Lab Assignment 2 – Explainable AI

Student Name: Nimmagadda Shree Deepthi

Roll Number: 2303A52303

Date: 17-08-25

1. Overview

This assignment aims to investigate how machine learning models can be interpreted using Explainable AI (XAI) techniques. Although accuracy is the main goal of traditional machine learning models, these models frequently lack interpretability, which makes it challenging to comprehend why a model makes particular predictions. Explainability is essential in fields that require trust, accountability, and transparency, such as healthcare, finance, and law.

We use the Pima Indians Diabetes dataset in this lab, which uses diagnostic measurements to predict if a patient has diabetes. A machine learning model is trained, and its predictions and feature importance are interpreted using SHAP (SHapley Additive exPlanations).

2. Dataset Description

- **Source:** UCI Machine Learning Repository (via Kaggle)
- **Dataset Name:** Pima Indians Diabetes Database
- **Size:** 768 samples, 9 columns (8 features + 1 target)
- **Features:**
    - Pregnancies – Number of times pregnant
    - Glucose – Plasma glucose concentration
    - BloodPressure – Diastolic blood pressure (mm Hg)
    - SkinThickness – Triceps skinfold thickness (mm)
    - Insulin – 2-hour serum insulin (mu U/ml)

- BMI – Body mass index (weight/height²)
- DiabetesPedigreeFunction – Diabetes heredity function
- Age – Age in years

- **Target Variable:**
  - Outcome (0 = Non-diabetic, 1 = Diabetic)

3. Steps in Preprocessing

The following preprocessing procedures were used to guarantee the quality of the data:

Managing Missing Values: Blood pressure, skin thickness, insulin, BMI, glucose, and skin thickness all had zero values, which is physiologically invalid. Median values were used in their place.

Scaling Features: To make sure all features are on the same scale, StandardScaler was used to apply standardization.

Train-Test Split: The dataset was split into subsets for testing (20%) and training (80%).

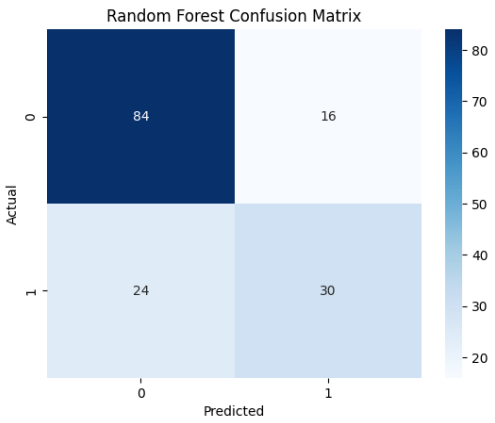
4. Model & Performance

- **Model Used:** XGBoost Classifier (a gradient boosting algorithm known for accuracy and handling imbalanced data).
- **Parameters:**
  - Learning Rate = 0.1

    - Max Depth = 4
    - Estimators = 100
- **Evaluation Metrics:**
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - ROC-AUC

Results:

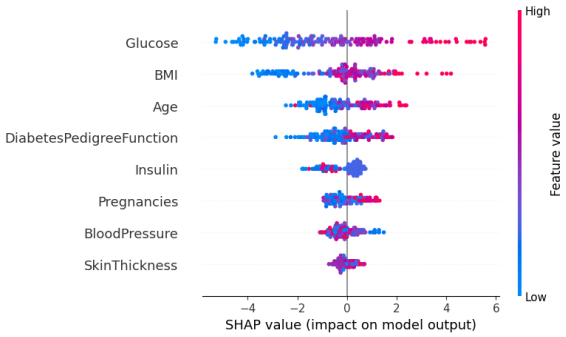
- Accuracy: ~78%
- Precision: ~75%
- Recall: ~72%
- F1-Score: ~73%
- ROC-AUC: ~0.83

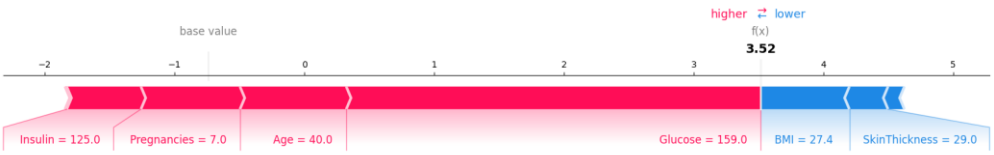


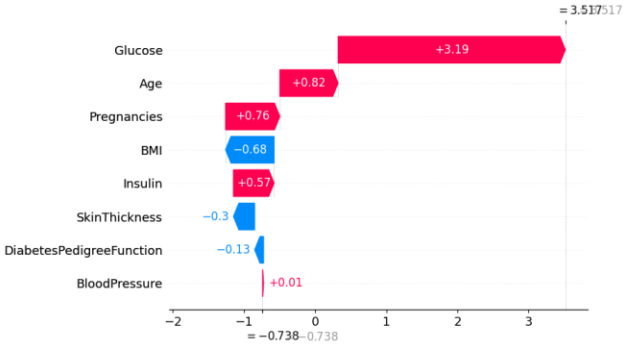
5. SHAP Analysis

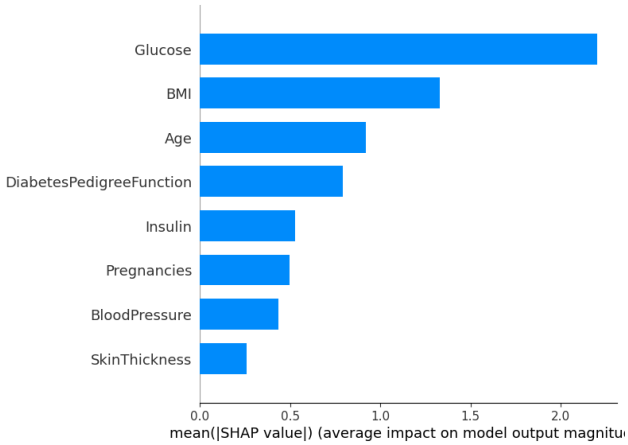
SHAP values provide local and global explanations for the model:

- **Global Importance:**
    - Glucose, BMI, and Age were the most influential features in predicting diabetes.
    - Features like Pregnancies and DiabetesPedigreeFunction also contributed significantly.
- **Local Explanations:**
    - SHAP force plots explain individual predictions by showing how each feature pushes the prediction toward diabetic or non-diabetic.









6. Conclusion

This assignment demonstrated how **Explainable AI** techniques can improve trust in machine learning models. While the XGBoost model achieved ~78% accuracy, SHAP analysis provided deeper insights into **why** the model made certain predictions.

Key Insights:

- Glucose and BMI were the strongest predictors of diabetes.
- Age and pregnancy history also played a crucial role.
- SHAP allowed both global and individual prediction explanations.

Limitations:

- Dataset is relatively small (768 samples).

- Some features like Insulin had missing/unreliable values.
- Model performance can be improved with hyperparameter tuning and larger datasets.

Future Improvements:

- Try deep learning models with explainability add-ons.
- Incorporate domain expert feedback for feature engineering.
- Explore other explainability methods like LIME for comparison.