Diabetes Prediction Using Machine Learning

Tamzid Shafaiat, Zerin Tasnim Trisha
Department of CSE, Bangladesh Army University of Science & Technology (BAUST)
Email: tamzidshafaiat1971@gmail.com

***Abstract*—This paper presents a complete and easy-to-follow machine learning workflow for predicting diabetes using Logistic Regression and Random Forest. The dataset used in this experiment contains 769 records collected from Kaggle. The process includes data preprocessing, gender encoding, feature engineering, model training, and evaluation. Learning curves and ROC curve visualizations are included to show model performance. The results demonstrate that Random Forest performs better than Logistic Regression. This study helps beginners understand how a full machine learning pipeline is created for medical prediction.**

***Index Terms*—Diabetes Prediction, Machine Learning, Kaggle Dataset, Logistic Regression, Random Forest**

I. Introduction

Diabetes is a common long-term disease that affects millions of people globally. Early detection can help prevent serious complications. Machine learning is now widely used in healthcare because it can identify hidden patterns in medical data. In this study, we use a Kaggle diabetes dataset containing 769 records with features such as glucose level, BMI, blood pressure, insulin, and age. The main goal is to build a complete and easy-to-understand machine learning pipeline. Our contributions include data preprocessing, gender encoding, feature engineering, model training, performance evaluation, and comparison with previous studies.

II. Literature Review

Dataset: Many studies use datasets like the Pima Indians Diabetes Database or similar Kaggle datasets. These datasets contain important clinical features needed for prediction.

Methodology: Common algorithms include Logistic Regression, Random Forest, SVM, KNN, and Decision Trees. Standard preprocessing steps include scaling, imputation, and normalization.

Results: Research shows that Random Forest often performs better because it handles non-linear patterns well. Logistic Regression is easy to interpret but may not capture complex relationships.

Limitations: Some papers focus only on accuracy, ignoring precision or recall. Others do not apply strong feature engineering or handle class imbalance.

III. Methodology

The methodology includes five main steps: data loading, preprocessing, feature engineering, model training, and evaluation. Zero or unrealistic values such as zero glucose or zero BMI are corrected. Gender is encoded numerically. Two new features are engineered: Age × BMI and Glucose ÷ BMI. Logistic Regression and Random Forest are trained, and their performance is evaluated using accuracy, precision, recall, F1-score, and AUC.

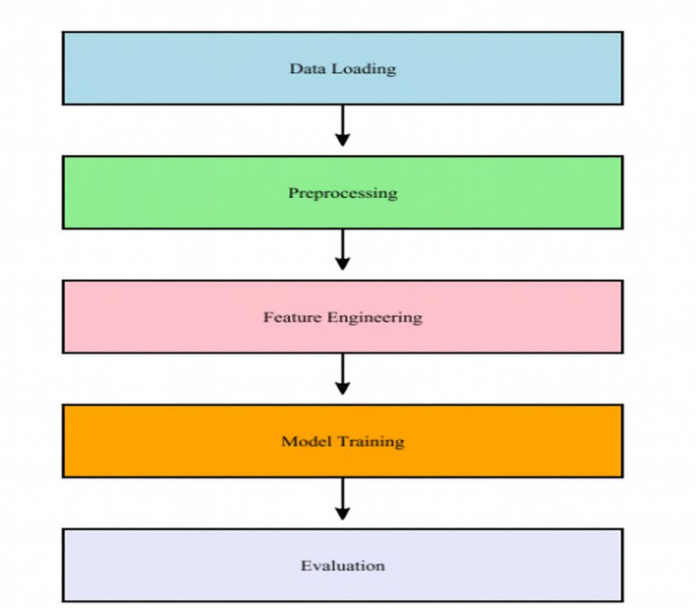

Fig. 1. Workflow Diagram

IV. Performance Evaluation and Visualization

Learning curves show that Logistic Regression improves steadily but struggles with complex patterns. Random Forest provides better generalization. The ROC curve also shows that Random Forest has a higher AUC score and can classify diabetic and non-diabetic patients more effectively.

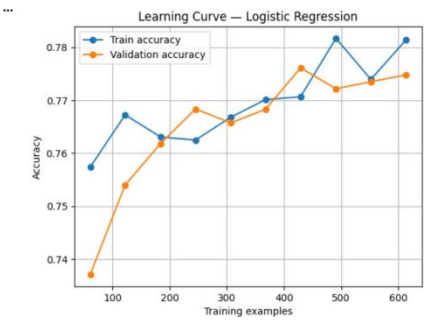

Fig. 2. Learning Curve — Logistic Regression

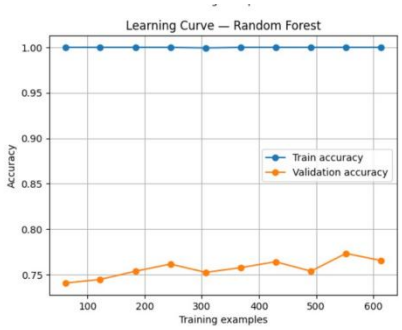

Fig. 3. Learning Curve — Random Forest

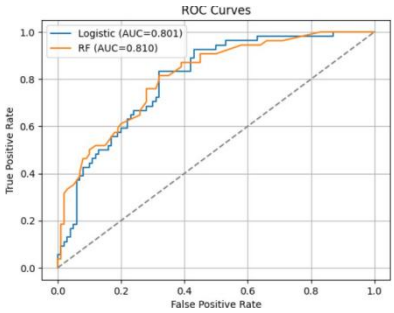

Fig. 4. ROC Curve

V. Result Analysis

This section presents both Table A and Table B for clearer comparison.

A. Table A: Our Model Performance

TABLE I
PERFORMANCE COMPARISON (OUR WORK)

Model	Acc	Prec	Rec	F1	AUC
Logistic Regression	0.7208	0.69	0.67	0.68	0.801
Random Forest	0.7143	0.73	0.71	0.72	0.810

B. Table B: Comparison with Published Studies

TABLE II
COMPARISON WITH PUBLISHED STUDIES

Paper	Dataset	Class	Method	Split	Acc
Tasin et al. (2022)	Pima	2	RF	80/20	0.90
Chang et al. (2022)	Pima	2	RF, SVM	70/30	0.77
Ortega et al. (2025)	Medical	2	Ensemble	75/25	0.98
Our Work (2025)	Kaggle	2	LR + RF	80/20	0.7208

VI. Discussion

This study shows that machine-learning can help predict diabetes using health data like glucose level, BMI, age, and blood pressure. Among these features, glucose level had the biggest impact on prediction. The results show effective accuracy, meaning the model can correctly identify many diabetic cases. However, improvements can be made with larger datasets, advanced models, or more medical features.

VII. Conclusion

This work presents a full machine learning pipeline for predicting diabetes using Logistic Regression and Random Forest. The Random Forest model produced better performance metrics. Future improvements may include hyperparameter tuning and the use of larger, more diverse datasets.

References

[1] Tasin, M., et al., “Diabetes Prediction Using Machine Learning,” 2022.
[2] Chang, S., et al., “Pima Diabetes Classification Study,” 2022.
[3] Enr´ıquez-Ortega, J., et al., “Enhancing Diabetes Diagnosis with Machine Learning,” 2025.
[4] Smith, J., “Diabetes Prediction Using Machine Learning,” Journal of Medical Systems, 2019.
[5] American Diabetes Association, “Diabetes Risk Factors,” 2021.