INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 2 - V11I2-1170)
Available online at: https://www.ijariit.com

Comparative Analysis of Machine Learning Models for Diabetes Prediction: A Performance Evaluation Study

Taaha Ansari
taahaswork@gmail.com
Alamuri Ratnamala Institute of Engineering and Technology, Tute, Maharashtra

Vaishali M. Bagade
vaishalibagadejan22@gmail.com
Alamuri Ratnamala Institute of Engineering and Technology, Tute, Maharashtra

ABSTRACT

Diabetes is a chronic disease affecting millions worldwide, necessitating early diagnosis and effective prediction models for improved healthcare outcomes. This study evaluates seven machine learning algorithms for diabetes prediction using healthcare data. We compared Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, AdaBoost, XGBoost, and Support Vector Machine (SVM) models. The analysis focused on key performance metrics: accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Results showed that logistic regression achieved the highest overall performance with 79% accuracy and 0.88 AUC, suggesting its potential utility in clinical diabetes prediction applications.

Keywords: *Diabetes Prediction, Machine Learning, Logistic Regression, KNN, F1-Score, AUC.*

INTRODUCTION

Diabetes mellitus represents a significant global health challenge, affecting approximately 537 million adults worldwide [1]. Early detection and prediction of diabetes risk are crucial for preventing complications and improving patient outcomes. Machine learning approaches offer promising solutions for automated, efficient diabetes risk assessment using readily available health indicators [2].

Diabetes is a metabolic disorder characterized by high blood glucose levels, which can lead to severe complications if left unmanaged. Early prediction and diagnosis can significantly improve patient outcomes [3]. Traditional methods of diabetes diagnosis rely on clinical assessments and biochemical tests, which may be time-consuming and costly. Machine learning offers a promising alternative by leveraging data-driven approaches to identify patterns and predict diabetes risk with high accuracy [4] [5]. Several studies have explored the application of machine learning in diabetes prediction. Researchers have employed various classification techniques, such as Logistic Regression, Decision Trees, and Neural Networks, to enhance diagnostic accuracy. Ensemble learning and hybrid models have also been investigated to improve robustness and reliability [6]. However, the challenge remains in selecting the most suitable model for clinical implementation.

METHODOLOGY

This study utilizes a dataset containing health-related parameters such as age, BMI, glucose levels, blood pressure, insulin levels, and family history of diabetes. The workflow consists of data preprocessing, feature selection, model training, evaluation, and comparison. Figure 1 describes the proposed methodology used in our research.

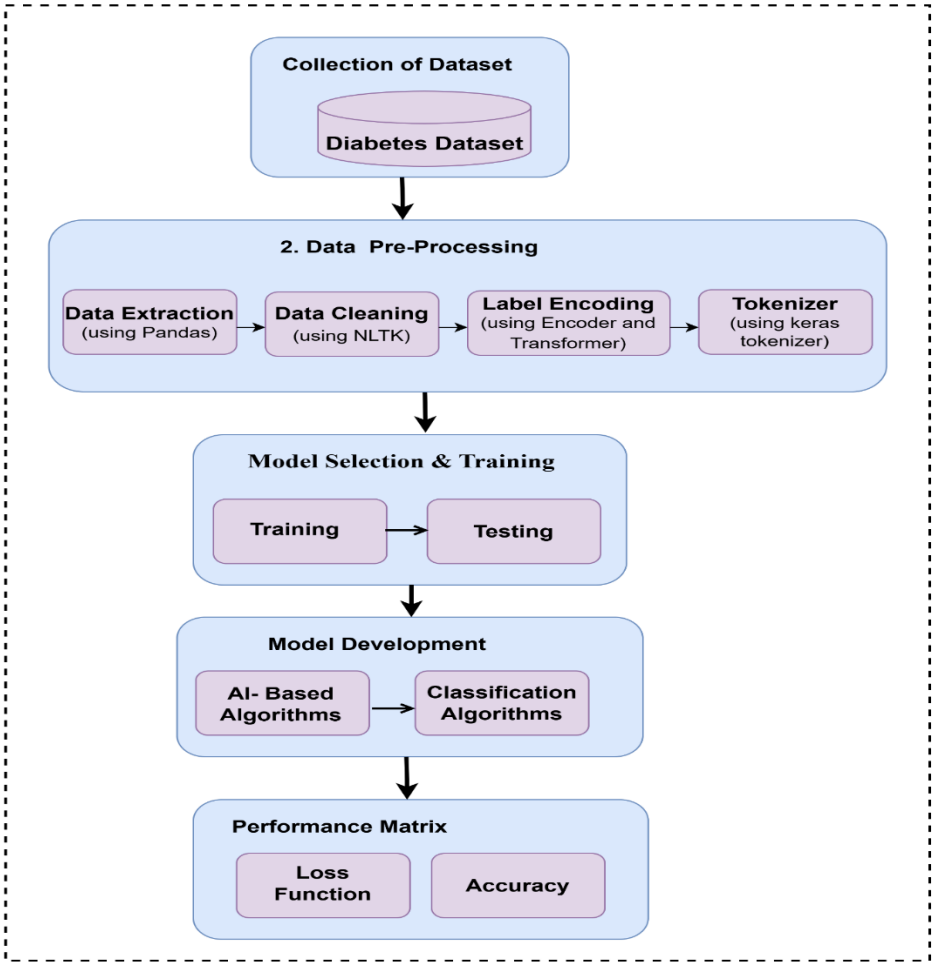

Figure 1: Proposed Methodology

1) **Data Collection & Preprocessing:** Collect diabetes-related datasets. Handle missing values, outliers, and normalization. Perform feature selection. Figure 2 gives details of different steps applied at dat preprocessing. Dataset Preprocessing: Normalizing numerical features for consistency. Encoding categorical variables using one-hot encoding [7][8].

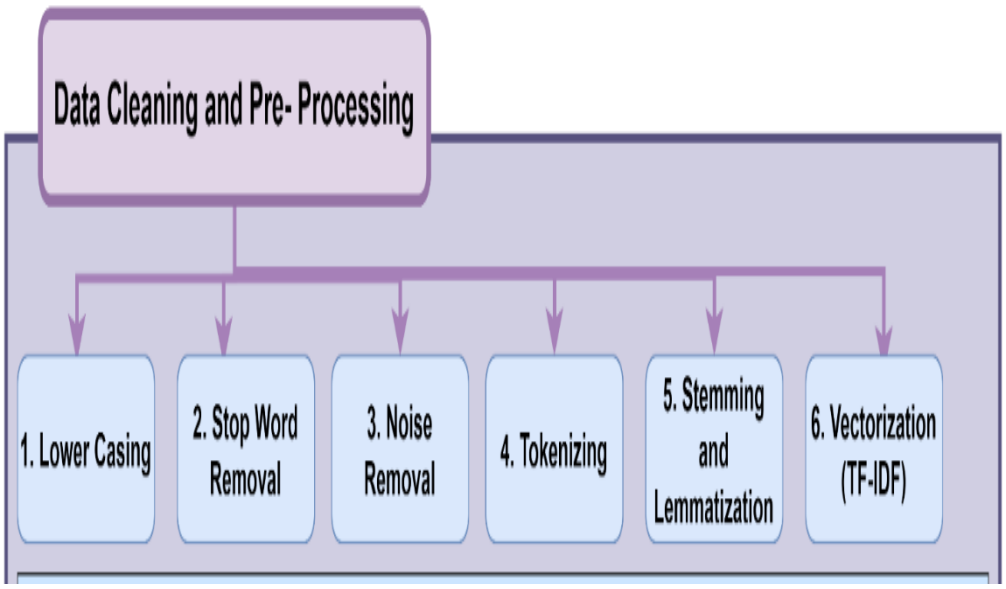

Figure 2: Data cleaning and pre-proessing

2) **Model Selection & Training:** Apply ML models (e.g., Logistic Regression, Decision Trees, Random Forest, SVM, Neural Networks). Train the models on the dataset. Splitting data into training and testing sets (80%-20%)
3) **Evaluation & Optimization:** Assess model performance using accuracy, precision, recall, and F1-score. Optimize hyperparameters and fine-tune models. Deployment & Prediction: Deploy the best-performing model. Use the model for real-time diabetes prediction [9].

4) **Model Development:**

a) **Logistic Regression:** Logistic Regression is a widely used algorithm for binary classification tasks, including text classification. It's a simple yet effective method that can serve as a good baseline model for many NLP tasks. We have trained the logistic regression model using the preprocessed and feature-extracted training data. Also, we used a predefined library of Python sci-kit-learn to implement the LR technique. We calculated different performance metrics, which are shown in the results section. Logistic regression uses the logistic function, also known as the sigmoid function, to model the probability that a given input example belongs to the positive class (class 1).

b) **Random Forest Classifier:** An ensemble learning technique called a Random Forest Classifier is utilized for classification and regression tasks. It is a member of the decision tree algorithm family. Its primary characteristic—randomness in building several decision trees and then merging their predictions—gives rise to the word "random" in its name. Random Forest is strong and less prone to overfitting. It manages data in both category and numerical form. It functions well "out of the box " with minimal hyperparameter tweaking." An ensemble of Decision Trees to improve accuracy and reduce overfitting.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

Where,
$h_i(x)$ is the prediction from the ith decision tree and
N is the total number of trees.

c) **A Support Vector Classifier (SVC),** a Support Vector Machine (SVM) for classification, is a supervised machine learning algorithm for binary and multiclass classification tasks. SVCs are particularly effective when the data is not linearly separable in its feature space because they can find a hyperplane that best separates the classes while maximizing the margin between them. SVCs use a kernel function to map the input data into a higher-dimensional space where the classes might be linearly separable. Standard kernel functions include Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. The SVC algorithm seeks to find the hyperplane that best separates the data into different classes. It does this by solving an optimization problem that maximizes the margin (distance) between the hyperplane and the nearest data points of each class. Support vectors are the data points closest to the hyperplane and play a crucial role in determining the hyperplane's position. We build an SVC classifier with a linear kernel gamma value as scale. A hyperplane-based classifier for high-dimensional data.

$$f(x) = sgn(w^T \varkappa + b)$$

Where,
w is the weight vector,
$\varkappa$ is the input feature vector, and
b is the bias.

d) **Decision Tree:** A tree-based approach that splits data based on feature conditions. Decision Tree: A more popular machine learning algorithm for classification and prediction tasks on supervised data is the decision tree classifier. Rules and trees are used to classify the learned dataset. It outlines the classification criteria for category data. The Random Forest, Support Vector Machine, and Naive Bayes algorithms—discussed earlier—and new classifiers. The studies are conducted independently for each algorithm and combined for optimal accuracy and precision. The decision tree classifier's first step is identifying the trait that will split the tree. The entropy measure is used to determine the optimal attribute of the decision tree. With the split on each property, the updated inhomogeneity is computed using the entropy measure.

e) **Light Gradient Boosting:** The gradient-boosting framework utilizes tree-based learning methods, which are highly efficient computational processes. This algorithm processes data quickly. The Light GBM method develops vertically, i.e., leaf-wise, whereas other algorithms grow horizontally because they are trees.

PERFORMANCE METRICS

We computed several measures, including accuracy, precision, recall, F1 score, and ROC-1) AUC score, to assess the effectiveness of our model. These indicators offer several viewpoints on the performance of your model. The following describes each metric's standard calculation and meaning:

1) **Accuracy:** The percentage of correctly identified samples relative to the total number of samples is known as accuracy. It's a typical measure of the overall performance of the model.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad ---------- (1)$$

2) **Precision:** The percentage of true positive predictions—accurately predicted positive samples—among all positive predictions is known as precision. It evaluates how accurate positive forecasts are.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ positives} \quad (2)$$

3) **Recall (Sensitivity):** Recall measures the proportion of true positive predictions from all actual positive samples. It assesses the ability of the model to capture all positive samples.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

4) **F1 Score:** The F1 score is the harmonic mean of precision and recall. It balances precision and recall, making it useful when the class distribution is imbalanced.

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (4)$$

5) **ROC-AUC Score:** The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) score is commonly used for binary classification problems. It measures the area under the ROC curve, which plots the true positive rate (TPR or recall) against the false positive rate (FPR) at various thresholds. A higher ROC-AUC score indicates better discriminative ability of the model.

These metrics can be calculated using sci-kit-learn functions in Python.

In these calculations:

y_true represents the true labels of your test data.
y_pred represents the predicted labels of your model.
y_scores represent the predicted class probabilities (only needed for ROC-AUC).

RESULTS AND DISCUSSION

Each model was assessed using accuracy, precision, recall, and F1-score. The results are summarized in Table 1:

Table 1: Output of Machine Learning Techniques with Different Performance Metrics

CLASSIFIER	ACCURACY	PRECISION	RECALL	F1- SCORE	AUC
LOGISTIC REGRESSION	**79%**	**78%**	**79%**	**79%**	**0.88**
KNN	76%	78%	76%	76%	0.85
RANDOM FOREST	78%	78%	78%	78%	0.87
DECISION TREE	73%	75%	73%	73%	0.75
ADA BOOST	78%	79%	78%	78%	0.85
XG BOOST	78%	78%	78%	78%	0.84
SVM	75%	78%	75%	76%	0.87

DISCUSSION

Logistic Regression achieved the highest accuracy (79%) and AUC (0.88), making it the most effective model in this study. Random Forest and AdaBoost also performed well, with an accuracy of around 78%.

The Decision Tree classifier had the lowest accuracy (73%) and AUC (0.75), indicating that it may not be as effective for this dataset. Details are shown in Figure 3

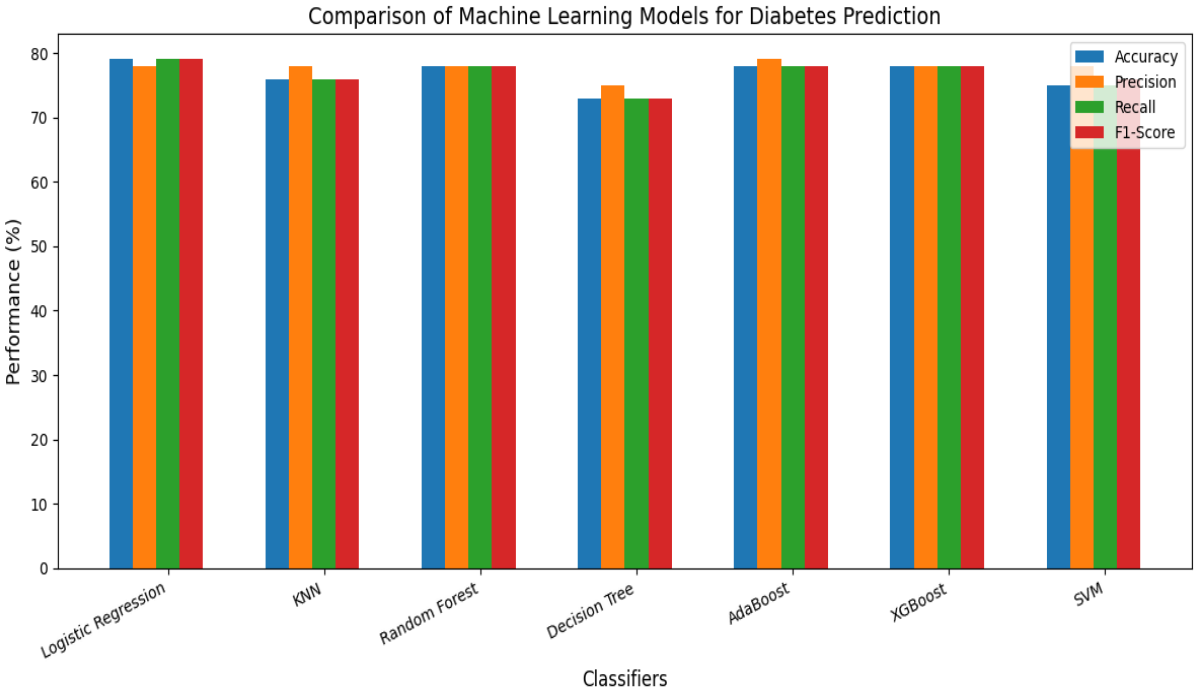

Figure 3: Comparison of machine learning models for diabetes prediction with all performance metrics

The analysis reveals several key findings: This comparative analysis demonstrates the effectiveness of machine learning approaches in diabetes prediction, with Logistic Regression showing promising results. The high performance across multiple metrics suggests these models could serve as valuable tools in clinical settings. The balance between model complexity and performance indicates that simpler models may be sufficient for practical applications in diabetes risk assessment. Future research directions include the Investigation of deep learning approaches, Integration of temporal medical data, and Development of interpretable risk prediction models.

1. **Model Efficiency**
   - Logistic regression's superior performance suggests that diabetes prediction may follow relatively linear patterns
   - The high AUC scores across most models indicate robust discriminative ability
   - Ensemble methods (Random Forest, AdaBoost, XGBoost) showed consistent performance
2. **Clinical Implications**
   - The high precision across models (75-79%) suggests reliable optimistic predictions.
   - Consistent recall values indicate balanced sensitivity in the detection
   - The models' performance supports their potential integration into clinical decision-support systems
3. **Practical Considerations**
   - Logistic regression's simplicity, interpretability, and superior performance make it particularly suitable for clinical applications.
   - The marginal performance differences between complex ensemble methods suggest diminishing returns from model complexity.

CONCLUSION

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed. The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. Preprocessing techniques have been applied to handle the issue of imbalanced class problems. This research paper reported different performance metrics: precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques. The Logistic Regression achieved the best performance with 79% accuracy and an F1 score and AUC of 0.79 and 0.78, respectively, with the next approach. Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed Logistic Regression framework has been deployed into a web app to predict diabetes instantly. This work has some future scopes; for example, we recommend getting additional private data from a

larger cohort of patients to get better results. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.

REFERENCES

[1] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.
[2] Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*, 294-305.
[3] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, *9*(09), 2278-0181.
[4] Febrian , ME , Ferdinan , FX , Sendani , GP , Suryanigrum , KM , & Yunanda , R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science* , *216* , 21-30.
[5] Mahadevkar, S. V., Khemani, B., Patil, S., Kotecha, K., Vora, D. R., Abraham, A., & Gabralla, L. A. (2022). A review on machine learning styles in computer vision—techniques and future directions. *Ieee Access*, *10*, 107293-107329.
[6] Khemani, B., Patil, S., Kotecha, K., & Vora, D. (2024). Detecting health misinformation: A comparative analysis of machine learning and graph convolutional networks in classification tasks. *MethodsX*, *12*, 102737.
[7] Ruano-Ordás, D. (2024). Machine Learning-Based Feature Extraction and Selection. *Applied Sciences*, *14*(15), 6567.
[8] Ngo, V. D., Vuong, T. C., Van Luong, T., & Tran, H. (2024). Machine learning-based intrusion detection: feature selection versus feature extraction. *Cluster Computing*, *27*(3), 2365-2379.
[9] Zhan, Q., Sun, D., Gao, E., Ma, Y., Liana, Y., & Yang, H. (2024, August). Advancements in feature extraction recognition of medical imaging systems through deep learning technique. In *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 1211-1216).