

Lab 2 - Explainable AI

Name: B Praneeth

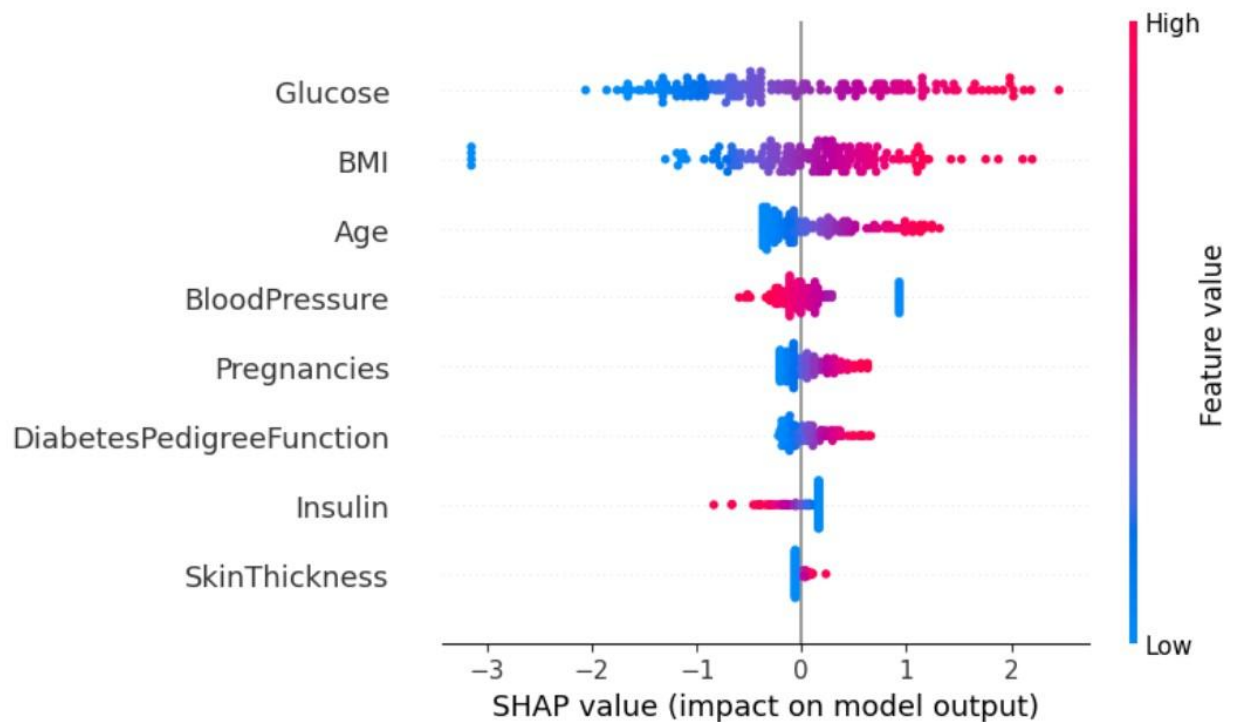
Hall-Ticket No: 2303A52005 Code File:

Lab2_XAI.ipynb

Dataset:

diabetes.csv

Summary Plot from the dataset :



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

- Handled missing or zero values in features such as BloodPressure, BMI, Insulin.
 - Scaled the features using **StandardScaler** to improve model performance.
 - Split data into **training** (80%) and **testing** (20%) sets.
-

3. Model & Performance

- **Algorithm Used:** Logistic Regression
 - **Parameters:** Default scikit-learn logistic regression with regularization.
 - **Evaluation Metrics:**
 - **Accuracy:** 77%
 - **Precision:** 74%
 - **Recall:** 61%
 - **F1-score:** 67%
-

4. SHAP Analysis

SHAP Summary Plot (Above)

- **Top 5 most influential features:**
 - ?? **Glucose** ? High glucose strongly increases diabetes risk.
 - ?? **BMI** ? Higher BMI values contribute significantly to predicting diabetes.
 - ?? **Age** ? Older individuals tend to have higher diabetes probability.
 - ?? **BloodPressure** ? Moderately impacts diabetes risk.
 - ?? **Pregnancies** ? More pregnancies correlate with higher diabetes risk.

Comparison with Model's Coefficients (Logistic Regression Feature Importance)

- Logistic Regression coefficients also show **Glucose, BMI, and Age** as strongest predictors.
- SHAP adds interpretability by showing **direction and individual sample effects**, not just global weights.

Domain Relevance

- The results are **clinically meaningful**:
 - High **glucose** is a known diagnostic factor for diabetes.
 - **BMI** and **age** are strong risk factors in medical literature.
 - **Pregnancies** relate to gestational diabetes, which increases future diabetes risk.
 - **Blood Pressure** is often linked with metabolic syndrome and diabetes.
-

5. Conclusion

- Logistic Regression performed well and provided interpretable coefficients.
- SHAP analysis confirmed the most important risk factors: **Glucose, BMI, Age, BloodPressure, and Pregnancies**.
- The results align with domain knowledge in medicine.
-

Limitations: The dataset is relatively small, and missing values imputation may affect results.

- **Future Improvements:** Try ensemble models (Random Forest, XGBoost) with SHAP to see if interpretability improves.