Pima Diabetes Prediction – Final Report

1. Problem Statement

Diabetes is a major global health challenge, with over 589 million adults affected worldwide.

This project aims to:

- Build a machine learning model to predict diabetes using the PIMA Indian Diabetes dataset.
- Identify high-risk groups, determine key clinical predictors, and develop interpretable features .

2. Model Outcomes or Predictions

This is a supervised classification problem (Outcome: 0 = non-diabetic, 1 = diabetic).

Models evaluated included Logistic Regression, Random Forest (Feature Importance), LightGBM, XGBoost, and Decision Tree (for interpretability). The expected output is a predicted class label.

3. Data Acquisition

Dataset: PIMA Indian Diabetes Dataset from Kaggle.

Features include: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.

Data visualization confirmed that Glucose, BMI and Insulin were strong predictors of diabetes.

4. Data Preprocessing / Preparation

• Zero-based missing values (Glucose, BP, SkinThickness, Insulin, BMI) were replaced with NaN.

• Median imputation applied within each Outcome group.

• Data split: 80% training, 20% testing (stratified).

• Decision-tree derived thresholds and statistics of features helped in generating new interpretable binary features (Feature_1 ... Feature_9, etc.).

5. Modeling

Models tested: LightGBM, XGBoost, XGBoost + KNN. LightGBM+ KNN

XGBoost performed the best due to strong non-linear learning, robustness, and regularization.

Hyperparameter tuning was performed with GridSearchCV,RandomizedSearchCV (learning_rate, depth, subsample, colsample, reg_alpha, reg_lambda, etc.).

6. Model Evaluation

5-fold cross-validation was used. XGBoost achieved:

Accuracy ≈ 0.884, Precision ≈ 0.85, Recall ≈ 0.814, Specificity ≈ 0.922,

F1-Score ≈ 0.831, ROC-AUC ≈ 0.95.

This demonstrates strong and reliable predictive performance.

7. Interpretation: High-Risk Groups

Decision-tree, statistics rules identified the following **high-risk** groups:

• Insulin > 121 (strongest predictor)

• Glucose > 127.5

• BMI ≥ 35

• Age > 28.5 combined with high Insulin

• Pregnancies ≥ 6 in older women

• SkinThickness > 31 (obesity pattern)

Low-risk groups include:

• Glucose ≤ 105 and BloodPressure ≤ 80

• BMI ≤ 30 and SkinThickness ≤ 31

• Pregnancies ≤ 5

• Insulin ≤ 121

8. Conclusion

XGBoost was the best-performing model. Interpretable rule-based features derived from the decision tree and statistics helped identify meaningful clinical groups with different diabetes risks. The final model supports early screening and targeted interventions for high-risk individuals. Furthermore, model can be tested on large dataset with balanced class, validation and improvement can be made for clinical features identified. New features can be extracted to further increase the models robustness.