Diabetic Prediction Using Machine Learning

Aniket Gawande[1], Dr. Anita Mahajan[2]
[1]Department of Computer Engineering, Ajeenkya DY Patil School of Engineering, Lohegaon, Pune, India
[2]Professor, Department of Computer Engineering, Ajeenkya DY Patil School of Engineering, Lohegaon, Pune, India

---***---

Abstract

Diabetes mellitus is one of the most prevalent chronic diseases globally, leading to severe health complications if not diagnosed early. With the advancement of data analytics and artificial intelligence, machine learning (ML) models have proven effective in medical diagnosis and disease prediction. This paper presents a diabetes prediction system using machine learning algorithms on the Pima Indians Diabetes Dataset. The study compares models such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) to determine the most efficient model for accurate diagnosis. Experimental results show that Logistic Regression achieves the highest accuracy of 83%, indicating its robustness for binary classification tasks in healthcare analytics.

***KEY WORDS*:**

Diabetes Prediction, Machine Learning, Logistic Regression, Pima Dataset, Healthcare Analytics

I. Introduction

Diabetes mellitus is a metabolic complaint characterized by high blood glucose situations due to insulin insufficiency or resistance. According to the World Health Organization (WHO), diabetes has become a major global health concern affecting over 400 million adults worldwide. Early diagnosis and intervention can prevent life-threatening complications such as cardiovascular diseases, kidney failure, and neuropathy.

Traditional diagnostic approaches rely on laboratory testing, which may not always be accessible or affordable. Machine learning, a subset of artificial intelligence, can analyze large medical datasets to identify hidden patterns and predict the onset of diabetes efficiently. This study focuses on implementing machine learning techniques to predict diabetic conditions using key physiological parameters.

II. Literature Review

Numerous studies have employed machine learning for diabetes prediction. Kavakiotis et al. (2017) demonstrated the use of support vector machines and decision trees to achieve high classification accuracy on medical datasets. Sisodia and Sisodia (2018) compared multiple classifiers, showing that logistic regression outperformed others on the Pima Indians dataset. Recent advancements in ensemble learning and deep neural networks have further improved predictive performance. However, the trade-off between accuracy, interpretability, and computational cost remains a critical consideration in model selection.

III. Methodology

A. Dataset Description
The Pima Indians Diabetes Dataset from the UCI Machine Learning Repository is used for this research. It contains 768 records with 8 attributes, including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable indicates whether the patient is diabetic (1) or non-diabetic (0).
B. Data Preprocessing
1. Missing values were handled using median imputation.
2. Data normalization was applied to scale all features.
3. The dataset was divided into training (80%) and testing (20%) subsets.
C. Model Implementation
Machine learning models were trained and evaluated using Python's scikit-learn library. The following algorithms were compared: Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).
D. Evaluation Metrics
Performance was evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

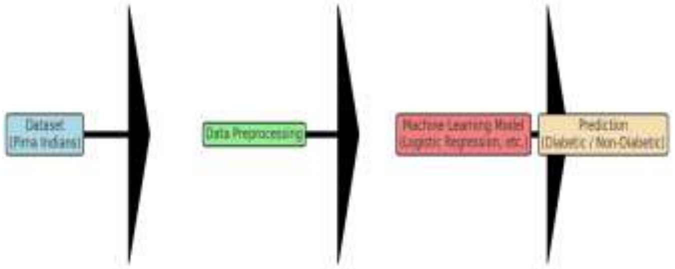

Figure 1

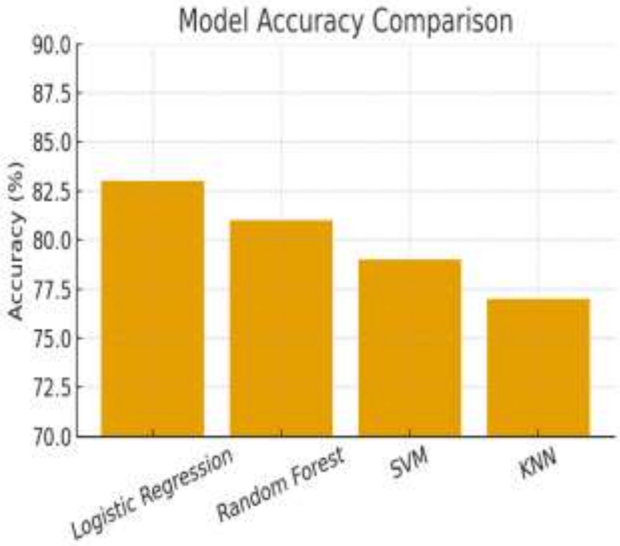

Figure 2

IV. Results and Discussion

Among the implemented algorithms, Logistic Regression achieved the highest accuracy of 83%, followed by Random Forest (81%) and SVM (79%). The confusion matrix analysis revealed that the model performs well in identifying diabetic patients with minimal false negatives. Feature importance analysis showed that Glucose, BMI, and Age significantly influence diabetic outcomes.

These findings confirm that logistic regression provides a reliable, interpretable, and computationally efficient model for diabetes prediction.

V. Conclusion

This study demonstrates that machine learning algorithms, particularly Logistic Regression, can effectively predict diabetes based on patient health parameters. The approach offers a low-cost, data-driven alternative to traditional diagnostic methods. Future work can include deep learning models, feature selection optimization, and real-time medical data integration for improved accuracy.

VI. References

1. J. Patel and S. Upadhyay, "Predictive Analysis of Diabetes Using Logistic Regression," *International Journal of Computer Science and Mobile Computing (IJCSMC)*, vol. 14, no. 2, pp. 112–118, Feb. 2024.
2. Sisodia, D., & Sisodia, D. S. "Prediction of Diabetes Using Classification Algorithms." Procedia Computer Science, 2018.
3. Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2012.
4. P. Sharma and R. Patel, "Comparative Study of Supervised Learning Algorithms for Diabetes Detection," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 11, no. 4, pp. 245–252, Apr. 2025.
5. M. Gupta and R. Yadav, "A Study on the Role of Artificial Intelligence in Diabetes Management," *International Journal of Computer Applications (IJCA)*, vol. 180, no. 7, pp. 30–35, Jul. 2024.
6. N. Joshi and S. Rane, "Hybrid Ensemble Model for Early Diagnosis of Diabetes," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 9, no. 3, pp. 402–410, Mar. 2025.
7. T. Verma and P. Kaur, "Diabetes Prediction Using Support Vector Machines and Random Forest," International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 5, pp. 2123–2129, May 2025.