Predicting Type 2 Diabetes Mellitus using Machine Learning Algorithms

Dr. Nisreen Sulayman[*]

(Received 18 / 8 / 2022. Accepted 3 / 10 / 2022)

□ ABSTRACT □

Purpose: to build an effective prediction model based on machine learning (ML) algorithms for the risk of type 2 (non-insulin-dependent) Diabetes Mellitus (T2DM).
Methods: I developed two machine learning prediction models based on extreme gradient boosting (XGBoost) and logistic regression (LR). To evaluate the ML prediction models I used the Pima Indian Diabetes dataset (PIDD). The dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases and consists of 500 non-diabetic patients and 268 diabetes patients.
Results: Models' performance was evaluated using six performance criteria. XGBoost model outperforms the logistic regression. The XGBoost model achieved: area under receiver operating characteristic curve (AUROC) = 85%, sensitivity = 71%, specificity = 81%, accuracy =77%, precision = 67%, and F1-score=69% respectively.
Conclusion: This study showed that the XGBoost ML algorithm can be applied to predict individuals at high risk of T2DM in the early phase, which has a strong potential to control diabetes mellitus.

Keywords: Type 2 Diabetes Mellitus, Machine Learning, XGBoost Model, Logistic Regression

[*] **Doctor- Biomedical Engineering Department, Mechanical and Electrical Engineering Faculty, Damascus University, Damascus, Syria. sulayman.nisreen@gmail.com**
nisreen.sulayman@damascusuniversity.edu.sy

التنبؤ بمرض السكري من النوع الثاني باستخدام خوارزميات تعلم الآلة

د. نسرين سليمان[*]

(تاريخ الإيداع 18 / 8 / 2022. قُبِل للنشر في 3/ 10 / 2022)

□ ملخّص □

الهدف: بناء نموذج فعال للتنبؤ بمرض السكري من النوع الثاني (غير المعتمد على الأنسولين) باستخدام خوارزميات تعلم الآلة.

طريقة البحث ومواده: تم تطوير نموذجي تنبؤ بمرض السكري من النوع الثاني باستخدام خوارزميتي تعلم الآلة: تعزيز التدرج الشديد والإنحدار اللوجستي. كما تم اختبار النموذجين باستخدام قاعدة بيانات لمرض السكري Pima Indian Diabetes dataset (PIDD) من المعهد الوطني للسكري وأمراض الجهاز الهضمي والكلى في الهند. تتألف قاعدة البيانات المستخدمة من 500 شخص غير مصاب بمرض السكري و268 مريض بالسكري من النوع الثاني.

النتائج والمناقشة: تم استخدام ستة بارمترات لتقييم أداء النموذجين. تفوق نموذج تعزيز التدرج الشديد على نموذج الإنحدار اللوجستي وكانت بارامترات أداءه على النحو الآتي: المساحة تحت المنحنى 85%، الحساسية 71%، النوعية 81%، الدقة 77%، الإحكام 67%، ومعامل F1 69% على التوالي. أظهرت الدراسة إمكانية استخدام نموذج تعزيز التدرج الشديد للتنبؤ بخطر الإصابة بمرض السكري من النوع الثاني.

الكلمات المفتاحية: مرض السكري من النوع الثاني، تعلم الآلة، نموذج تعزيز التدرج الشديد، الإنحدار اللوجستي

[*] عضو هيئة تدريسية-قسم الهندسة الطبية – كلية الهندسة الميكانيكية والكهربائية – جامعة دمشق- دمشق- سورية.
sulayman.nisreen@gmail.com, nisreen.sulayman@damascusuniversity.edu.sy

Introduction:

Diabetes mellitus is a chronic, metabolic disease characterized by excess levels of blood glucose. The most common is type 2 (non-insulin-dependent) Diabetes Mellitus (T2DM), usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. There is a globally agreed target to cease the rise in diabetes by 2025. About 422 million people have diabetes, the major part living in low-and middle-income countries, and 1.5 million deaths are attributed to diabetes each year. Both the number of cases and the prevalence of diabetes patients have been increasing over the past few decades [1].

The number of individuals with diabetes patients rose from 108 million in 1980 to 422 million in 2014. Prevalence has been rising sooner in low- and middle-income countries than in high-income countries. Diabetes is a major reason of kidney failure, heart attacks, blindness, stroke, and lower limb amputation [2]. 537 million adults are living with diabetes. This number will rise to 643 million by 2030, and 783 million by 2045. Over 3 in 4 adults with diabetes are in low- and middle-income countries [3].

The rising incidence of diabetes imposes a significant burden on the individuals, health system, and the whole society [4, 5]. T2DM is an irreversible disease but preventable [6]. Therefore, it is essential to have an effective model to predict the onset of T2DM in individuals, which helps in the early identification of people at high risk of T2DM. The drastic increase in the rate of individuals suffering from diabetes mellitus makes the demand to make a system using the most effective available technology such as machine learning algorithms which provide accurate diabetes prediction results very essential to avoid or reduce common comorbidities and complications of diabetes. Although plenty of research has been conducted on T2DM prediction, there are still existing obstacles, due to the study population disparity and the difference in dataset sources and features. Thus, further studies are still required to be done in this area.

The rest of this paper is arranged as follows: The related studies are discussed in section 2. A detailed description of the materials and methods is shown in section 3. Section 4 demonstrates the results. Section 5 presents a discussion about the results and compares them with the previously obtained from the literature**.** This paper is concluded in section 6.

1-Related work

In recent years, Machine Learning (ML) algorithms have been applied in the medical field. They have proven to be efficient in disease diagnosis [7,8], treatment [9,10], and prognosis [11,12]. Predictive models based on ML algorithms can be useful in the identification and prediction of the risk of T2DM in individuals [13].

Pronab Ghosh et al. (2021) compared different ML algorithms for detecting diabetes. They used four ML algorithms: Gradient Boosting (GB), Support Vector Machine (SVM) AdaBoost (AB), and Random Forest (RF). ML algorithms are evaluated using seven different types of performance metrics with a 10-fold cross-validation approach. The best results were obtained with the RF approach after the features were selected with the minimal redundancy maximal relevance feature selection approach [14].

Chen et al. (2017) proposed a hybrid prediction model to help the diagnosis of type 2 diabetes. In the proposed model, the K-means clustering algorithm is used for data reduction with J48 decision tree as a classifier for classification. To get the experimental result, they used the Pima Indians Diabetes Dataset (PIDD) from the UCI machine learning repository. The result shows that the proposed model has reached 90% accuracy compared to other studies [15].

Sisodia, D. and Sisodia, DS designed a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. They used three machine learning classification algorithms namely Decision Tree (DT), SVM, and Naive Bayes to detect diabetes at an early stage. Assessment was performed on Pima Indians Diabetes Database (PIDD). The performances of all three algorithms are evaluated on various measures like precision, accuracy, F-Measure, and sensitivity. Results obtained show the Naïve Bayes outperforms with the highest accuracy comparatively to other algorithms [16].

Karthikeyani, V., and Begum, I. P. (2013) compared the results of ten supervised data mining algorithms using five performance criteria. He used partial least squares (PLS) to extract features of PIDD, and Linear Discriminate Analysis (LDA) method to build a model for predicting T2DM. The PLS-LDA was the best one among the ten algorithms with an accuracy of 74.40%. The Best results are achieved by using the Tanagra tool (a data mining matching set) [17].

Materials and Methods:

The proposed procedure is summarized in figure 1. It shows the flow of the study conducted in constructing the machine learning predictive model.

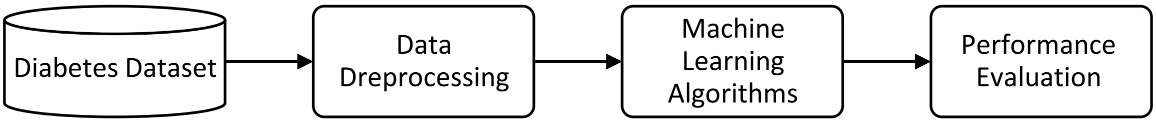


Figure 1. Diabetes Prediction Model.

1- Dataset

The dataset used in this study was obtained from Pima Indian Diabetes Dataset (PIDD) heritage. It is from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of the PIDD is to diagnostically predict whether or not an individual has diabetes. Several constraints were placed on the selection of the instances from a larger database. In particular, all patients in the dataset are females at least 21 years old. The dataset is available at the Kaggle repository [18].

The dataset consists of 500 non-diabetic patients and 268 diabetes patients. Each patient had eight medical predictor features and one target variable. Predictor features include the number of times pregnant, plasma glucose concentration at 2 Hours in oral Glucose Tolerance Test (GTT), diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (µh/ml), Body Mass Index (BMI) (Weight in kg / (Height in In)), diabetes pedigree function, and age (years). The target variable has a binary value of either zero or one indicating a non-diabetic\diabetes patients.

Table 1 represents descriptive statistics of the medical predictor features of the PIDD participants. Figure 2 shows a pair plot matrix of medical predictor features of the PIDD participants. It is helpful to clarify the pair-wise relationships of the medical features preliminarily.

Table 1. Descriptive statistics of the medical predictor features of the PIDD participants.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Mean	3.84	120.89	69.10	20.53	79.79	31.99	0.47	33.24
Std	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76
Min	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.078	21.00
Max	17.00	199.00	122.00	99.00	846.00	67.100	2.42	81.00

Std: Standard Deviation; Min: minimum value of the feature; Max: maximum value of the feature

2- Data preprocessing

For the successful use of the ML algorithms, data preprocessing is applied. Looking at table 1, the following features: glucose, blood pressure, skin thickness, insulin and BMI have an invalid zero as a minimum value which indicates missing value. Dealing with inconsistent values for the aforementioned features is done as follows: first replacing the zero values with Not a Number (NaN), then features distribution was examined by drawing a histogram of each feature: glucose concentration and diastolic blood pressure had a left-skewed distribution while skin thickness, insulin, and body mass index had a right-skewed distribution, and finally imputing NaN values with mean for glucose concentration and diastolic blood pressure; and with the median for skin thickness, insulin, and body mass index. To standardize the input features, the data were normalized using Python to mean 0 and variance 1 using the StandardScaler function from the Sklearn preprocessing library.

3- Machine learning algorithms

This section briefly discusses the ML algorithms which have been used in this study.

3-1 Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable, an efficient implementation of the gradient boosting ensemble algorithm. It is the leading machine learning algorithm for regression, classification, and ranking problems and is known as one of the best machine learning algorithms utilized for supervised learning. Data scientists prefer XGBoost because of its high performance and computational speed. A detailed description of how XGBoost works is available at [19].

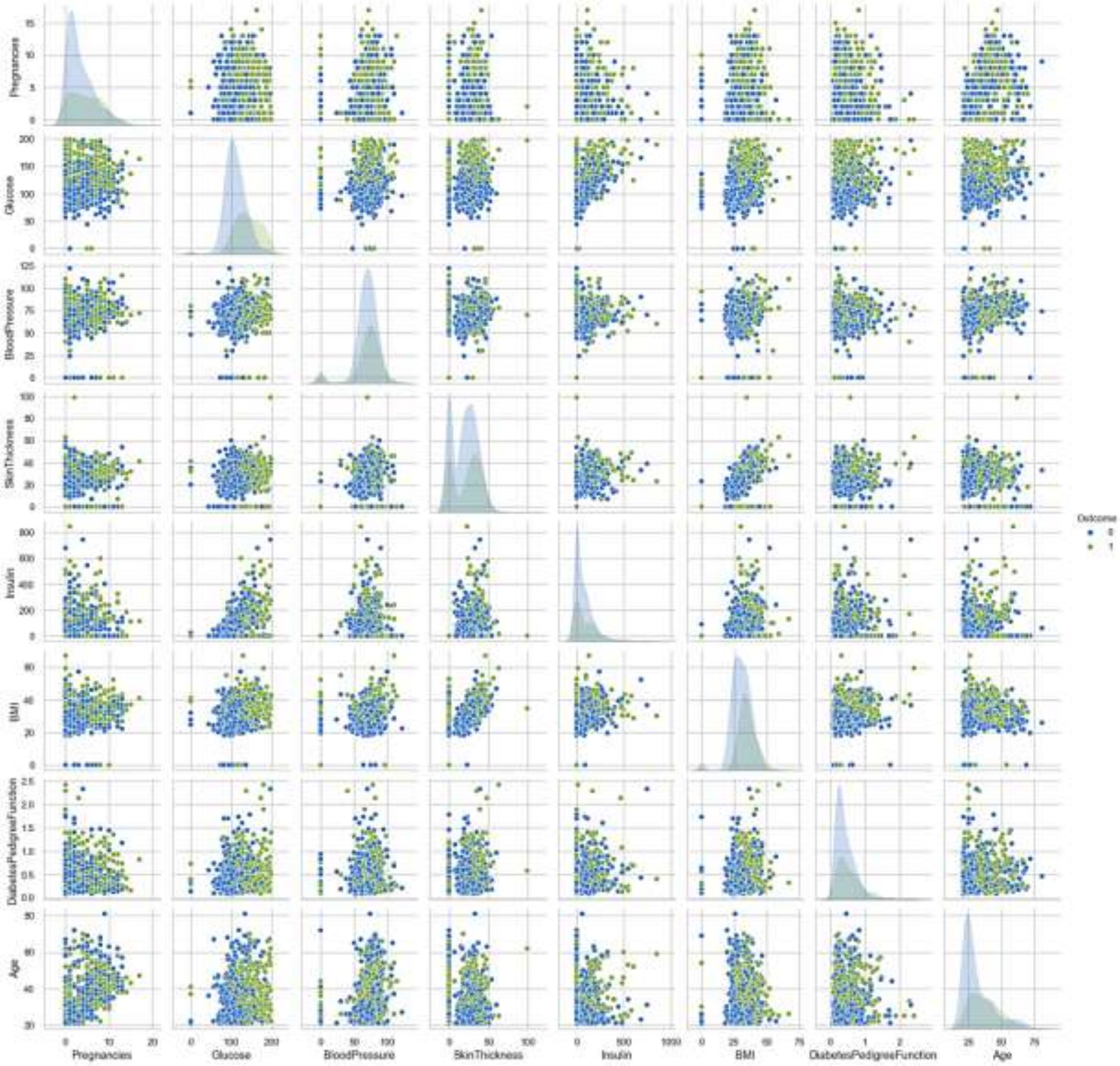


Figure 2. Pair plot matrix of medical predictor features.

3-2 Logistic Regression

Logistic regression (LR) is a traditional classification algorithm that measures the relationship between a categorical dependent variable (input features) and one or more independent variables (outcome(s)) based on the sigmoid function [20]. LR is a simple method for prediction that provides baseline accuracy values to compare with other non-parametric machine learning algorithms [21].

3-4 Evaluation metrics

The dataset was randomly divided into two parts: the training set accounted for 80% (n = 614) and the test set accounted for 20% (n = 154). The training set is used to train the logistic regression and XGBoost machine learning algorithms and the test set is used to evaluate the models. The training set is independent from the test set. The hyperparameters for XGBoost were as followed: learning_rate = 0.1, max_depth = 5, n_estimators = 10, seed=42.

Different performance metrics are considered for evaluating the prediction performance of logistic regression and the XGBoost ML models. The evaluating metrics include accuracy, sensitivity, precision, specificity, the area under receiver operating characteristic curve (AUROC), and F1-score. Sensitivity is the percentage of diabetes patients who are correctly predicted as having diabetes. Specificity is the percentage of non-diabetic patients who are correctly predicted as having no diabetes. Equations (1)-(5) refer to the definition of each metric.

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \qquad (1)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \qquad (2)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \qquad (3)$$

$$F1 - score = \frac{2TP}{2TP + FN + FP} \times 100\% \qquad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \qquad (5)$$

where TP is called true positive, denoting the number of diabetes patients who are correctly predicted as having diabetes, FN is called false negative, which determines the number of diabetes patients who are misclassified as having no diabetes, and (TP+FN) is the total number of diabetes patients. TN is called true negative and denotes the number of non-diabetic patients who are correctly predicted as having no diabetes, FP is called false positive, denoting the number of non-diabetic patients who are misclassified as having the diabetes, and (TN+FP) is the total number of non-diabetic patients. Accuracy is the percentage of correct predictions, and F1 score is the balance between precision and sensitivity.

Results and discussion:

Results:

Table 2 represents different performance metric values of XGBoost and logistic regression ML algorithms calculated on various measures. It illustrates that XGBoost and logistic regression have the same accuracy but XGBoost has a higher sensitivity and F1score compared to logistic regression.

Table 2. Prediction results using XGBoost and Logistic Regression machine learning algorithms.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)
XGBoost model	**77**	**71**	**81**	**67**	**69**
Logistic Regression	75	62	80	67	64

Figure 3 and figure 4 show the confusion matrix and receiver operating characteristics (ROC) curve of XGBoost and logistic regression respectively. The area under the ROC curve (AUROC) provides a vital performance measurement for classification models and represents the degree of separability of classes. AUROC of XGBoost model is 85% compared to 83% of logistic regression.

The advantage of using the XGBoost ML algorithm is that an importance score for each feature can be obtained. In general, the importance score measures the value of the feature in the construction of the model. Figure 5 shows the contributions of the eight features on the XGBoost ML model output ranked by the average absolute SHAP value. Glucose, body mass index, diabetes pedigree function, and age were the top four important features.

Discussion

In this study, I applied two machine learning algorithms to build a prediction model for the risk of T2DM among PIDD participants. It is found that the XGBoost ML model with eight features demonstrated good performance for predicting T2DM. This suggested that the prediction model derived in this study could be applied to predict individuals at high risk of T2DM, which could benefit the control of type 2 diabetes mellitus and hence the prevention of it. Table 3 presented the results of performance of XGBoost and LR machine learning models compared to other studies in the field on the same dataset.

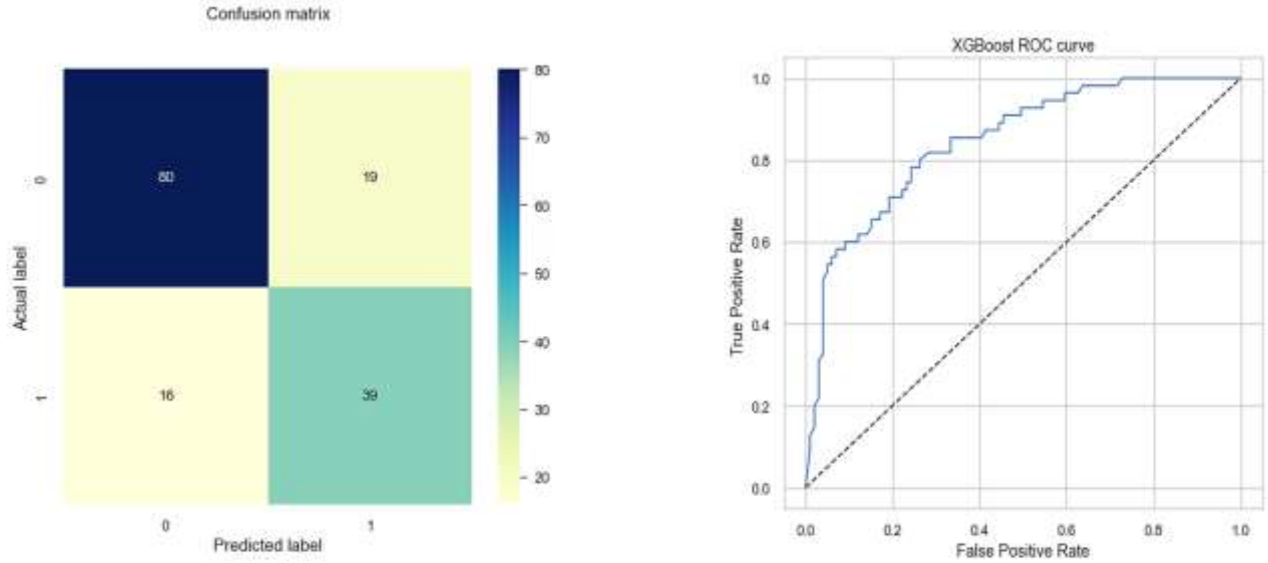


Figure 3. The confusion matrix and receiver operating characteristics (ROC) curve of the XGBoost machine learning model with AUROC of 85%.

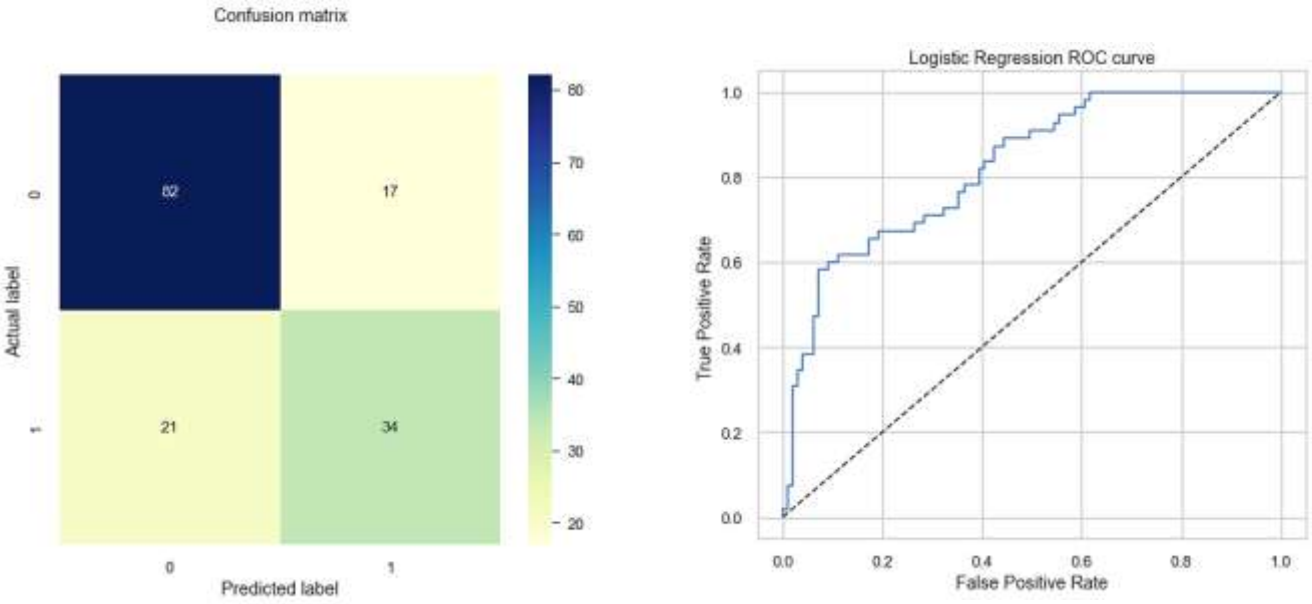
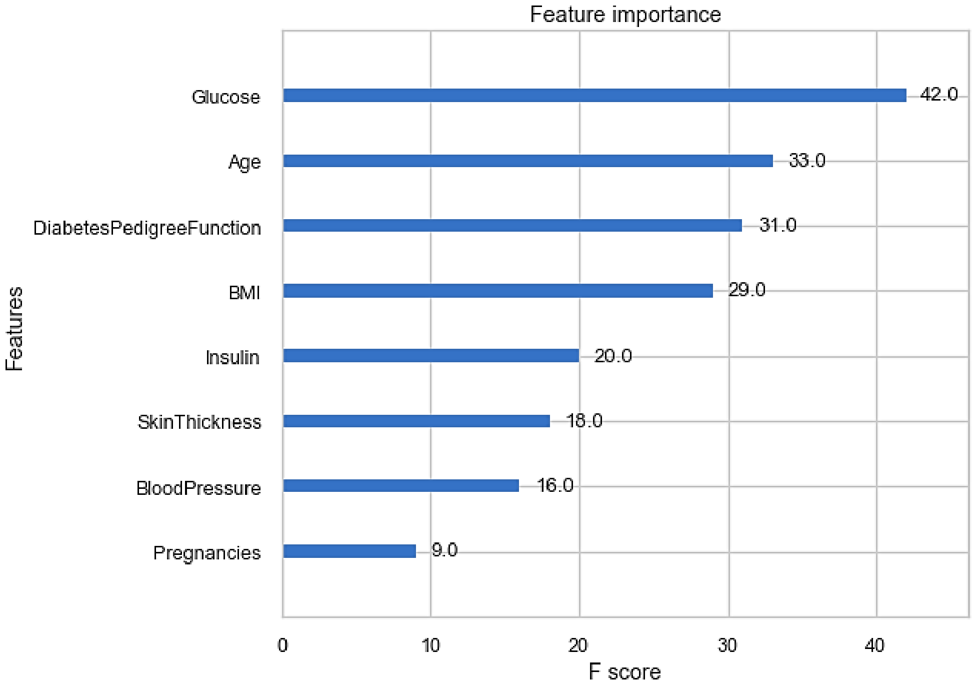
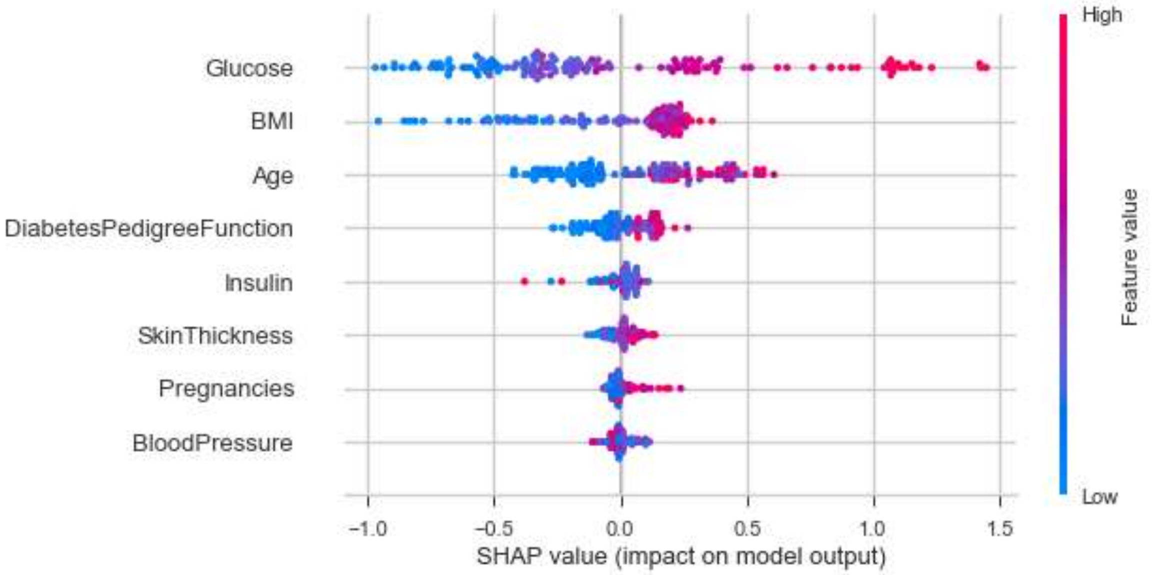


Figure 4. The confusion matrix and receiver operating characteristics (ROC) curve of the logistic regression machine learning model with AUROC of 82%.



A



B

Figure 5. The interpretation for the XGBoost model. (A) The feature importance ranking by the SHAP value; (B) SHAP summary plot of the XGBoost ML model.

Each dot represents an instance, with blue indicating a low feature value and red indicating a high feature value. The higher the value of a feature, the higher the risk of incident T2DM.

The prediction results confirmed that the XGBoost ML model performed best with the highest AUROC value of 85% on the test set in predicting the probability that an individual develops type 2 diabetes mellitus T2DM. It is a good example of success in the research of diabetes risk prediction. This finding was consistent with earlier studies [16,22], which identified the good prediction power of the XGBoost ML model, with AUROC values of 82% and 83% respectively.

Table 3. Performance metrics of XGBoost and logistic regression predictive models compared to other studies.

Reference	Prediction model	Accuracy (%)	Sensitivity (%)	AUROC (%)
[16]	Naïve Bayes	76	**76**	82
	Support Vector Machine	65	65	50
	Decision Tree	74	74	75
[23]	Random Forest	76	76	-
	J48	73	72	-
	Neural Network	76	78	-
[24]	Logistic Regression	76	63	76
	Random Forest	75	64	75
	XGBoost	75	65	75
This study	**XGBoost**	**77**	**71**	**85**
	Logistic Regression	75	62	82

There are several limitations in this study: The dataset used in this study is PIDD and it is believed that there are race/ethnic differences with type 2 diabetes mellitus [25], which might limit the extrapolation of the results. World Health Organization (WHO) has confirmed that a healthy diet, tobacco, and regular physical activity, are also important features to prevent or delay the onset of T2DM [2]. However, PIDD does not contain the aforementioned features of participants.

Conclusions and Recommendations:

The current study developed predictive models using XGBoost and logistic regression ML algorithms for the risk of incident T2DM. Glucose, age, diabetes pedigree function, and body mass index were the strongest medical predictors in the T2DM prediction model, which would benefit clinical practice in developing targeted T2DM prevention and control interventions. In the future, this work can be extended by taking into consideration additional predictor features such as education, healthy diet, smoking, and exercise to find how likely nondiabetic people can have diabetes in the next few years.

Conflicts of Interest

No conflict of interest to declare.

References:

[1] World Health Organization. Diabetes. https://www.who.int/health-topics/diabetes#tab=tab_1 (Accessed on 12 July 2022).

[2] World Health Organization. Diabetes. Fact sheets. https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed on 12 July 2022).

[3] International Diabetes Federation IDF Atlas. Diabetes around the world in 2021. https://diabetesatlas.org/ (Accessed on 12 July 2022).

[4] Ma RC, Tsoi KY, Tam WH, Wong CK. Developmental origins of type 2 diabetes: a perspective from China. European journal of clinical nutrition. 2017 Jul;71(7):870-80. https://doi.org/10.1038/ejcn.2017.48

[5] Huang Y, Vemer P, Zhu J, Postma MJ, Chen W. Economic burden in Chinese patients with diabetes mellitus using electronic insurance claims data. PLoS One. 2016 Aug 29;11(8):e0159297. http://dx.doi.org/10.1371/journal.pone.0159297

[6] Li Y, Wang DD, Ley SH, Vasanti M, Howard AG, He Y, Hu FB. Time trends of dietary and lifestyle factors and their potential impact on diabetes burden in China. Diabetes care. 2017 Dec 1;40(12):1685-94. https://doi.org/10.2337/dc17-0571

[7] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications. 2017;9(01):1. http://dx.doi.org/10.4236/jilsa.2017.91001

[8] Jain R, Chotani A, Anuradha G. Disease diagnosis using machine learning: A comparative study. InData Analytics in Biomedical Engineering and Healthcare 2021 Jan 1 (pp. 145-161). Academic Press. http://dx.doi.org/10.1016/B978-0-12-819314-3.00010-0

[9] McConnell KJ, Lindner S. Estimating treatment effects with machine learning. Health services research. 2019 Dec;54(6):1273-82. http://dx.doi.org/10.1111/1475-6773.13212

[10] McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, Gospodarowicz M, Helou J, Isfahanian N, Kong V, Lam T. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nature medicine. 2021 Jun;27(6):999-1005 http://dx.doi.org/10.1038/s41591-021-01359-w

[11] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17. http://dx.doi.org/10.1016/j.csbj.2014.11.005

[12] Diller GP, Kempny A, Babu-Narayan SV, Henrichs M, Brida M, Uebing A, Lammers AE, Baumgartner H, Li W, Wort SJ, Dimopoulos K. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. European heart journal. 2019 Apr 1;40(13):1069-77. http://dx.doi.org/10.1093/eurheartj/ehy915

[13] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019 Jan 1;165:292-9. http://dx.doi.org/10.1016/j.procs.2020.01.047

[14] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, *192*, 467-477. http://dx.doi.org/10.1016/j.procs.2021.08.048

[15] Chen W, Chen S, Zhang H, Wu T. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In2017 8th IEEE International conference on software engineering and service science (ICSESS) 2017 Nov 24 (pp. 386-390). IEEE. http://dx.doi.org/10.1109/ICSESS.2017.8342938

[16] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018 Jan 1;132:1578-85. http://dx.doi.org/10.1016/j.procs.2018.05.122

[17] Karthikeyani V, Begum IP. Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease. International journal on computer science and engineering. 2013 Mar 1;5(3):205.

[18] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[19] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794). http://dx.doi.org/10.1145/2939672.2939785

[20] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology. 2020 Jan 28;9(2):14-. https://doi.org/10.1167/tvst.9.2.14

[21] Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological). 1958 Jul;20(2):215-32. http://dx.doi.org/10.1111/j.2517-6161.1958.tb00292.x

[22] Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC medical informatics and decision making. 2019 Dec;19(1):1-5. http://dx.doi.org/10.1186/s12911-019-0918-5

[23] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 2018 Nov 6;9:515. http://dx.doi.org/10.3389/fgene.2018.00515

[24] Liu Q, Zhang M, He Y, Zhang L, Zou J, Yan Y, Guo Y. Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques. Journal of Personalized Medicine. 2022 Jun;12(6):905. http://dx.doi.org/10.3390/jpm12060905

[25] Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. Current diabetes reports. 2013 Dec;13(6):814-23. http://dx.doi.org/10.1007/s11892-013-0421-9