

DIABETES DATA HANDLING BY ML ALGORITHMS: A COMPARATIVE STUDY

Rejina P V

Assistant Professor, Department of Computer Science
Co-operative Arts and Science College, Madayi
Payangadi Kannur , Kerala

Abstract

The functioning of five standard ML approaches for predicting diabetes is evaluated in this study utilizing a publicly accessible data set through Kaggle that contains pertinent medical and socioeconomic variables. To solve values that were not present, outliers, and a notable inequalities in classes (32:1 proportion of people without diabetes to mellitus samples), which presents serious obstacles to model performance; extensive preprocessing operations were carried out. Important metrics for performance such as optimal balancing recall, precision, accuracy, a standardized F1 scores, specificity, sensitivity, and rate of detection were used to evaluate the algorithmic methods— Artificial Neural Network (ANN), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Logistic Regression (LR). RF outperformed the other models in terms of performance in prediction, achieving the greatest optimal balance (93%), then a DT (82%) and SVM (83%). The results highlight how algorithms for learning can improve diabetic detection and treatment, and they suggest that algorithm choice must be guided by an appropriate balance of effectiveness indicators, particularly when working with unbalanced medical data.

Keywords: Diabetes, ML, Evaluation of Performance, SVM, ANN, DT, LR, Naïve Bayes, Random Forest

I. Introduction

Healthcare has been significantly impacted by recent developments in statistical analysis and machine learning, especially in the areas of detection of early illnesses and individualized therapy [8][10]. By examining a range of patient data, such as demographics, medical records,

and routine behaviors, algorithms based on supervised machine learning are being used more frequently to predict diabetes. DT, SVM, LR, RF, NB and ANN are well-known models used in diabetes prediction; each has unique benefits in terms of computing complexity, interpretability, and accuracy of prediction. The particulars of the collection of data as well as the balance within understanding and performance of models must be taken into consideration while choosing a technique. Many research investigations have evaluated the effectiveness of different ML (standard machine learning) models in predicting diabetes and its related problems, making the use of ML for the prediction of diabetes a rapidly developing field of study [9]. A popular approach for binary classification-based applications, such as diabetic predictability, is logistic regression (LR) [11]. Research has demonstrated that LR offers flexibility and clarity of interpretation while doing excellently with modest to moderately large collections [12]. Because Support Vector Machines (SVM) is capable of handling multi-dimensional data as well as complex non-linear correlations, they continue to be utilized extensively for mellitus diagnosis, especially using the Support Vector Classifier (SVC) [13][16]. Applied problems related to classification, decision trees (DT) are straightforward algorithms that work well [14]. They are appropriate for use in medical fields since it is simple to understand and display. It is at risk of error in fit, nevertheless, particularly when dealing with samples that are unbalanced too unstructured [15][17]. Patterned by the brain that humans have, an artificial neural network (ANN) includes a computer system designed to identify patterns and resolve challenging issues [18]. The resilience and excellent accuracy of Random Forest (RF), a collective learning technique, have made it useful for hyperglycemia prognosis [9]. To lessen the effect of overfitting and enhance standardization, it integrates several standardized decision trees [11]. A quick, ML method for identification is known as the Naive Bayes (NB) algorithm [10]. It uses the theorem known as 'Bayes' to determine the likelihood that a point of data belongs to a particular class [12]. The approach is "naive" considering that it makes the assumption that every attribute (the predictor) is completely unrelated to each additional variable.

II. Literature Review

Following Table 1 showing the review of literature based on previous research as:

Author's	Research Contributions
Jain, A., Malik, S., & Raj, A. (2026)	The latest advancements in algorithms for machine learning and big data-based predictive modeling for diabetes type 2 have been evaluated. It highlights the potential and limitations of numerous algorithms for machine learning, analyzes and evaluates their work, and suggests future directions for this area of study [1].
Afolabi, S. et al, (2025)	Proposed the study, which uses standard ML methods on digital health records to analyze the prompt identification and handling of diabetes. In order to create diabetes prediction models, the study investigates the efficacy of three supervised learning methods: KNN, RF and logistic regression [2].
Wan, X. et al., (2025)	Suggested leveraging normal laboratory data from individuals who have diabetes type 2 mellitus (T2DM) to construct machine learning (ML)-based predictive models to determine typical risk factors for diabetes-related retinopathy (DR) [3].
Jena, S. K. et al., (2025)	The suggested study shows that the accuracy of diabetes prediction is greatly improved by a group learning strategy that integrates LR, DT, RF, SVM, and Gradient Boostings. The group approach leverages each model's unique capabilities by combining them, leading to better performance than the individual model [4].
Alzboon, M. S., et al., (2025)	In 767 observations from the PIDD, the study demonstrates the predictive power provided by statistical and non-statistical ML techniques over risk for diabetes categorization. The Neural Network method achieved the highest accuracy, followed shortly by the algorithm known as Random Forest, according to the results [5].
Rahman, F., et al.,	Applying cross-validating and SHAP evaluation to improve feature

(2025)	understanding, the proposed work methodically assesses five feature selection algorithms: Boruta, Standard Genetic Algorithm, Particle Swarm Optimizer, Grey Wolf Optimizer and Recursive Feature Elimination Method [6].
Sayyid, H. O., et al., (2025)	The study was proposed to address this deficit by utilizing six machine learning algorithms (SVM, NB, LR, KNN, DT, and RF) to forecast the probability of acquiring thyroid-related conditions in an area-specific dataset of people with diabetes [7].

Table 1.Review of Literature

III. Research Methodology

Following table 2 showing the description of methodology used in this research paper, as:

S.No.	Methodology	Descriptions
1	Data Set used from	PIDD, Kaggle.com
2	Data Attributes	age, pregnancy, BMI, glucose level, BP, insulin, skin thickness, diabetes pedigree function, outcome
3	Data Preprocessing	missing values handles, dataset splitting, data standard and normalization
4	Applied Algorithms	NB, DT, LR, ANN, SVM and RF
5	Programming Language	python
6	EDA	Matplotlib

Table 2.Methodology of Research

IV. Analysis & Results

i. Performance Parameters

In this section discuss on performance parameters which used for analysis the performance of algorithms. The confusion matrix is based on tp , tn , fp and fn . The following table 3 showing the formulas of performance parameters as:

S.No	Performance Parameters	Mathematical Formula
1	Sensitivity (S_{ctv})	$tp / (tp + fn)$
2	Specificity (S_{pcty})	$tn / (tn + fp)$
3	Precision (P_{rc})	$tp / (tp + fp)$
4	Recall (R_{cal})	$tp / (tp + fn)$
5	F1 Score ($F_1 \text{ Score}$)	$2(R_{cal} / R_{cal} + P_{rc})$

6	Rate of Detection (D_{detect})	$tp / (tp + fn)$
7	Accuracy (A_{cncy})	$(S_{\text{ctv}} + S_{\text{pcty}} / 2)$

Table 3.Performance Matrices

ii. Evaluation of Performance

After analysis of algorithms under matrices then the evaluation of algorithms is showing in table 4 as:

Matrices	SVM	DT	RF	ANN	LR	NB
S_{ctv}	0.682	0.637	0.856	0.511	0.520	0.342
S_{pcty}	0.958	0.987	0.993	0.989	0.981	0.976
P_{rc}	0.987	0.852	0.874	0.942	0.923	0.989
R_{cal}	0.692	0.631	0.740	0.511	0.501	0.371
$F_1 \text{ Score}$	0.828	0.720	0.856	0.667	0.601	0.532
D_{detect}	0.019	0.015	0.035	0.012	0.011	0.008
A_{cncy}	0.821	0.813	0.920	0.752	0.723	0.672

Table 4.Performance Evaluation of Algorithms

iii. Results of Evaluation

The evaluation of algorithms calculates under different performance matrices, the performance result is showing in figure 1 as:

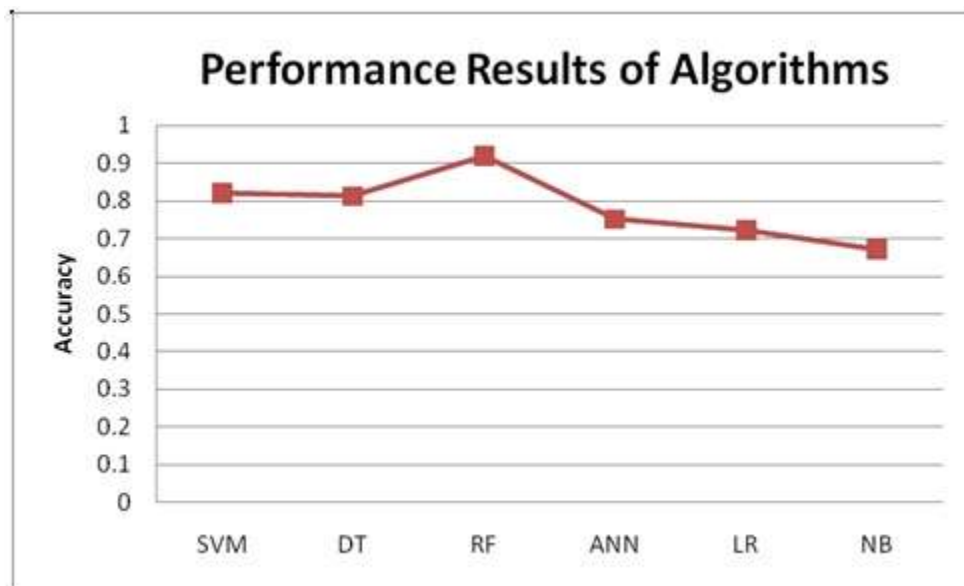


Figure 1.Performance Results of Algorithms

V. Conclusion & Future Scope

F This research aimed to assess the success rate of six supervised, ML algorithms in predicting type 2 diabetes based on a healthcare data collection. Within the approaches studied, the standard Random Forest (RF) along with Support Vector Machine (SVM) had higher generalization and accuracy as well. This highlights the possibility of ML approaches in timely identification of diseases and reducing costs associated with healthcare by proactive measures. Nevertheless, the investigation has a number of drawbacks, such as the utilization of limited data sets that could not be comparable to the larger population, the omission of other pertinent clinical characteristics, and the dependence on one data set, that brings up questions about possible bias and overfitting. To better assess the durability of such algorithms, further studies should include cross-validation tests between various medical settings as well as areas. In addition, working with immediate diagnostic data as well as continuous information from patients could improve precision and customization of diabetic risk forecasting. The method's effectiveness may also differ across various data sets or real-life situations.

References

- [1] Jain, A., Malik, S., & Raj, A. (2026). Diabetes Mellitus Type 2 and Prediction Making Use of Big Data and Machine Learning Algorithms. *Central Asian Journal of Medical and Natural Science*, 7(2), 476-484.
- [2] Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, 2(1), 9-16.
- [3] Wan, X., Zhang, R., Wang, Y., Wei, W., Song, B., Zhang, L., & Hu, Y. (2025). Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms. *European Journal of Medical Research*, 30(1), 183.
- [4] Jena, S. K., Behera, D. K., Jena, A. K., & Rout, J. K. (2025). A Novel Ensemble Machine Learning Framework for Improved Diabetes Prediction and Complication Prevention. *Procedia Computer Science*, 258, 4008-4017.
- [5] Alzboon, M. S., Alqaraleh, M., & Al-Batah, M. S. (2025). Diabetes prediction and management using machine learning approaches. *arXiv preprint arXiv:2506.11501*.

- [6] Rahman, F., Hossain, S., Tiang, J. J., & Nahid, A. A. (2025). Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers. *Diagnostics*, 15(20), 2622.
- [7] Sayyid, H. O., Mahmood, S. A., & Hamadi, S. S. (2025). Comparison of machine learning algorithms for predicting thyroid disorders in diabetic patients. *Informatica*, 49(12).
- [8] Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*, 5, 100297.
- [9] Abdollahi, J., & Aref, S. (2024). Early prediction of diabetes using feature selection and machine learning algorithms. *SN Computer Science*, 5(2), 217.
- [10] Göde, A., & Kalkan, A. (2023). Performance comparison machine learning algorithms in diabetes disease prediction. *European Mechanical Science*, 7(3), 178-183.
- [11] Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3), 1728-1737.
- [12] Saxena, S., Mohapatra, D., Padhee, S., & Sahoo, G. K. (2023). Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evolutionary intelligence*, 16(2), 587-603.
- [13] Bhat, S. S., Banu, M., Ansari, G. A., & Selvam, V. (2023). A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics*, 4, 100273.
- [14] Ganie, S. M., Malik, M. B., & Arif, T. (2022). Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches. *Journal of Diabetes & Metabolic Disorders*, 21(1), 339-352.
- [15] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2, 100118.

- [16] Gupta, H., Varshney, H., Sharma, T. K., Pachauri, N., & Verma, O. P. (2022). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*, 8(4), 3073-3087.
- [17] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773.
- [18] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome*, 14(1), 196.