Lab 2 - Explainable AI

Name: BELLAMPALLY KOWSHIK REDDY

Enrollment No: 2303A52002

Code File: XAI_2303A52002_Lab_Assignment_2.ipynb

1. Dataset Description

Dataset: PIMA Indians Diabetes Dataset (Kaggle / UCI ML Repository)

Size: 768 rows × 9 columns

Features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction,

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

2. Preprocessing Steps

Verified dataset structure using .info()

Standardized features using StandardScaler

Data split: 80% training, 20% testing

3. Model & Performance

Model Used: Logistic Regression (default parameters)

Evaluation Metrics:

Accuracy: ~76%

Precision: ~72%

Recall: ~63%

F1-score: ~68%

4. SHAP Analysis

Applied SHAP values for interpretability

Top Influential Features:

Glucose - Strong positive impact on diabetes risk # Replaced en dash with hyphen

BMI - High BMI increases probability # Replaced en dash with hyphen

Age - Older age associated with higher diabetes risk # Replaced en dash with hyphen

BloodPressure - Moderate influence # Replaced en dash with hyphen

Pregnancies - More pregnancies linked with higher risk # Replaced en dash with hyphen

Comparison with Logistic Regression Coefficients: Both SHAP and coefficients confirm Glucose, BMI, and

5. Conclusion

Logistic Regression gave interpretable and consistent results

SHAP confirmed domain knowledge: Glucose, BMI, and Age are key risk factors

Limitations: Small dataset size, missing value imputation may affect accuracy

Future Work: Try Random Forest or XGBoost with SHAP for improved performance