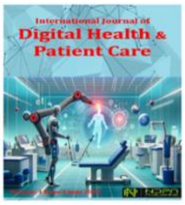
International Journal of
Digital Health & Patient Care
e-ISSN : 3023-851X
https://ndpapublishing.com/index.php/



Predicting Prediabetes Risk From Electronic Health Records Using Machine Learning

Nursena Bilgin[*1], **Ekber Gülpınar**[1] **and Yasin Karakuş**[1]

[1]*Kutahya Health Sciences University, Department of Computer Engineering, Türkiye*

Article Info
Received: 08 December 2025
Revised: 19 December 2025
Accepted: 23 December 2025
Published: 30 December 2025

Keywords
Prediabetes
Machine Learning
Health Data Analysis
Classification Algorithms
EHR

ABSTRACT

Prediabetes, also known as 'hidden sugar,' is a public health priority because of its risk of progressing to diabetes if left untreated. While studies regarding diabetes have been extensive, research focusing on the early detection of prediabetes is limited. For that reason, it is extremely important that efforts toward early identification be conducted. This study develops a prediabetes prediction model using machine learning algorithms from electronic health record data. Data were retrieved from the Korean National Health and Nutrition Examination Survey (KNHANES), examining associations between prediabetes and multiple factors among adults. The dataset consisted of 16 attributes and included clinical health information, socioeconomic indicators, physical activity, and dietary habits for 16,137 individuals. Non-contributory features were removed during preprocessing, while values normalization was performed with a Standard Scaler. To evaluate model performance, the dataset was split into an 80% training set and a 20% test set. Four different machine learning methods were applied: SVM, KNN, Logistic Regression, and Random Forest. After training, their performance was tested on the test set. Accuracy, precision, recall, F1-score, and ROC-AUC were measured. Among all models, the Random Forest algorithm demonstrated 68% accuracy and 61% precision, while SVM demonstrated 75% recall. Logistic Regression showed a performance of 64% for the F1-score with 75% ROC-AUC. These are very promising results for the detection of prediabetes. In the future, prediction will be improved by using larger datasets and advanced feature selection, including deep learning techniques.

1. INTRODUCTION

Prediabetes, which is often seen in people who do not have the symptoms necessary for a diagnosis of Type 2 Diabetes, commonly known as diabetes, but whose blood sugar levels are abnormally high, is a dangerous condition for human health. Although it is not defined as a disease from a medical point of view, it is a sign of serious health problems that may develop in the future. Prediabetes is a significant metabolic disorder that affects a large segment of the population but often goes unnoticed because it does not present obvious symptoms.

Given that symptoms are typically unnoticeable, identifying prediabetes requires an assessment of specific risk factors. High-risk populations include individuals who are overweight or obese, suffer from hypertension, exhibit reduced high-density lipoprotein (HDL), or show elevated triglyceride concentrations. Additional indicators involve a family history of diabetes, history of gestational diabetes, delivery of infants exceeding 4 kg, and polycystic ovary syndrome (PCOS).

Should prediabetes evolve into Type 2 Diabetes, it may result in severe medical complications, including cardiovascular pathology, renal dysfunction, neurological impairments, and stroke. Consequently, detecting the condition prior to the onset of full-blown diabetes is critical. Fortunately, for those identified with prediabetes, it is feasible to halt or reverse this trajectory through dietary management, consistent physical activity, and regular medical oversight.

According to the International Diabetes Federation's latest report [1], Approximately 589 million adults aged 20–79 worldwide are living with diabetes. This number is expected to reach 853 million by 2050. Prediabetes can be associated with the prevalence of Impaired Fasting Glucose (IFG) and Impaired Glucose Tolerance (IGT) observed in the population. As of 2024, it is estimated that approximately 635 million adults have IGT and 488

***Corresponding author**

*e-mail: nursena.bilgin@ogr.ksbu.edu.tr
ORCID ID: 0009-0001-7790-5524

Research Article/ DOI: 10.5281/zenodo.18055607

How to cite this article

Bilgin, N., Gülpınar, E., & Karakuş, Y. (2025). Predicting Prediabetes Risk From Electronic Health Records Using Machine Learning. Int. J. Digital Health & Patient Care, 2(2), 103-110.

million have IFG. By 2050, this number is projected to be approximately 847 million for IGT and 648 million for IFG.

2.1. Literature Review

Studies in the literature on the diagnosis of diabetes and prediabetes have demonstrated significant advances using different clinical data and machine learning methods.

Zhang et al. [2], aimed to predict prediabetes and diabetes (AGM) in individuals with normal fasting blood glucose (normoglycaemic) using machine learning methods. They utilised health screening data from Shandong First Medical University in China between 2019 and 2023. Twenty-one features selected using the LASSO method (age, BMI, fasting glucose, haemoglobin, erythrocyte count, and triglyceride-glucose index, etc.) were used to train the model. Seven different machine learning algorithms were used in the study: logistic regression, random forest, support vector machine, XGBoost, multilayer perceptron, LightGBM, and CatBoost. According to the test results, the CatBoost model showed the highest accuracy (ACC = 0.731 / 0.718) and AUC value (auROC = 0.806 / 0.794) in both internal and external validation sets.

De Silva et al. [3], conducted a study using data from the 2013–2014 US National Health and Nutrition Examination Survey (NHANES) to improve the prediction of prediabetes. The models used included logistic regression, Artificial Neural Network (ANN), Random Forests (RF), and Gradient Boosting (GB). Forty-six variables selected through feature selection were used, and class imbalance was addressed using SMOTE, ROSE, oversampling, and undersampling methods. As a result of the study, the Random Forest (RF) model (with the minority class oversampled) showed the highest performance, and the AUC value of this model was reported as 71.59%.

Severeyn et al. [4], used five-point OGTT data collected from 188 individuals in Venezuela to predict prediabetes and diabetes using Support Vector Machines (SVM). Models based on the AUCG variable demonstrated high success, with 94.1% accuracy and a 95.2% F1 score. The findings suggest that OGTT-based SVM models may be effective in early diagnosis.

Bashar et. al. [5], used machine learning to predict prediabetes using NHANES 2011–2014 data. Sixteen variables, including age, gender, weight, education, race, lifestyle, and dietary habits, were included in the model. Decision Tree, SVM, Gradient Boosting, Random Forest, Logistic Regression, and Artificial Neural Network algorithms were compared; a 65% training, 25% validation, and 10% test split was used. The Random Forest model demonstrated the best performance with 89% accuracy, 0.115 ASE, 0.593 ROC area, and 0.1298 KS score.

Kanbour et al. [6], investigated the prediction of microvascular complications (retinopathy, kidney disease, neuropathy) in type 2 diabetes patients using machine learning methods. Data from 74 longitudinal studies published in PubMed between 1990 and 2023 were analysed. Fifteen variables were used, including age, gender, BMI, A1C, blood pressure, and kidney function indicators; Linear and Logistic Regression, Random Forest, XGBoost, LightGBM, SVM, and other ML algorithms were evaluated. Diabetic kidney disease models showed the highest performance (internal validation c-statistic 0.81, external validation 0.74); retinopathy and neuropathy models remained less accurate. XGBoost, Random Forest, and Logistic Regression were the most successful algorithms.

Islam and Khanam [7], performed diabetes classification using machine learning algorithms on the Pima Indian Diabetes dataset. SVM, Naive Bayes, Logistic Regression, Decision Trees, KNN, and Random Forest algorithms were tested; Gaussian Naive Bayes yielded the best result with 79.87% accuracy. SVM achieved 78.5% accuracy with a polynomial kernel and 77.9% accuracy with a linear kernel. The findings suggest that machine learning-based models may be effective in the early diagnosis of diabetes.

Mujumdar and Vaidehi [8] performed diabetes prediction using a special dataset containing 800 samples and 10 features. The data underwent preprocessing, missing values were imputed, and pre-labelling was performed using K-means clustering. Logistic Regression was the most successful base model with 96% accuracy, while the AdaBoost classifier using the pipeline method yielded the best result with 98.8% accuracy. The findings demonstrate that machine learning algorithms can achieve high success in diabetes prediction with appropriate data preparation and modelling steps.

2.1. Observations and Contributions

While the literature abounds with diabetes prediction models, there is a notable scarcity of research dedicated to the early prognosis of prediabetes. Furthermore, the comparative efficacy of specific machine learning algorithms in detecting this condition within the Korean demographic remains largely unexplored. To address this gap, this study leverages the Korean National Health and Nutrition Examination Survey (KNHANES) dataset to differentiate between normal and prediabetic subjects. We implement Support Vector Machines

(SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest classifiers, selected specifically for their robustness in handling the non-linear complexities inherent in health data.

The methodology begins with a rigorous preprocessing pipeline applied to a cohort of 16,137 individuals, focusing on feature optimization and data standardization. We then evaluate the diagnostic capability of these models through comprehensive performance metrics. The remainder of this paper is organized as follows: Section 2 details the dataset characteristics and the methodological framework; Section 3 reports the comparative results including accuracy, sensitivity, and ROC-AUC analyses; and Section 4 discusses the clinical implications of these findings alongside directions for future research.

2. MATERIALS AND METHODS

The work flow diagram for the study is shown in Figure 1.

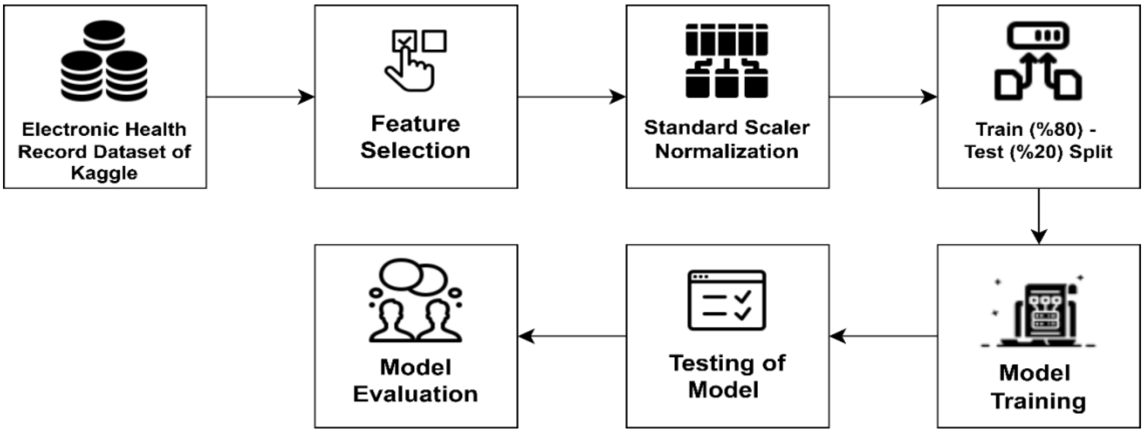

Electronic Health Record Dataset of Kaggle
Feature Selection
Standard Scaler Normalization
Train (%80) - Test (%20) Split
Model Training
Testing of Model
Model Evaluation

Figure 1. Work flow diagram of the study

2.1. Dataset

The data set used in this study is a subset of the Korea National Health and Nutrition Examination Survey (KNHANES). It was obtained from the Kaggle website [9]. The data set investigates the interaction between prediabetes and environmental factors in individuals aged 19 and older. The dataset consists of 16,137 records with 16 features. Detailed information about the dataset is provided in Table 1. The correlation heatmap for the dataset is shown in Figure 2.

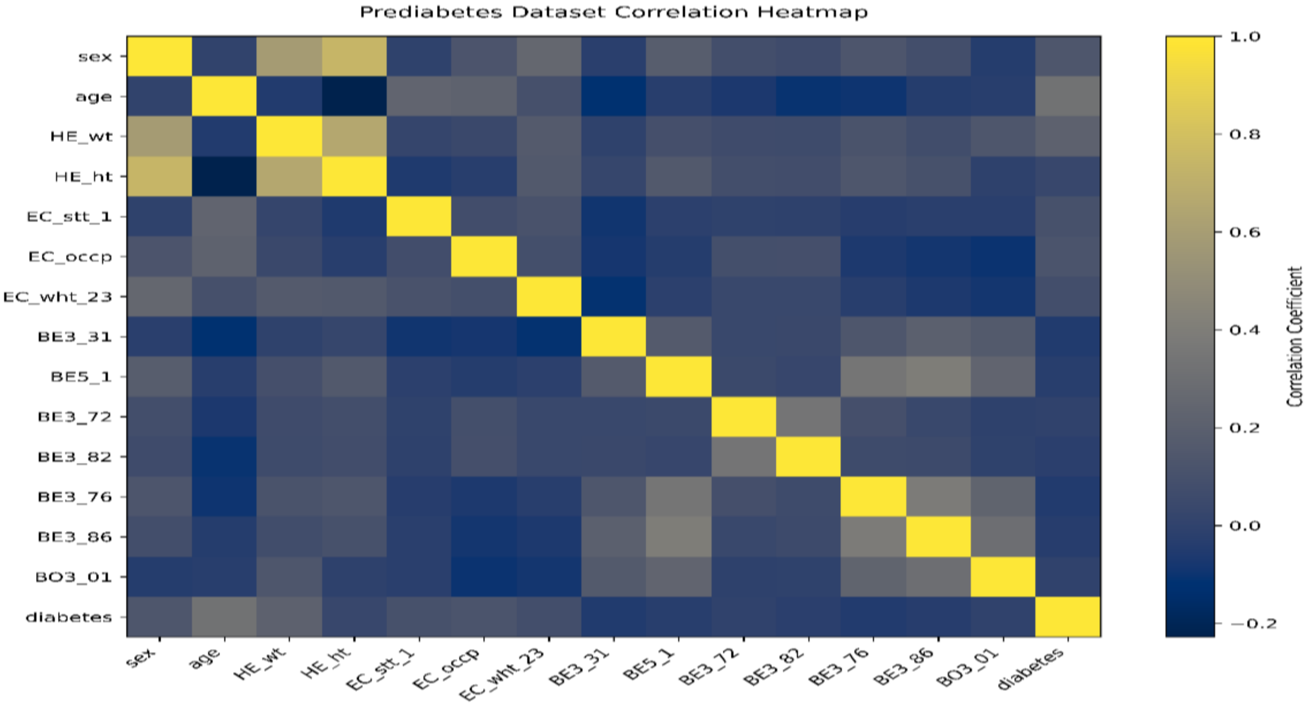


Figure 2. The correlation heatmap for the dataset

Table 1. Detailed information about the dataset

Feature Name	Data Type	Explanation	Value Range
Sex	int64	gender	0: Female, 1: Male
Age	int64	age	19 – 60
HE_wt	float64	Body Weight (kg)	30.6 – 138.1 kg
HE_ht	float64	Height (cm)	128.8 – 195.0 cm
EC1_1	int64	Employment Status	1 (Fixed Value)
EC_stt_1	int64	Type of Employment	1: Elementary school – 4: University
EC_occp	int64	Occupational Category	1 – 10
EC_wht_23	int64	Weekly Working Hours	0 – 145
BE3_31	int64	Number of Days Walking (Weekly)	0: Never – 7: Every day
BE5_1	int64	Number of Days Strength Training (Weekly)	0: Never – 7: Every day
BE3_72	int64	Number of Days High-Intensity Work Activity (Weekly)	0: Never – 7: Every day
BE3_82	int64	Number of Days of Moderate-Intensity Work Activity (Weekly)	0: Never – 7: Every day
BE3_76	int64	High-Intensity Leisure Activity (Weekly)	0: Never – 7: Every day
BE3_86	int64	Number of Days of Moderate-Intensity Leisure Activity (Weekly)	0: Never – 7: Every day
BO3_01	int64	Exercise for Weight Control	0: No, 1: Yes
Diabetes	int64	Prediabetes Status	0: No, 1: Yes

2.2. Data preprocessing

The data preprocessing stage is the stage where data is processed and optimized to obtain results. In this study, the data preprocessing stage consists of two different steps: feature selection and standard scaling. Some columns are present in the dataset even though they do not carry information relevant to the prediction. These unnecessary features negatively affect the model's learning, so they are removed. For this reason, the variances of the features were first calculated. Then, features with a variance of 0, i.e., those with a constant value that do not contribute to learning and increase the computational load, were removed from the dataset. Therefore, the "EC1_1" feature was removed from our dataset to enable the model to produce better results. Features with a variance below 0.3 are shown in Figure 3.

Standardization is the process of rescaling data so that its mean is 0 and its standard deviation is 1, in order to improve the calculation process and ensure that the model evaluates the variables in a balanced manner. Standard Scaler is a scaling method that transforms each numerical feature in the dataset to have a mean of 0 and a standard deviation of 1. This brings the features to the same scale and enables algorithms to learn more consistently. The distributions of the features before and after the standard scaler are shown in Figure 4.

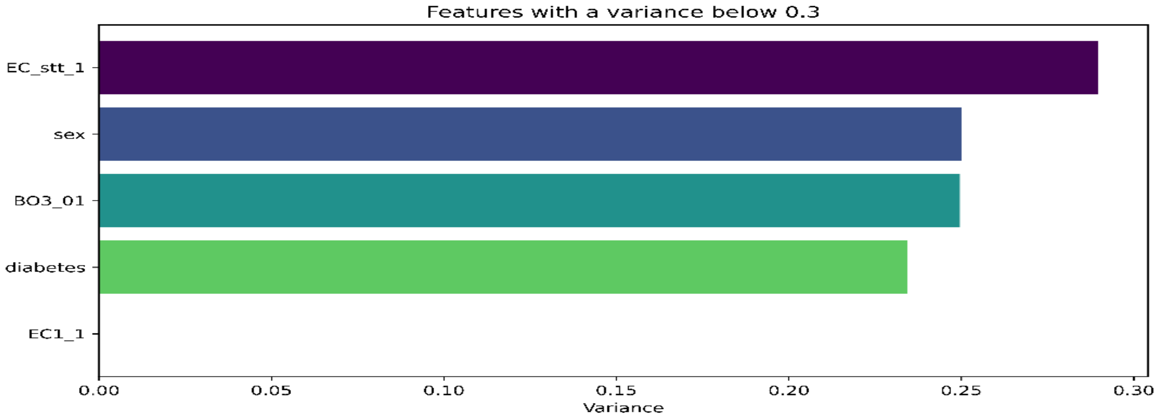


Figure 3. Features with a variance below 0.3.

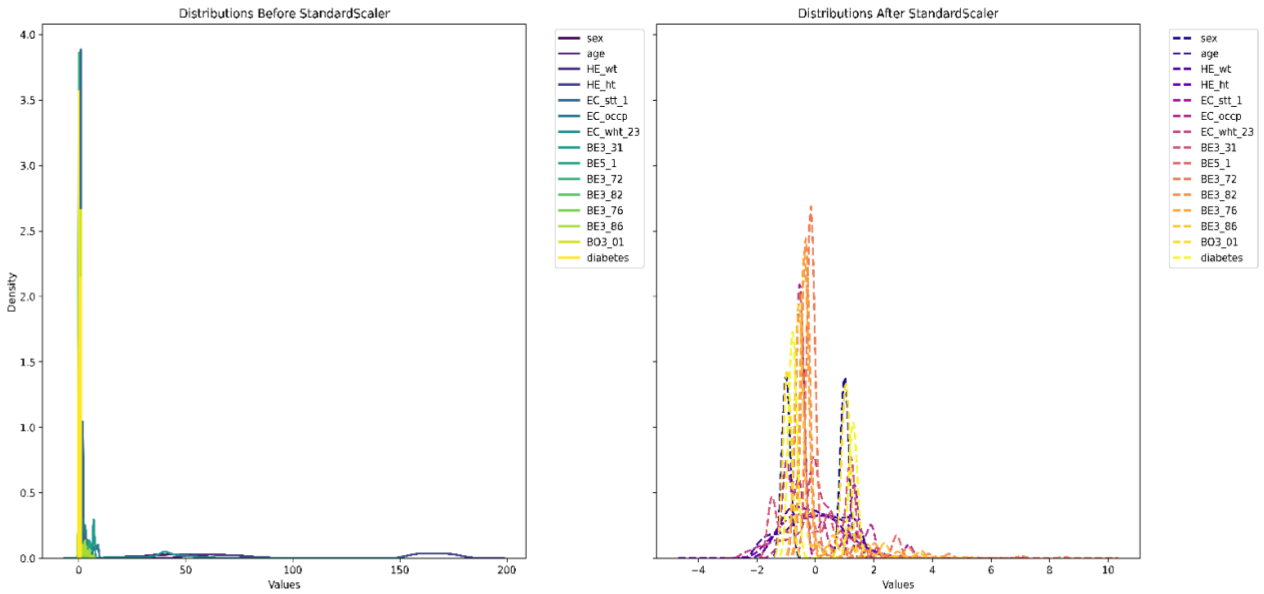


Figure 4. Distributions of features before and after the standard scaler

2.3. Prediabetes prediction model

Following feature selection and standard scaling steps, the dataset was split into two parts: 80% for training and 20% for testing. Four different machine learning methods were used to predict the laptop price range. These are: random forest, support vector machine, k-nearest neighbor and logistic regression.

The performance of the models was evaluated on the test dataset. Each model was tested independently of the other models under the same conditions; therefore, the result of one model does not affect the performance of the other models. The parameters used when defining the models are given in Table 2. Parameters not specifically mentioned were used with their default values.

Table 2. Detailed information about the dataset

Model	Parameter	Value
Random Forest	Class_Weight	Balanced
	N_Estimators	100
	Random_State	42
SVM	Class_Weight	Height (Cm)
	Probability	True
	Kernel	Rbf
	Random_State	42
KNN	N_Neighbors	5
Logistic Regression	Class_Weight	Balanced
	Random_State	42

2.3.1. *Random Forest*

Random Forest is a randomised multi-tree (ensemble) classification method developed to overcome the problem of traditional decision trees being unable to increase complexity without incurring generalisation loss. It achieves higher accuracy by combining the classifications of numerous decision trees trained independently on randomly selected subsets of the feature space [10].

2.3.2. *Support Vector Machine*

A Support Vector Machine is a two-class learning method that transforms input vectors into a high-dimensional feature space using a non-linear function and establishes a linear decision surface in this space; it achieves high generalisation capability thanks to the special geometric properties of the decision surface; initially defined only for perfectly separable data, it was later extended to non-separable cases [11].

2.3.3. *K-Nearest Neighbor*

The k-Nearest Neighbour (KNN) algorithm, which belongs to the supervised learning class, relies on labelled training data to solve classification and regression problems. The algorithm performs the computation at query time rather than going through a training phase, thus requiring the storage of the entire dataset; this makes the method highly dependent on memory usage. Classification operations are based on majority voting among the closest neighbours to the target data point. In other words, the algorithm's output is the class that is most frequently observed among the nearest neighbours (the mode).

2.3.4. *Logistic Regression*

Logistic regression is a parametric statistical method that models the relationship between independent variables and the probability of the dependent variable occurring when the dependent variable is binary. In this model, the linear combination of independent variables is converted to probability via the logit transformation (log-odds), thereby limiting the predicted probabilities to between 0 and 1. The model provides probability estimates, particularly in classification problems, and allows for the direct interpretation of the effects of the independent variables [12].

3. RESULTS

The work was carried out using Python 3.12.7 in the JupyterLab 4.2.5 IDE. It was successfully run on a computer with an Intel Core i5-12600HX 2.50 GHz CPU and 24 GB RAM.

In this study, accuracy, precision, recall, F1-score, and ROC-AUC score criteria were used to evaluate the model's success. TP, used in the following formulas, is the number of individuals with prediabetes who were correctly predicted. TN is the number of individuals without prediabetes (healthy) who were correctly predicted. FP is the number of individuals who are actually healthy but were incorrectly predicted to have prediabetes. FN is the number of individuals who actually have prediabetes but were incorrectly predicted to be healthy. The confusion matrix obtained from the test results of the trained models is shown in Figure 5. Accuracy is calculated according to Equation 1, Precision according to Equation 2, Recall according to Equation 3, and F1 score according to Equation 4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

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

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

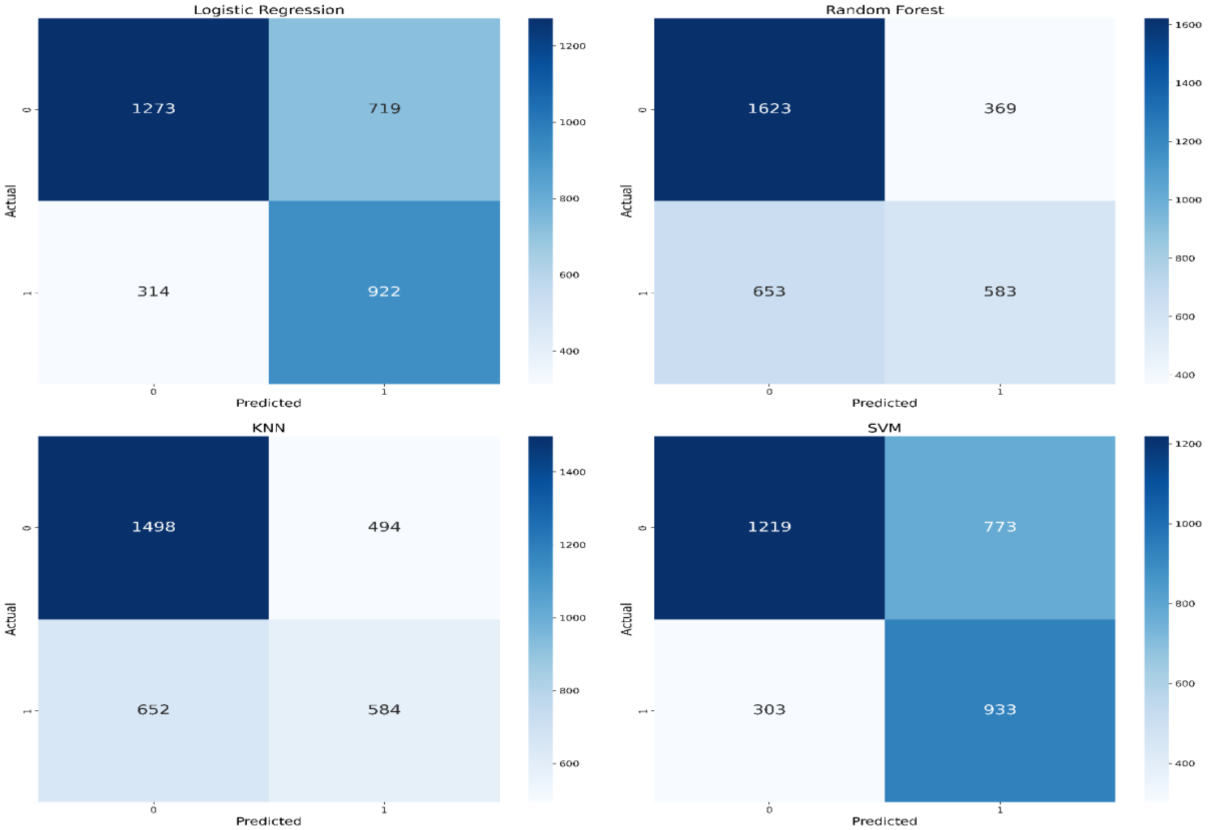


Figure 5. Confusion matrix obtained from the test results of trained models.

The evaluation results show that the models achieved similar success scores. As a result of the measurements, Random Forest stands out with 68% accuracy and 61% precision, SVM with 75% sensitivity, and Logistic Regression with 64% F1-score and 75% ROC-AUC score. In terms of overall performance, the Logistic Regression model provided the most balanced result. The evaluation results are shown in Table 3.

Table 3. Evaluation results

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
Logistic Regression	0.679988	0.561853	0.745955	**0.640945**	**0.755987**
Random Forest	**0.683395**	**0.612395**	0.471683	0.532907	0.733399
KNN	0.644981	0.541744	0.472492	0.504754	0.667100
SVM	0.666667	0.546893	**0.754854**	0.634262	0.744877

4. Conclusion

Although diabetes prediction is an established field, research focused on the early detection of prediabetes remains relatively scarce. The gap in this regard is filled by this study, as for the first time, a machine learning framework is used on the KNHANES dataset targeting the asymptomatic phase of the disease. Following a structured preprocessing pipeline that emphasized feature optimization and normalization, we set up a sound basis for model training.

Our comparative analysis among the four classifiers-SVM, KNN, Logistic Regression, and Random Forest-showed different performance profiles. Remarkably, though Logistic Regression provided the most balanced metrics overall, Random Forest achieved both the highest precision and accuracy, while SVM had the highest sensitivity. These results highlight the clinical feasibility of electronic health records for automated screening, representing a great opportunity for early intervention at a point when disease progression is still limited.

However, some limitations regarding sample size and the cross-sectional nature of the dataset do exist, and findings should be replicated in larger and longitudinal cohorts to enhance generalisability. A very promising avenue for further investigation lies in developing hybrid models that make use of data from multiple modalities, such as the combination of routine clinical records with new biomarkers or with speech features including vocal characteristics. Expansion of the feature space with variables known to relate to metabolic risks stays among the top priorities for enhancing the predictive power of such systems.

Conflict of Interest

No conflict of interest is declared by tehe authors. In addition, no financial support was received.

Author Contributions

Study Design, NB, EG, YK; Data Collection, NB, EG, YK; Statistical Analysis, NB, EG; Data Interpretation, NB, EG, YK; Manuscript Preparation, NB, EG, YK; Literature Search, NB, EG. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. International Diabetes Federation. (**2025**). IDF Diabetes Atlas (11th ed.). Brussels, Belgium: International Diabetes Federation. http://www.diabetesatlas.org
2. Zhang, X., Yao, W., Wang, D., Hu, W., Zhang, G., & Zhang, Y. (**2024**). Development and validation of machine learning models for identifying prediabetes and diabetes in normoglycemia. *Diabetes Metab Res Rev, 40(8), e70003.* [CrossRef] [PubMed]
3. De Silva, K., Jönsson, D., & Demmer, R. T. (**2019**). A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *Journal of the American Medical Informatics Association*, *27*(3), 396–406. [CrossRef] [PubMed]
4. Severeyn, E., Velásquez, J., La Cruz, A., & Huerta, M. (**2024**). Leveraging support vector machines for enhanced diagnosis of diabetes and prediabetes. *2024 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE. [CrossRef]
5. Bashar, A. K. M. R., Goudarzi, M., & Tsokos, C. P. (**2024**). A machine learning classification model for detecting prediabetes. *Journal of Data Analysis and Information Processing*, *12*(03), 462–478. [CrossRef]
6. Kanbour, S., Harris, C., Lalani, B., Wolf, R. M., Fitipaldi, H., Gomez, M. F., & Mathioudakis, N. (**2024**). Machine learning models for prediction of diabetic microvascular complications. *Journal of Diabetes Science and Technology*, *18*(2), 273–286. [CrossRef] [PubMed]
7. Islam, N. U., & Khanam, R. (**2021**). *Classification of diabetes using machine learning. In 2021* International Conference on Computational Performance Evaluation (ComPE) (pp. 185-189).

8. Mujumdar, A., & Vaidehi, V. (**2019**). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. [CrossRef]
9. Prediabetes and Health Dataset. (**2025,** April 15). Retrieved from https://www.kaggle.com/datasets/jesusdeleon19/prediabetes-and-health-dataset
10. Ho, T. K. (**1995**). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. [CrossRef]
11. Cortes, C., & Vapnik, V. (**1995**). Support-vector networks. *Machine Learning, 20(3)*, 273–297. [CrossRef]
12. Cox, D. R. (**1958**). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B (Statistical Methodology), 20(2),* 215–232. [CrossRef]