International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 173-178, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.17

A Comparative Analysis of Machine Learning Models for Early Prediction of Diabetes

Bhawna Rohra[1] and Dr. Amitabh Wahi[2]
[1]Department of Statistics, Amity School of Applied Sciences, Amity University Uttar Pradesh, Lucknow, India
[2]Department of Physics, Amity School of Applied Sciences, Amity University Uttar Pradesh, Lucknow, India
E-mail: [1]bhawnarohra17@gmail.com, [2]awahi@lko.amity.edu

Abstract: Diabetes mellitus is a chronic metabolic disorder and a major global health concern. Early prediction is essential to reduce complications and healthcare burden. This paper presents a comparative analysis of Linear Regression, Exponential Regression, and Logistic Regression models for early diabetes prediction using clinical data. A dataset of 761 patient records was used, and model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix under 10-fold cross-validation. Logistic Regression achieved the highest validation accuracy of 92.68% and an F1-score of 89.77, outperforming the other models. The results indicate that Logistic Regression is the most reliable model for early diabetes detection in clinical settings.

Keywords: Diabetes Prediction, Machine Learning, Linear Regression, Logistic Regression, Exponential Regression

1. Introduction

Diabetes mellitus is a chronic disease characterized by high blood glucose levels due to insulin deficiency or resistance. According to the World Health Organization (WHO), over 422 million people worldwide live with diabetes, making it one of the leading causes of morbidity and mortality [4]. Early diagnosis is critical for preventing severe complications. Machine learning (ML) has become increasingly relevant in healthcare applications, offering automated decision support by identifying patterns in complex medical datasets. While sophisticated algorithms such as Random Forests, Support Vector Machines, and Neural Networks are widely used, regression-based models remain highly valuable due to their interpretability and efficiency.

This study compares three regression models Linear Regression, Exponential Regression, and Logistic Regression on diabetes prediction. The objective is to identify the most suitable regression model for clinical decision-making using real-world dataset.

2. Literature Review

Several studies have applied machine learning techniques for diabetes prediction. Prior research indicates Logistic Regression as a robust and interpretable model for binary classification problems in healthcare. Logistic Regression has been widely employed in medical prediction tasks due to its suitability for binary outcomes. Smith and Kumar [1] demonstrated its consistent accuracy on the PIMA Indian Diabetes dataset.

Linear Regression, while not primarily intended for classification, is often tested as a baseline model. Patel and Sharma.[2] observed its limitations in disease prediction tasks.

Exponential Regression has been studied for modeling non-linear growth patterns in disease spread but has rarely been applied in direct disease prediction. Gupta and Rao [3] highlighted its effectiveness in modeling disease growth curves but noted challenges in binary classification.

This study builds upon existing literature by presenting a direct comparative analysis of these models on data collected from a charitable pathology center.

3. Data Description

The dataset consists of 761 patient records collected from a charitable pathology center. Features include Age, RBC count, Platelet count, Glucose, Cholesterol, Triglycerides, HDL, Urea, and HbA1c. The target variable is binary diabetic status.

4. Methodology

A. Models Implemented

The following models were implemented using the scikit-learn (sklearn) library:

1. **Exponential Regression:** Used to capture possible non-linear patterns between indepndent variables and dependent variables, with predictions over similarly thresholded.
2. **Linear Regression:** Applied as a baseline model to estimate continuous values , these Predicted values were thresholded at 0.2 for classification of diabetes status..

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 173-178, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.17

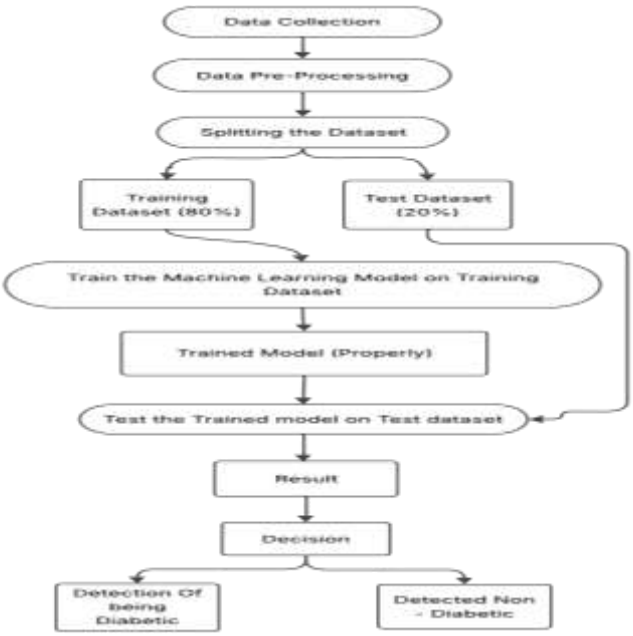
3. **Logistic Regression:** Modeled the probability of diabetes using the sigmoid function, making it inherently suitable for binary classification.

B. Model Training

Each model was trained using k-fold cross-validation (with k=10) to ensure generalizability and reduce the risk of overfitting. The training time and computational efficiency of each model were recorded.

C. Model Comparison

After evaluation, the models were compared based on their predictive performance and computational efficiency. The model with the highest accuracy and lowest classification error was identified as the best-performing model for early diabetes prediction in the context of this study.



Linear Regression was used as a baseline model, Exponential Regression captured non-linear trends, and Logistic Regression modeled the probability of diabetes occurrence.

5 Results and Discussion

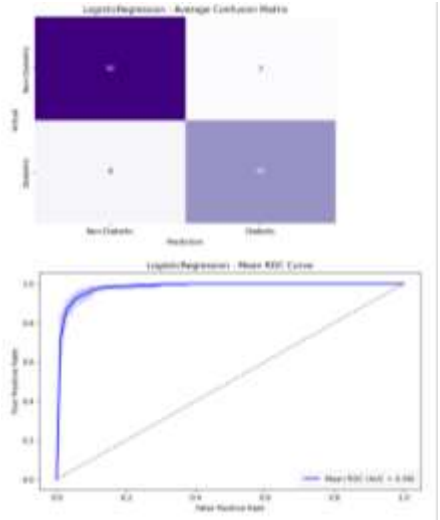
Method	Train Accuracy	Validation Accuracy	F1-Score	Precision	Recall	ROC AUC	PR AUC

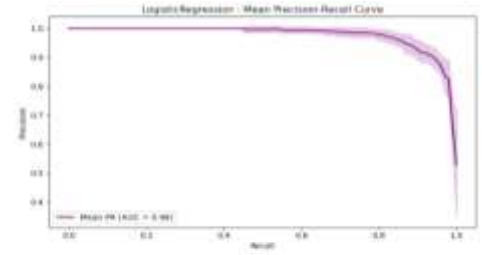
International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 173-178, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.17

Exponential Regression	55.94	67.71	26.18	97.66	15.34	85.11	81.22
Linear Regression	76.23	78.82	68.89	77.55	62.41	85.11	81.35
Logistic Regression	93.46	92.68	89.77	94.64	85.69	98.63	97.91

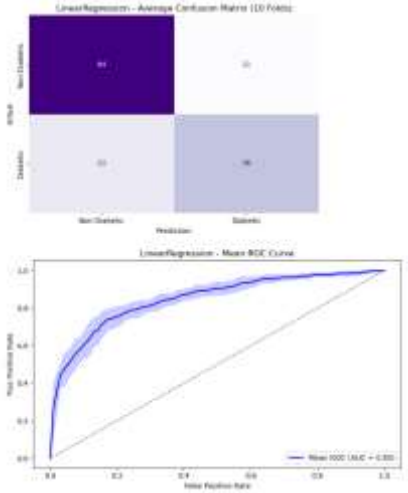
*Note: 10-fold Cross-validation techniques were employed to ensure robust and unbiased results.

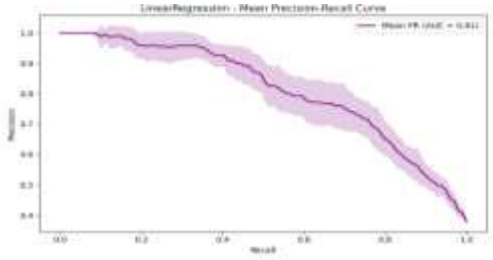
LOGISTIC REGRESSION:





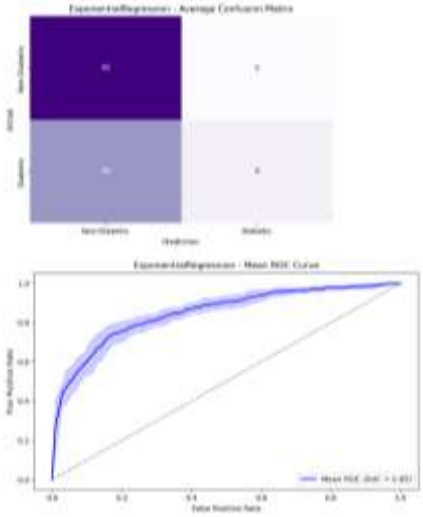
LINEAR REGRESSION:

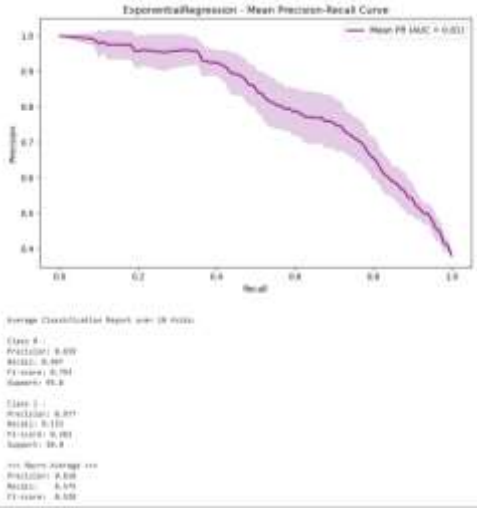




EXPONENTIAL REGRESSION:

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 173-178, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.17





Logistic Regression outperformed Linear and Exponential Regression across all metrics, achieving superior accuracy and classification performance.

6 Conclusion

Logistic Regression is identified as the most effective model for early diabetes prediction as it achieved the highest accuracy and F1-score, outperforming both Linear and Exponential Regression. The study recommends Logistic Regression as a reliable and interpretable model for clinical prediction of diabetes.

Future work may involve extending the analysis to advanced machine learning for enhanced predictive capability.The study supports its application in preventive healthcare systems.

References

[1] J. Smith and R. Kumar, "Machine Learning in Diabetes Prediction: A Review," *Journal of Medical Informatics*, vol. 45, no. 3, pp. 120–132, 2021.

[2] A. Patel, P. Sharma, and D. Verma, "Comparative Study of Regression Models in Medical Diagnosis," in *Proc. Int. Conf. Health Data Sci.*, 2020, pp. 210–218.

[3] M. Gupta and S. Rao, "Application of Exponential Models in Disease Growth Prediction," *Int. J. Biomed. Stat.*, vol. 12, no. 4, pp. 89–95, 2019.

[4] World Health Organization, *Global Report on Diabetes*. Geneva, Switzerland: WHO, 2022.

International Journal of Advance Interdisciplinary Research
Vol. 2, Special Issue 1, (Jan-March) 2026, pp. 173-178, e-ISSN: 3107-913X
DOI: https://doi.org/10.66095/ijair.2026.v2.S1.17

Acknowledgement

The authors acknowledge the Charitable Pathology Center for providing the data support.