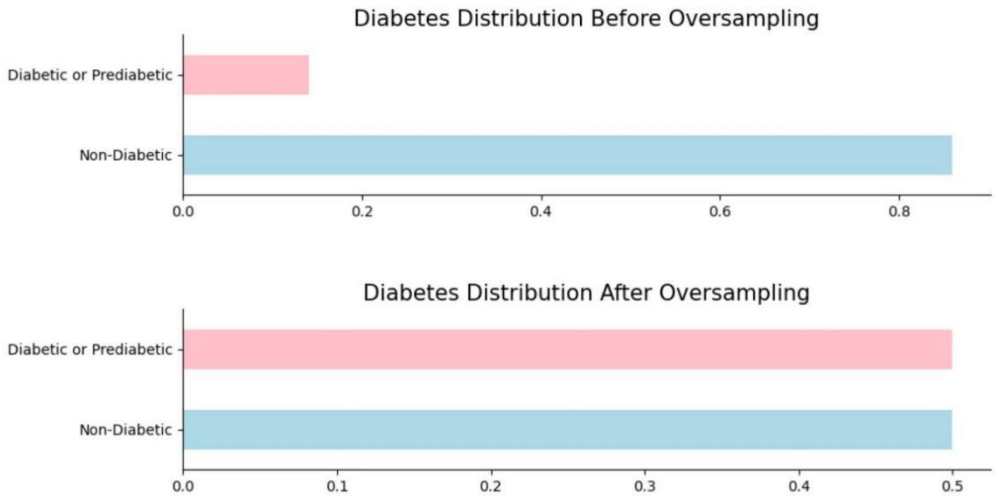

Figure 4. Diabetes Distribution Before and After Oversampling.

3.2.3. Feature scaling. The final step for data preprocessing is feature scaling which can transform features of a dataset to improve the performance of machine learning models and reduce the time for training models. If the original index value is directly used to train models, the features with a large value will be emphasised, and the features with a small value will be weakened. The feature scaling ensures that all features contribute equally to the model and prevents the domination of features with larger values [9]. The feature scaling method used in this project is standardisation scaling.

3.3. Models and evaluation

3.3.1. Model training and results. Several supervised machine learning classifiers have been applied to predict diabetes with the processed training set and testing set, including Gaussian Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, Linear Discriminant, and Cat Boost classifiers. The accuracy, precision, and mean square error measures are applied for analysing the performance of the models above. In order to calculate accuracy and precision for training and testing data, there are four cases that need to be considered [10].

True positive (TP): record is classified as positive and is actually positive.
False positive (FP): record is classified as positive and is actually negative.
True negative (TN): record is classified as negative and is actually negative.
False negative (FN): record is classified as negative and is actually positive.
The accuracy is the proportion of the total number of predictions that were correct:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (2)$$

The precision is the proportion of positive records that were correctly predicted as positive:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (3)$$

The four classifiers with an accuracy of more than 80% are the Cat Boost Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Decision Tree Classifier. As seen in Table 2, the training accuracy of the Decision Tree and Random Forest Classifier is close to 100%, but the accuracy of the two classifiers decreases for the test set, and the precision is very low. Therefore, these two models may be overfitting. Overfitting refers to matching the data of the training set too closely and precisely so that it cannot fit the data of the testing set well. The Cat Boost Classifier has the highest accuracy (86.6%) and highest precision (55.6%). Thus the selected model is the Cat Boost Classifier.

DOI: 10.54254/2755-2721/32/20230214

Table 2. Summary of Models.

Model	Train Accuracy	Train Precision	Test Accuracy	Test Precision
Cat Boost	0.925	0.978	0.866	0.556
Random Forest	0.990	0.994	0.849	0.420
Gradient Boosting	0.886	0.901	0.838	0.421
Decision Tree	0.990	0.997	0.802	0.298
Gaussian Naïve Bayes	0.717	0.731	0.741	0.308
Logistic Regression	0.752	0.740	0.729	0.307
Linear Discriminant	0.752	0.734	0.721	0.302

3.3.2. Permutation feature importance. Permutation Feature Importance (PFI) is a method applied to calculate the importance of features independent of the model type. As a result of the Cat Boost classification model, feature importance measures how much each feature affects the target. The PFI was helpful to know which features are more important in the selected model and which features affect the prediction results, thus interpreting the performance of the model. The following steps are applied to calculate the feature importance score [11]:

1. Use the selected model and record the original score of the model.
2. Shuffle the value of a feature, use the model to predict again, and calculate the score on the test set. The reduction of model performance represents the importance of this feature.
3. Restore the value of the disrupted feature, and repeat step 2 on the next feature until the importance of each feature is obtained.

Accuracy was selected as a representation of model performance, which is the "score" for permutation. In that case, the features were sorted from high to low based on the decrease in accuracy, and the top feature has the highest importance. For each feature variable, the reduction in accuracy was to calculate and record for 30 random shuffles [11]. To visualize these records, boxplots (Figure 5 and Figure 6) are created for the train set and test set respectively.

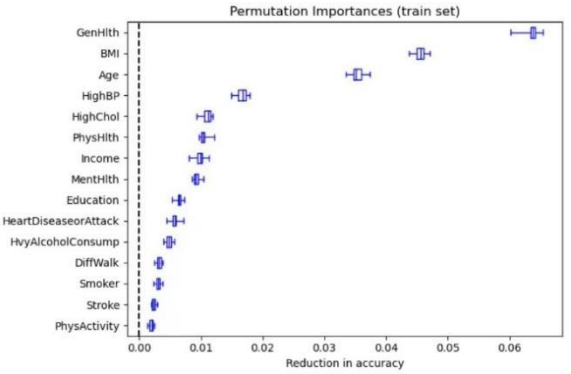

Figure 5. PFI for Train Set.

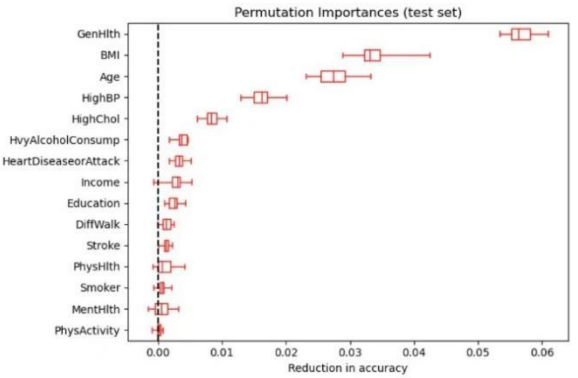

Figure 6. PFI for Test Set.

According to the boxplot of the permutation importance feature for the Cat Boost model, the most important feature was General Health (GenHlth), followed by BMI, Age, high blood pressure (HighBP), and high cholesterol (HighChol) variables. The permutation feature importance in the train set and test set is very similar.

DOI: 10.54254/2755-2721/32/20230214

4. Conclusion

This paper applies a dataset that represents the distribution of diabetes disease from BRFSS in 2015. The goal of this paper is to find the features that are correlated with diabetes and train a predictive model with a reasonable level of accuracy. The correlation diagram and the Chi-square test have been applied to select the best 15 features from 21 features in total. After the data preprocessing and model training, the results suggest that the Cat Boost classifier is suitable for predicting diabetes with the selected features. However, there are still some limitations existing in the model. The precision of the model is not as expected. One of the possible reasons is that the data set is unbalanced. The SMOTE algorithm was implemented in the training set to rebalance the data, but the testing set is still unbalanced (it is incorrect to use SMOTE for the full set, which affects the purity of the data set and leads to overfitting). The imbalance of the test set may cause the model to be less precise on the test set. In the future, the performance of the Cat Boost Classifier can be improved by collecting more data on diabetic patients.

References

[1] Kumari, V. A. and Chitra, R. (2013). Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA), 3, 1797-1801.

[2] U.S. Department of Health and Human Services Centers for Disease Control and Prevention. (2020). National Diabetes Statistics Report Estimates of Diabetes and Its Burden in the United States. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

[3] Fang, M., Wang, D., Coresh, J. and Selvin, E. (2021). Trends in Diabetes Treatment and Control in U.S. Adults, 1999-2018. The New England journal of medicine, 384(23), 2219–2228.

[4] National Center for Chronic Disease Prevention and Health Promotion. Division of Population Health. https://www.cdc.gov/brfss/index.html.

[5] Teboul, A. (2021). Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-datase t/code.

[6] Bahassine, S., Madani, A., Al-Sarem, M. and Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University - Computer and Information Sciences Volume, 32(2), 225-231.

[7] Kumar Gajawada, S. (2019). Chi-Square Test for Feature Selection in Machine learning. Published in Towards Data Science. https://j-pcs.org/temp/JPractCardiovascSci1169-9537648_023857.pdf.

[8] Satwik, M. (2017). Handling Imbalanced Data: SMOTE vs. Random Undersampling. International Research Journal of Engineering and Technology (IRJET), 4(8), 317-320.

[9] Hanan, A., Yuan, X. H., Esterline, A., Khorsandroo, S. and Lu, X. C. (2021). Studying the Effects of Feature Scaling in Machine Learning. Ph.D. Dissertation. North Carolina Agricultural and Technical State University. Advisor(s) Xu, Jinsheng. Order Number: AAI28772109.

[10] Gürsoy, M. İ. and Alkan, A. (2022). Investigation Of Diabetes Data with Permutation Feature Importance Based Deep Learning Methods. Karadeniz Fen Bilimleri Dergisi. The Black Sea Journal of Sciences. ISSN (Online): 2564-7377.

[11] Li, S. (2022). Best Practice to Calculate and Interpret Model Feature Importance: With an example of Random Forest model. Published in Towards Data Science. https://towardsdatascience.com/best-practice-to-calculate-and-interpret-model-feature-importance-14f0e11ee660.