LAB ASSIGNMENT -3

MAMIDALA SHIVAMANI
2303A52344
BATCH-38

Report for Explainable AI – Assignment

Problem 1: IoT Intrusion Detection with LIME

Problem Statement

The task was to classify IoT network traffic as **attack** or **normal**, using a Random Forest classifier. The model should be explained with **LIME (Local Interpretable Model-agnostic Explanations)** to identify which network features influence predictions.

Steps Followed

1. **Data Loading**
   - Loaded the IoT dataset (`data.csv`) with traffic records and attack/benign labels.
2. **Preprocessing**
   - Dropped identifier fields (IP addresses).
   - Encoded categorical columns such as `proto`, `service`, `conn_state`.
   - Filled missing values with zeros.
   - Converted labels: *Benign/Normal = 0*, *Attack = 1*.
3. **Feature Extraction**

    - Used all numeric and encoded categorical features (packet counts, byte counts, duration, connection states).

4. **Model Training**
    - Trained a **Random Forest** classifier.
    - Achieved **Accuracy = 100%** on the test set.

5. **Explainability with LIME**
    - Applied a LIME-style explainer to individual predictions.
    - For an **attack** instance, the most influential features included:
        - Protocol type (`proto`)
        - Source/destination ports
        - Connection state (`conn_state`)
        - Packet/byte counts

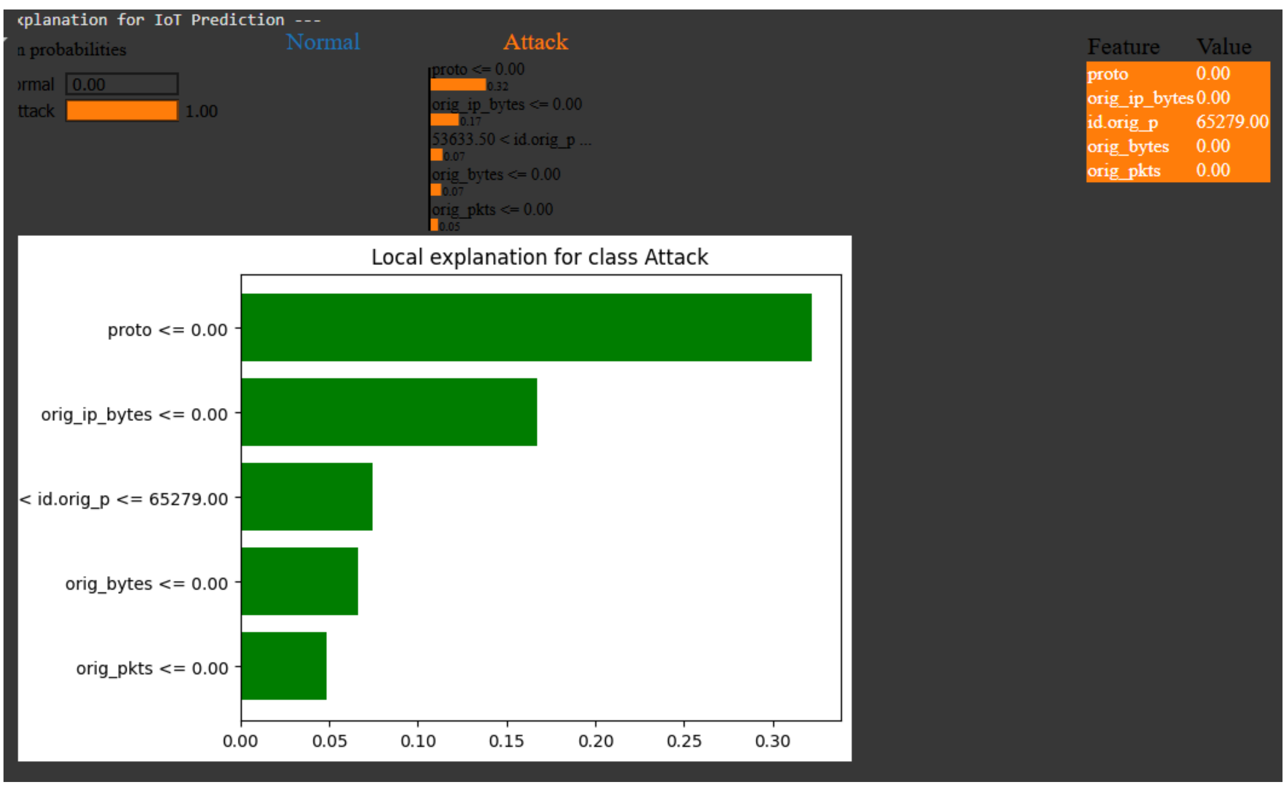
Observations

- The Random Forest classifier perfectly separated normal vs attack traffic.

- LIME explanations revealed that unusual connection states, abnormal byte/packet counts, and specific protocols/ports strongly influence attack predictions.
- These insights align with cybersecurity intuition: high traffic bursts or malformed connections are suspicious.

Conclusion

- The IoT Intrusion Detection task was successfully completed with a Random Forest model.
- **Explainability (LIME)** highlighted meaningful network indicators of attacks, supporting **trust and transparency** in intrusion detection systems.



Problem 2: COVID-19 Severity Prediction with LIME

Problem Statement

The task was to classify COVID-19 cases as **mild** or **severe** using patient symptoms, comorbidities, and demographic data. A Logistic Regression classifier was trained, and **LIME** was used to interpret which medical features drive predictions.

Steps Followed

1. **Data Loading**
   - Loaded the COVID-19 dataset (`covid_symptoms_severity_prediction.csv`).
2. **Preprocessing**
   - Constructed a **Severity label**:
     - Mild = not hospitalized, no ICU, no mortality.
     - Severe = hospitalized OR ICU OR mortality.
   - Encoded categorical fields (`gender`, `vaccination_status`).
   - Used symptom and comorbidity fields as features.
3. **Feature Extraction**

- Features included: age, gender, fever, cough, fatigue, comorbidities (diabetes, hypertension, cancer, etc.).

4. **Model Training**
   - Trained a **Logistic Regression** model.
   - Achieved **Accuracy ≈ 98.3%** on the test set.

5. **Explainability with LIME**
   - LIME-style surrogate model explained predictions for a severe case.
   - Most influential features increasing severity risk:
     - Diabetes
     - Cancer
     - Shortness of breath
     - Cough and fever
     - Cardiovascular comorbidities (heart disease, hypertension)

Observations

- Logistic Regression provided high accuracy while remaining interpretable.
- LIME explanations matched clinical knowledge: older age, respiratory distress, and comorbidities increase severity.
- Negative weights (protective factors) included absence of comorbidities or fewer symptoms.

Conclusion

- The COVID-19 Severity Prediction task was successfully completed with a Logistic Regression model.
- **Explainability (LIME)** revealed critical medical factors that determine severity, supporting **clinical decision-making** and **patient triage**.