Name – Shruthi Addagudi

2303A52133 Batch – 37

Report for Explainable AI – Assignment 2

1. Objective

The aim of this assignment was to apply **Explainable AI (XAI) techniques** to a machine learning model trained on an **IoT Intrusion Detection dataset**. The goals include:

- Preprocessing the dataset
- Building a classification model
- Evaluating performance metrics
- Using **SHAP (SHapley Additive exPlanations)** for explainability

2. Dataset

- **Source:** `IoT_Intrusion.csv`
- **Shape:** Rows × Columns (based on initial load)
- **Target:** `label` (categorical, later encoded into numeric form)

Preprocessing Steps

1. Removed duplicates and missing values.
2. Encoded categorical target labels with `LabelEncoder`.
3. Scaled features using **StandardScaler**.

3. Model Development

- **Algorithm Used:** Random Forest Classifier
- **Parameters:** `n_estimators=100`, `random_state=42`, `n_jobs=-1`
- **Train-Test Split:** 80-20

4. Performance Evaluation

Metrics calculated:

- **Accuracy**

- **Precision, Recall, F1-score**
- **ROC-AUC Score (macro, multi-class)** using One-vs-Rest

This ensures balanced evaluation of the classifier beyond just accuracy.

5. Explainability (XAI)

- Used **SHAP library** to interpret model predictions.
- **Global Explanation:** Feature importance plot (summary plot).
- **Local Explanation:** Force plots for individual predictions.

These insights highlight:

- Which features influenced classification most.
- Transparency in model decisions for intrusion detection.

6. Key Findings

1. Random Forest achieved **good performance** on IoT intrusion dataset.
2. **SHAP values** revealed which features were most impactful in detecting intrusions.
3. The workflow demonstrates how **XAI techniques enhance trust and interpretability** in security-related ML applications.

7. Conclusion

This lab successfully integrated:

- **Data preprocessing**
- **Model training & evaluation**
- **Explainable AI (SHAP) insights**

Such methods are vital in domains like **cybersecurity**, where **interpretability** of model predictions is crucial for trust and decision-making.