Experiment 5

Aim: To apply explainable AI (XAI) methods (SHAP & LIME) for interpreting model predictions and evaluate fairness using Fairlearn.

Objective

- Generate **global** (SHAP feature importance) and **local** (LIME) explanations of model predictions.
- Audit bias in ML models across sensitive features (e.g., gender, race).
- Propose fairness mitigation strategies if bias is detected.

Detailed Steps

1. **Dataset & Model Selection**
    - Use dataset with potential fairness aspects (e.g., COMPAS, Adult Income, or synthetic dataset).
    - Train a classifier (Logistic Regression, Random Forest, XGBoost).
2. **Explainability with SHAP**
    - Install and apply SHAP (`shap.TreeExplainer` or `KernelExplainer`).
    - Generate **global explanations**: feature importance summary plots, dependence plots.
    - Interpret which features most influence predictions.
3. **Explainability with LIME**
    - Install and run LIME (`lime.lime_tabular`).
    - Generate **local explanations** for individual predictions.

- Visualize contribution of features for specific samples.

4. **Fairness Audit with Fairlearn**
    - Define sensitive attribute(s) (e.g., gender, race).
    - Use **Fairlearn's metrics** (demographic parity difference, equalized odds difference).
    - Generate a fairness report comparing performance across groups.
5. **Bias Mitigation**
    - If bias detected, propose strategies such as:
        - Pre-processing (reweighting data, sampling).
        - In-processing (fairness constraints in training).
        - Post-processing (threshold adjustments).

Open-Source Tools

- **SHAP** – Global feature importance.
- **LIME** – Local explanation of predictions.
- **Fairlearn** – Fairness metrics and bias mitigation.

Deliverables

- SHAP plots (summary, dependence).
- LIME local explanation visualizations.
- Fairness audit report (metrics, visualizations).

Conclusion