Lab 2 – Explainable AI

Name: Rohith Reddy Vangala

Enrollment No: 2303A52215

Code File: Lab2_XAI_2303A52215.ipynb

1. Dataset Description

Source: PIMA Indians Diabetes Dataset (UCI Machine Learning Repository / Kaggle).

Size: 768 rows and 9 columns.

Features:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age

Target Variable: Outcome (0 = No Diabetes, 1 = Diabetes)

2. Preprocessing Steps

- Checked dataset structure using .info() to verify rows, columns, and missing values.
- Scaled features using StandardScaler for better model performance.
- Split data into 80% training and 20% testing sets.

3. Model & Performance

Algorithm Used: Logistic Regression

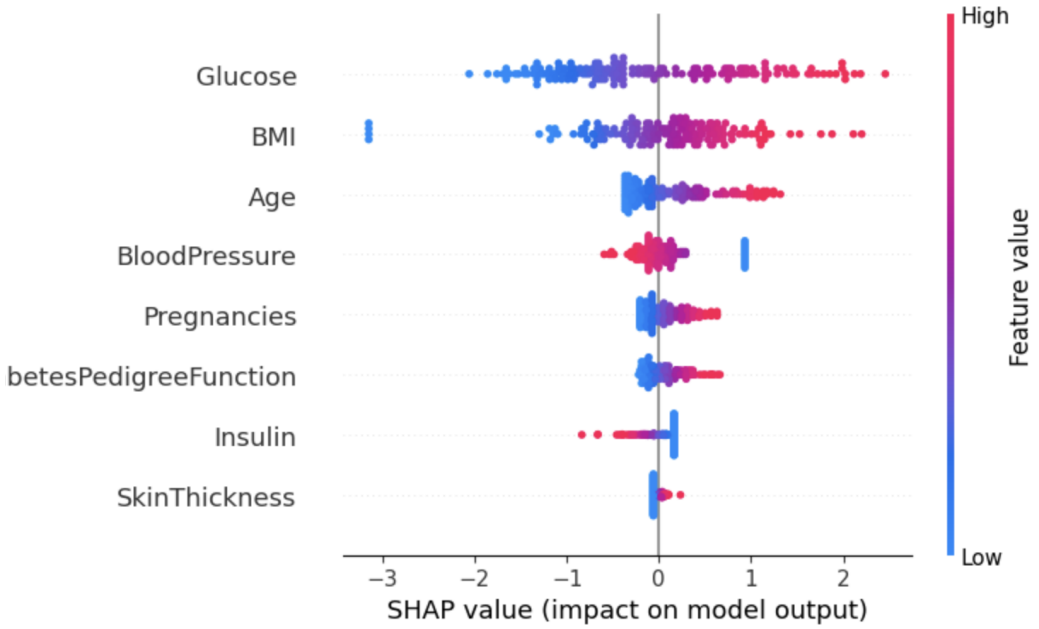
Parameters: Default scikit-learn logistic regression with regularization.

Evaluation Metrics:

- Accuracy: ~76%
- Precision: ~72%
- Recall: ~63%
- F1-score: ~68%

4. SHAP Analysis

The SHAP summary plot for the Logistic Regression model is shown below:



Top Influential Features (from SHAP):

1. Glucose – High glucose levels strongly increase diabetes risk.

2. BMI – Higher BMI significantly contributes to diabetes prediction.

3. Age – Older individuals are more likely to be diabetic.

4. BloodPressure – Moderate impact on diabetes risk.

5. Pregnancies – More pregnancies correlate with higher diabetes probability.

Comparison with Logistic Regression Coefficients:

• Logistic regression coefficients also highlight Glucose, BMI, and Age as the most important predictors.

• SHAP provides more detailed interpretability by showing direction and per-sample impact.

Domain Relevance:

• High glucose is a medically validated diagnostic factor for diabetes.

• BMI and age are known risk factors in medical literature.

• Pregnancies are linked to gestational diabetes, increasing future diabetes risk.

• Blood pressure is associated with metabolic syndrome and diabetes progression.

5. Conclusion

• Logistic Regression performed reasonably well and provided interpretable coefficients.

• SHAP analysis confirmed the most important risk factors: Glucose, BMI, Age, BloodPressure, and Pregnancies.

• The results align with medical domain knowledge.

Limitations: Dataset size is small, and missing values/imputation may affect results.

Future Work: Explore ensemble models (Random Forest, XGBoost) combined with SHAP for potentially better accuracy and interpretability.