International Journal of Scientific & Engineering Research, Volume 15, Issue 4, April-2024
ISSN 2229-5518

Diabetes Prediction Using Machine Learning Algorithm

Sumaia Rahman, Suraiya Jabin, Most. Ruckaya Haque, Ayasha Akter Lija, Habiba Ferdausi Ritu

Abstract— Diabetes is a major public health concern, demanding accurate early warning systems for efficient care. This study diagnoses diabetes using the PIMA Indian database, which includes medical predictor variables as well as lifestyle factors. Five ML algorithms—support vector machine, random forest, logistic regression, decision tree, and K-nearest neighbors—are rigorously evaluated for predictive usefulness using criteria such as accuracy, sensitivity, and specificity. The results demonstrate the Random Forest (RF) algorithm's extraordinary performance, with a staggering 98% accuracy in diabetes prediction. This study highlights the transformative potential of machine learning in illness management by providing a data-driven method for identifying at-risk patients and implementing preventative actions. The findings add significant insights to the field of diabetes prediction, emphasizing the importance of machine learning in increasing early identification and proactive healthcare measures.

Index Terms— SVM, Machine learning, Random Forest Classifier, Data Mining, Diabetes

1 Introduction

Diabetes (DM), a prevalent and escalating chronic disease, hinders the body's ability to produce and respond to insulin, resulting in elevated blood glucose levels Associated conditions such as Coronary Kidney Disease, Hypertension, Coronary Artery Disease, COPD, and Hypothyroidism often manifest without clear symptoms, emphasizing the critical need for early detection and appropriate treatment [1] Further classifying into insulin-dependent glucose dysregulation disorder (T1D) and adult-onset glucose dysregulation disorder (T2D), T1D commonly affects younger individuals, presenting clinical indications like increased thirst and frequent urination, requiring insulin treatment. T2D, prevalent in middle-aged and older individuals, associates with overweight, abnormally high blood pressure, high cholesterol, gastritis, and other disorders.[2] According to WHO, nearly four hundred twenty-two million are impacted by glucose dysregulation disorder, with projections anticipating a rise to 490 million by 2030. As a major cause of mortality worldwide, the initial identification and monitoring of disorders such as diabetes are critical to preventing deaths. [3] In the contemporary landscape, ML approaches play a crucial part in the identification and prediction of life-threatening illnesses, including diabetic. This study focuses on forecasting diabetes using disease-related factors from Pima Indian Diabetic Data. We intend to create robust predictive models by employing a variety of Machine Learning (ML) categorization and combined methods. The study looks at the performance of Random Forest, K-NN, Decision Tree, SVM, and Logistic Regression algorithms in forecasting diabetes, which contributes to ongoing attempts to improve healthcare outcomes through sophisticated predictive analysis.

2 LITERATURE REVIEW

Glucose dysregulation is a medical disorder that results in increased amounts of glucose in the blood due to a lack of insulin, which interferes with sugar metabolism.[2] Individuals with diabetes face challenges in converting ingested carbohydrates into glucose efficiently, the primary energy source for daily activities. This results in a gradual buildup of sugar in the bloodstream, as glucose fails to reach all body cells.[4] A recent study investigated a Diabetic database with eight hundred samples and ten features. To improve the model's accuracy, researchers corrected inconsistencies and missing values in important variables such as age, BP, insulin level, body mass index, and skin thickness. They used imputation and subsequently data scaling for this motive. The application of K-means clustering classified patients into diabetic or non-diabetic categories, identifying highly correlated attributes (Glucose and Age) before clustering.[5] Various classification algorithms, including SVM Classifier, RF, DT Classifier, were evaluated. LR emerged with the highest accuracy at 96%. Additionally, a pipeline implementation revealed the AdaBoost classifier as the top-performing model, achieving an accuracy of 98.8%. The study included a comparative analysis of machine learning algorithm accuracies, considering two distinct datasets.[5] In another analysis involving the Pima Indians diabetes dataset, three methods were tested, with Artificial Neural Networks (ANN) demonstrating superior performance. Association rule mining uncovered a notable correlation between BMI and glucose levels with diabetes. While the study focused on structured data, future research aims to explore unstructured data and extend predictive methods to various medical conditions such as cancer, psoriasis, and Parkinson's disease. Planned considerations involve incorporating additional attributes like physical inactivity, family history of diabetes, and smoking habits to enhance diabetes diagnosis.[6][7]

3 METHODOLOGY

In this part, we delve into various classifiers employed in ML for predicting diabetes, accompanied by our proposed methodology geared towards enhancing accuracy. Five distinct methods were employed in this study, each outlined

below, with the resulting accuracy metrics of the machine learning modelsThe overall goal of this research is to investigate models capable of predicting diabetes with increased accuracy. To attain this purpose, experiments were carried out with various classification and ensemble techniques. ML, a critical subset of AI, commonly referred to as artificial intelligence, employs algorithms designed to learn and adapt from data without explicit programming. Its transformational potential stems from its capacity to recognize patterns, extract insights, and generate predictions based on processed data. Algorithms that use machine learning analyze previous and current clinical data to reveal detailed patterns and correlations related to the beginning of diabetes. ML has a significant influence by allowing for the construction of prediction models that not only improve early detection but also contribute to personalized interventions. ML is at the forefront of technological advancement because it constantly refines its understanding through the incorporation of fresh data, offering a future in which healthcare decision-making is increasingly information-driven, precise, and proactive. The initial objective of this study is to develop and apply Diabetes Prediction utilising a variety of ML methodologies. Furthermore, an Output Evaluation of these approaches is performed to determine the optimum classifier with the maximum accuracy. The next sections provide a brief overview of the phases involved, with Figure 1 representing the diagram for the suggested diabetes prediction methodology.

3.1 Dataset Description

The glucose dysregulation disorder database was obtained from Kaggle datasets. The dataset used in this work has been collected by Dr. Shruti Garg and Neha Prerna Tigga from the Department of CSE at BIT Mesra, Ranchi, for non-profit experimental reasons. The dataset contains nine hundred and fifty-two instances, 17 individual predictive factors, and a binary target or dependent variable connected to diabetes. The collection of a large dataset is a primary emphasis in data mining and machine learning research, and getting data from credible sources with a complete understanding is crucial. In this investigation, Kaggle was used to find robust and reliable data sources. [9][10]

TABLE 1
DATASET DESCRIPTION

Number	Attributes	Description
1.	Gender	Male(61%), Female(39%)
2.	Age	Age (in years)
3.	Family_Diabetes	True 454 (48%) False 498 (52%)
4.	highBP	True 228 (24%) False 724 (76%)
5.	Physically Active	more than 30min (29%) less than 30min (35%) Other (344) 36%
6.	BMI	The index of body mass
7.	Smoking	True 108 (11%) False 844 (89%)
8.	Alcohol	True 192 (20%) False 760 (80%)
9.	Sleep	Sleeping time in hr
10.	SoundSleep	Sound Sleep time in hr
11.	RegularMedicine	No 65% Yes 35% Other (1) 0%
12.	JunkFood	Occasionally 71% Often 19% Other (96) 10%
13.	Stress	Sometimes 59% very often 17% Other (224) 24%
14.	BPLevel	Normal 74% High 22% Other (34) 4%
15.	Pregancies	Number of times pregnant
16.	Pdiabetes	0 98% Yes 1% Other (2) 0%
17.	UriationFreq	Not much 70% Quite often 30%
18.	Diabetic	yes 266 (28%) no 684 (72%) [null] (10%) ("1" means yes and "0" means no)

The diabetes data set consists of almost 1000 data points, with 18 features each.

•*Sumaia Rahman is currently working as a lecturer in the department of Computer Science & Engineering in Varendra University, Bangladesh, PH-+8801795381332.E-mail: ononnaontora@gmail.com*
•*Suraiya Jabin, Student of Computer Science & Engineering in University of Rajshahi, Bangladesh, PH- +880 1710251866. E-mail: surai-yajabin765@gmail.com*
•*Most. Ruckaya Haque, Student of Computer Science & Engineering in Varendra University, Bangladesh, PH- +880 1600-356769. E-mail ruckaya.saba@gmail.com*
•*Ayasha Akter Lija, Student of Computer Science & Engineering in Varendra University, Bangladesh, PH- +880 1303869390. E-mail: ayashaakter-li-ja55@gmail.com*
•*Habiba Ferdausi Ritu, Student of Computer Science & Engineering in Varendra University, Bangladesh, PH- +8801719392461. E-mail: habibaferdausiri-tu@gmail.com*

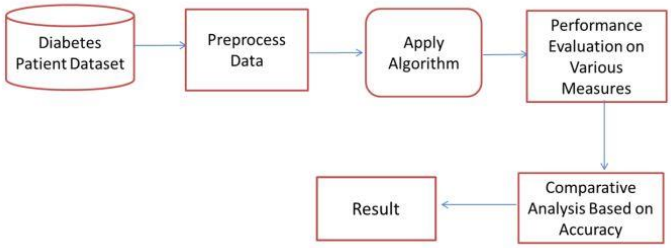

Fig. 1. Methodological workflow.

4 Method used for Identification

After preprocessing the data, applied a variety of ML techniques for classification. K-Nearest-Neighbors (K-NN), Random Forest(RF), Logistic Regression(LR), Decision Tree (DT), and Support Vector Machine (SVM) algorithms are utilized in this study, considering features such as A person's gender, age, family diabetes, high blood pressure, physical activity, body mass index, alcohol, tobacco use, sleep, quality of sleep, routine medicine, , stress Unhealthy foods, blood pressure levels, pregnancies, Urine Frequency, and Diabetes.It's worth noting that among these features, the RegularMedicine attribute, alongside Age, emerges as the second most informative feature overall. Notably, leveraging these features, the Random Forest classification algorithm achieves the highest accuracy, reaching 98%.

4.1 SVM or Support Vector Machine

The primary goal of this classifier is to create a hyperplane that maximizes the separation between classes by fine-tuning the distance between data points and the hyperplane[12]. Different kernels are utilized to determine this hyperplane. In this experimentation, I tested four kernels: Linear, Poly, Rbf, and Sigmoid.

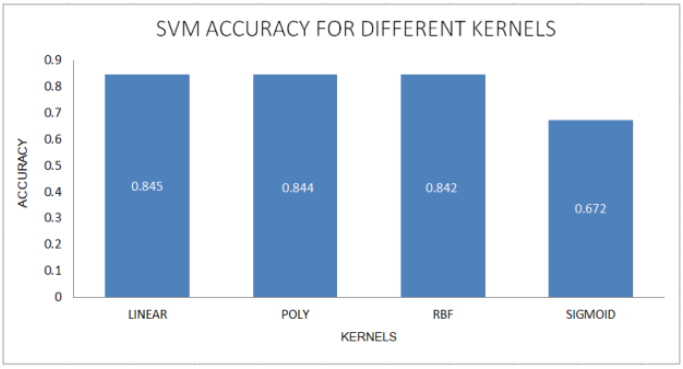

Fig. 2. SVM accuracy of different kernels

Upon analysis of the results, it's evident from the plot that the linear kernel outperformed the others for this dataset, achieving a notable score of 85%.

4.2 K-Nearest Neighbor

The k-NN algorithm is often considered one of the most straightforward machine learning approaches. Its model construction involves merely storing the training dataset. When predicting for a new data point, the algorithm identifies the closest data points within the training dataset, known as its "nearest neighbors."

In our approach, we employ Cross Validation to determine the optimal value for k. For this purpose, we utilize scikit-learn's cross_val_score function. This involves passing an instance of the kNN model, along with our dataset, and specifying the number of splits to generate.

4.3 Decision Tree

A decision tree serves as a fundamental classification method and falls under the category of supervised learning. This method is employed when the response variable is categorical. Feature importance is a metric that assesses the significance of each feature in the decision-making process of a tree. Represented as a number between 0 and 1 for each feature, where 0 implies "not used at all" and 1 denotes "perfectly predicts the target," the feature "Regular Medicine" emerges as the most crucial. The accuracy achieved through this algorithm is 97%.

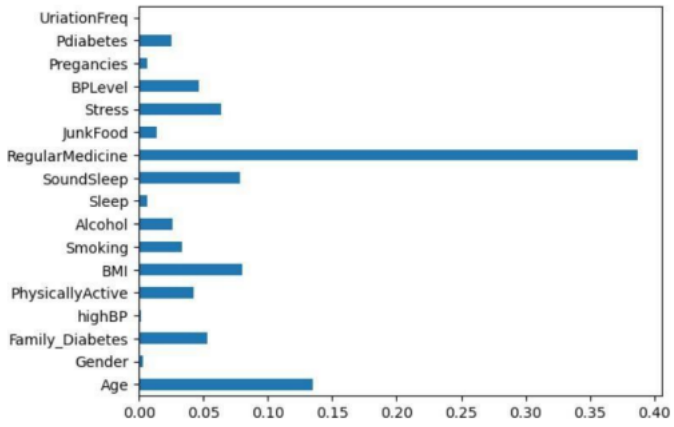

Fig. 3. Feature Importance in Decision Trees.

4.4 Random Forest

The random forest [13], a form of ensemble learning, is applied for both classification and regression tasks. Its accuracy surpasses that of other models. In parallel to a single decision tree, the random forest assigns significant importance to the "Regular Medicine" feature. Additionally, it identifies "Age" as the second most informative feature overall. The accuracy achieved through this algorithm stands at an impressive 98%.

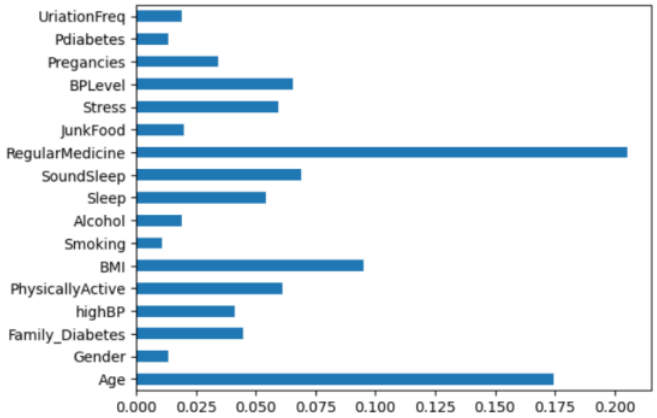

Fig. 4. Feature Importance in Random Forest.

4.5 Logistic Regression

Logistic regression [14], like decision trees and random forests, is a supervised learning classification algorithm. It's employed to estimate the probability of a binary response based on one or more predictors, which can be either continuous or discrete variables. This method is particularly useful when classifying data items into categories, typically in binary form (0 and 1), such as distinguishing patients as positive or negative for diabetes. The primary goal of logistic regression is to find the best-fitting relationship between the target and predictor variables. It's rooted in the linear regression model but utilizes the sigmoid function to predict the probabilities of positive and negative classes. Despite its effectiveness, the accuracy attained from this algorithm amounts to 88%.

5 EXPERIMENTAL RESULTS

In this paper, we employed diverse classification techniques for predicting diabetes. The proposed approach incorporates various classification algorithms, including Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), and Logistic Regression (LR). The features considered in the analysis comprise A person's gender, age, family diabetes, high blood pressure, physical activity, body mass index, alcohol, tobacco use, sleep, quality of sleep, routine medicine, , stress Unhealthy foods, blood pressure levels, pregnancies, Urine Frequency, and Diabetes—attributes precisely available in the dataset. The final outcomes of these five machine learning models revealed the Random Forest algorithm achieving the highest accuracy of 98%. The results for the different classification techniques are presented in Table 2, encompassing all available features.

TABLE 2
ACCURACY ACCURACY VALUES FOR ALL FIVE USED MACHINE-LEARNING ALGORITHMS

Classifier Name	Accuracy
SVM	85%
KNN	88%
Decision tree	97%
Random forest	98%
Logistic regression	88%

6 CONCLUSIONS

ML techniques play a crucial role in disease diagnosis, particularly in achieving early detection, which enables patients to receive prompt medical attention. In this study, various machine learning algorithms were applied to the dataset, and classification was performed using multiple algorithms, with Random Forest yielding the highest accuracy of 98%. A comparative analysis of machine learning algorithm accuracies was conducted using large datasets, demonstrating the effectiveness of the model in improving the accuracy and precision of diabetes prediction compared to existing datasets.

One limitation of the study is the utilization of structured data; however, future research will explore unstructured data for further insights. Additionally, the models developed in this study can be applied or adapted to other healthcare domains for predicting diseases such as cancers, Parkinson's disease, heart disease, and Corona. Furthermore, the research scope can be expanded to consider additional attributes such as family history of diabetes, smoking habits, drinking habits, and physical inactivity for more comprehensive diabetes prediction.

ACKNOWLEDGMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCES

[1] J H. Kang, "The prevention and handling of the missing data," Korean Journal of Anesthesiology, vol. 64, no. 5. pp. 402–406, May 2013. doi: 10.4097/kjae.2013.64.5.402.

[2] . D. Accili, "Insulin action research and the future of diabetes treatment: The 2017 banting medal for scientific achievement lecture," Diabetes, vol. 67, no. 9. American Diabetes Association Inc., pp. 1701–1709, Sep. 01, 2018. doi: 10.2337/dbi18-0025.

[3] S. NAHZAT and M. YA ANO LU, "Makine Ö renimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini," European Journal of Science and Technology, Apr. 2021, doi: 10.31590/ejosat.899716.

[4] J. O. Healthcare Engineering, "Retracted: A Novel Diabetes

Healthcare Disease Prediction Framework Using Machine Learning Techniques," Journal of healthcare engineering, vol. 2023. NLM (Medline), p. 9872970, 2023. doi: 10.1155/2023/9872970.

[5] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," in Procedia Computer Science, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[6] T. Mahboob Alam et al., "A model for early prediction of diabetes," Inform Med Unlocked, vol. 16, Jan. 2019, doi: 10.1016/j.imu.2019.100204.

[7] Al Helal, M., Chowdhury, A.I., Islam, A., Ahmed, E., Mahmud, M.S. and Hossain, S., 2019, February. An optimization approach to improve classification performance in cancer and diabetes prediction. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

[8] Iancu, I., Mota, M., & Iancu, E. (2008, May). Method for the analysing of blood glucose dynamics in diabetes mellitus patients. In 2008 IEEE international conference on automation, quality and testing, robotics (Vol. 3, pp. 60-65). IEEE.

[9] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716. DOI: https://doi.org/10.1016/j.procs.2020.03.336

[10] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[11] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

[13] Breiman, L. (2001). Random forest, vol. 45. Mach Learn, 1.

[14] Tabaei, B. P., & Herman, W. H. (2002). A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care, 25(11), 1999-2003.

[15] Dutta, D., Paul, D., & Ghosh, P. (2018, November). Analysing feature importances for diabetes prediction using machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 924-928). IEEE.

[16] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET) (pp. 1-6). Ieee.

[17] Pinky, P., & Verma, G. S. A MACHINE LEARNING APPROACH FOR DETECTION OF DIABETIC SYMPTOM ON HUMAN USING IRIS.

[18] Lucaccioni, L., & Iughetti, L. (2016). Issues in diagnosis and treatment of type 1 diabetes mellitus in childhood. Journal of Diabetes Mellitus, 6(02), 175-183.

[19] Piero, M. N., Nzaro, G. M., & Njagi, J. M. (2015). Diabetes mellitus-a devastating metabolic disorder. Asian journal of biomedical and pharmaceutical sciences, 5(40), 1.

[20] Punthakee, Z., Goldenberg, R., & Katz, P. (2018). Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. Canadian journal of diabetes, 42, S10-S15.