A Comparative Study of Logistic Regression and Other Classification Models for Predicting Diabetes in Pima Indian Women

Author : Porawat Katiya & Jaturapat Pinchai **Advisors** : ACTING CAPT. DR.Chalermrat Nontapa
Master of Science Program in Data Science, Faculty of Engineering, Chaing Mai University

Abstract

Diabetes mellitus is a major global health concern and one of the most common chronic diseases, this study developed a Logistic Regression model to predict its occurrence in Pima Indian women. Using a dataset of 769 patients, the analysis identified glucose level, BMI, and age as significant predictors. The model demonstrated strong predictive accuracy and interpretability, making it valuable for clinical screening. Future work should compare its performance against advanced machine learning models, such as Random Forest or LightGBM, to enhance prediction accuracy.

Introduction

Diabetes Mellitus is a leading global health concern, with the Pima Indian population exhibiting one of the highest prevalence rates worldwide. This makes their health data a critical benchmark for research. Early prediction is essential for effective patient management and preventing severe complications. This study develops and evaluates a Logistic Regression model, valued for its interpretability, to predict diabetes in this high-risk group. Its performance is then systematically compared against other advanced classification algorithms to assess its effectiveness and utility in a clinical context.

Methodology

Data Collection and Preparation

data collected and made available by the National Institute of Diabetes and Digestive and Kidney Diseases, comprises 769 female patients of Pima Indian descent, all aged 21 years and above.It contains a total of nine variables, namely: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the Outcome. The dataset was appropriately cleaned and preprocessed prior to analysis.

Logistic Regression analysis and Forecasting

1. Check that the dependent variable is binary
2. Check for linearity in the logit between continuous predictors and the log odds of the outcome
3. Ensure no multicollinearity among independent variables
4. Select independent variables using the stepwise regression base on AIC technique
5. Verify the appropriateness of the obtained model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

where:

- y is dependent variable as binary and $y_i \sim \text{Bernoulli}(p_i)$
- $x_1, x_2, \ldots, x_j$ is independent variable where j = 1,2, ... ,k
- β_0 is Intercept (the value of y when all independent variables are zero)
- $\beta_1, \beta_2, \ldots, \beta_k$ is regression coefficients represents the change in y for a one-unit increase in x_k

Model Adequacy and Model Accuracy

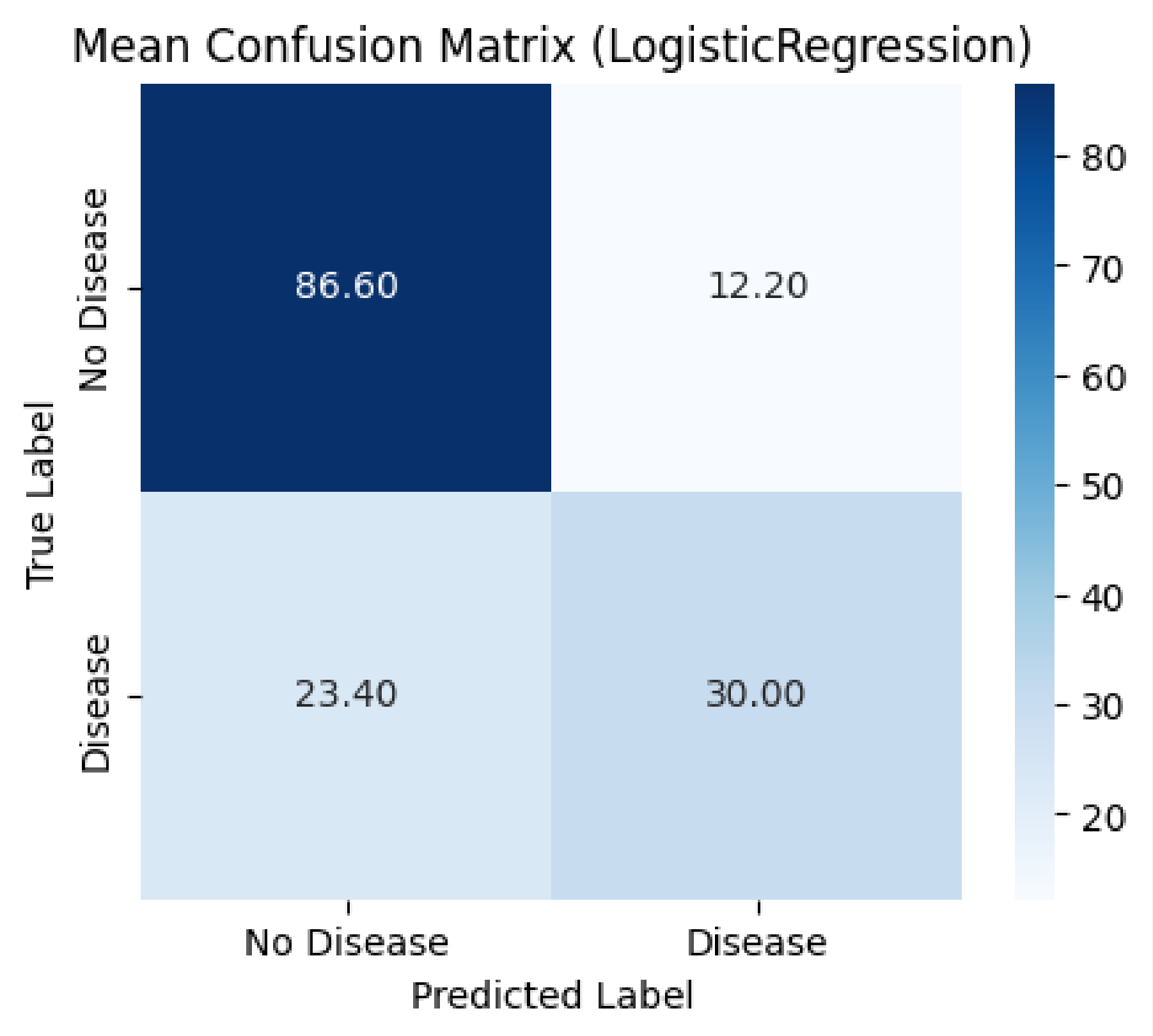
After obtaining a verified model, the forecasting model's performance and predictive capability will be evaluated using several classification metrics, including Accuracy, Precision, Recall, F1-score, AUC-ROC, and Confusion Matrix.

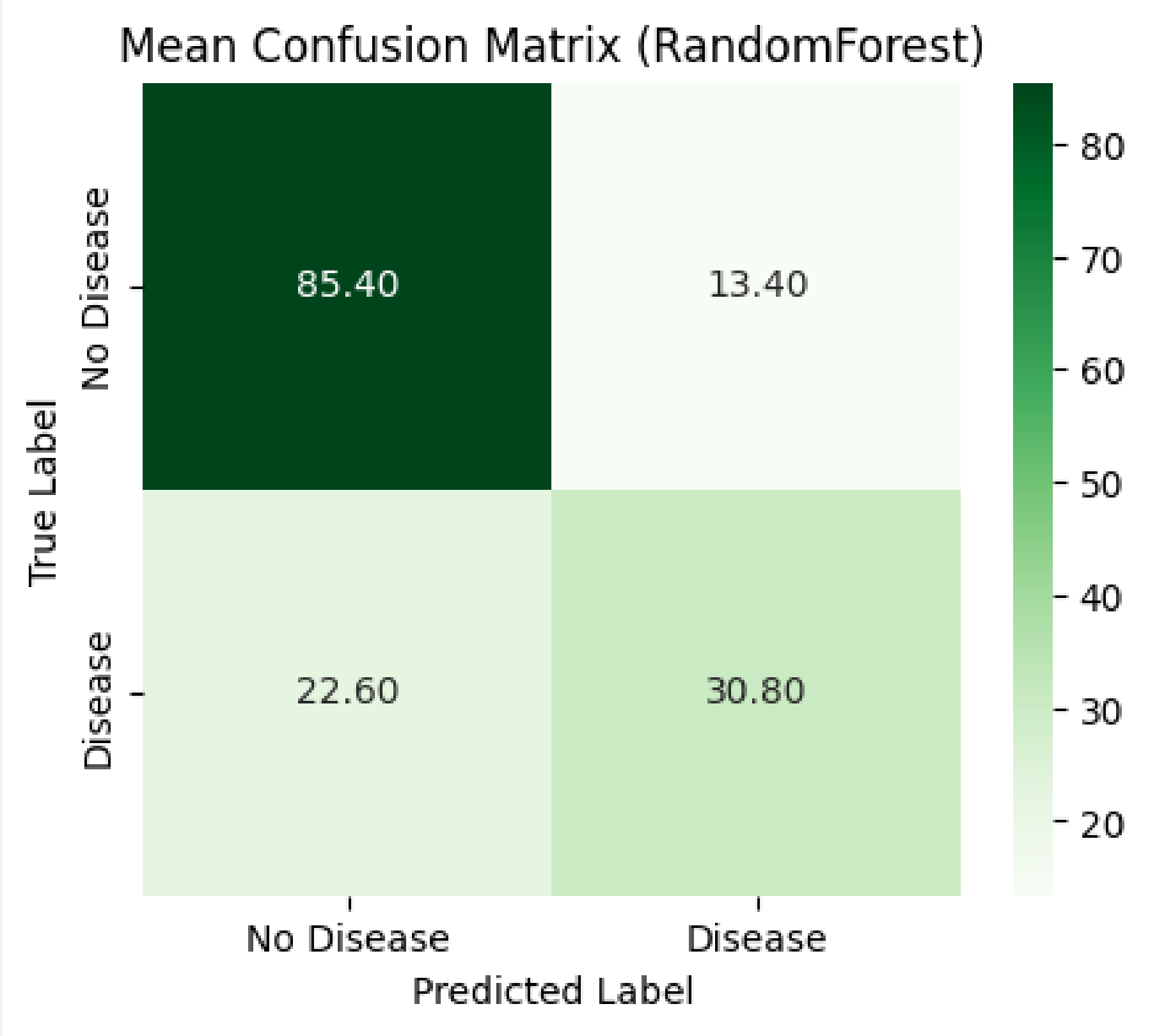
Objectives

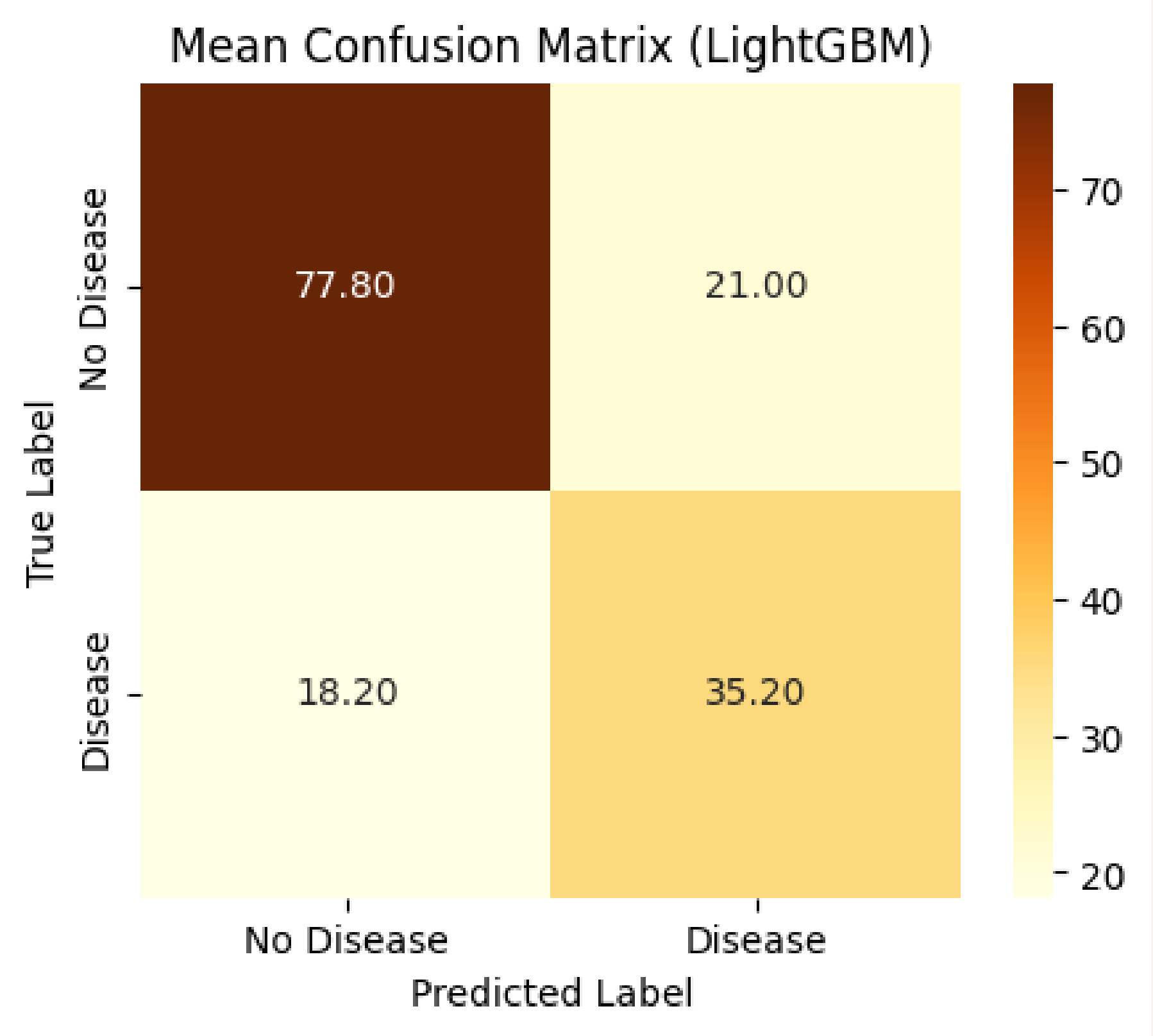
- To analyze factors affecting diabetes in Pima Indian women
- To develop a predictive model and evaluate its performance against other classification models

Results

Forecasting model	β	Std.Err.	z	p-value
(Constant)	-9.834	0.835	-11.784	0.000
Glucose	0.038	0.004	9.706	0.000
BMI	0.095	0.017	5.614	0.000
Pregnancies	0.142	0.031	4.611	0.000
DPedigreeFunction	1.756	0.549	3.196	0.000







Model Performance	Logistic Regression	Random Forest	LightGBM
Accuracy	0.766	0.764	0.743
Precision	0.712	0.699	0.631
Recall	0.562	0.577	0.66
F1-score	0.626	0.627	0.639
AUC-ROC	0.838	0.836	0.814

Reference

- Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. Critical Care, 9(1), 112–118.
- Rahman, M. M., & Islam, M. Z. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches based on the Pima Indians dataset. Journal of Biomedical Research and Environmental Sciences, 2(8), 745–753.

Conclusion and discussion

The Logistic Regression model effectively identified key factors associated with diabetes among Pima Indian women—namely glucose level, BMI, and age. It showed strong predictive performance across metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC, confirming reliable classification of diabetic and non-diabetic cases. Its interpretability adds clinical value by clarifying how each factor influences diabetes risk. However, performance may be constrained by sample size and class imbalance. Future work could improve accuracy and robustness by comparing with advanced models like Random Forest, Gradient Boosting, or LightGBM.