Report - Explainable AI (xAI) Dashboard for UCI Adult Dataset

1. Problem Understanding & Rationale

AI models, especially black-box types like XGBoost, achieve high accuracy yet lack interpretability. Understanding the rationale for a model's specific prediction is crucial for:

- Trust & Transparency: Users can comprehend the rationale behind choices.
- Equity & Ethics: Personal characteristics like age, gender, and ethnicity can be monitored for bias.
- Cybersecurity: Monitoring abuse or abnormal forecasts strengthens system robustness.
- Goal: Develop an interactive dashboard that demonstrates the reasoning for an AI model's predictions using SHAP explanations on the UCI Adult dataset.

2. Dataset Description

Dataset: UCI Adult Income Dataset

- **Samples:** 32,561
- **Features:** 14 input features + 1 target (income >50K or <=50K)
- **Feature Types:**
    - **Categorical:** workclass, education, marital-status, occupation, relationship, race, sex, native-country
    - **Numerical:** age, fnlwgt, education-num, capital-gain, capital-loss, hoursperweek

Target Variable:

- 0 → Income ≤ 50K
- 1 → Income > 50K

Notes on Indices:

Sample index in the dashboard (0,1,2,…) refers to the **row in the test set**. Selecting an index shows the model prediction and SHAP explanations for that specific sample.

3. Design & Implementation Approach

Models Trained:

1. Logistic Regression

2. Decision Tree

3. Random Forest

4. XGBoost (**Best-performing model**)

Evaluation Metrics:

Accuracy, ROC-AUC, F1-score

Model	Accuracy	ROC-AUC	F1-score
XGBoost	0.87	0.93	0.71
Random Forest	0.85	0.90	0.67
Logistic Regression	0.84	0.90	0.66
Decision Tree	0.84	0.89	0.62

SHAP (SHapley Additive exPlanations):

- Delivers overall feature significance for all test samples.
- Provides localized insights for specific predictions.
- Produces clear statements that demonstrate how features affect the likelihood of predictions rising or falling.
- Note: Explanations for Decision Tree and LIME are part of the code but not present in the deployed prototype, which emphasizes XGBoost and SHAP for ease and clarity.

Dashboard Features:

1. **Sample Selection:** Choose any test sample index to inspect its prediction.

2. **Model Prediction & Comparison:** Shows XGBoost prediction with probability.

3. **SHAP Explanations:**

   - Global feature importance plots.
   - Local explanation table with features, SHAP values, and impact.

4. **Trust & Safety:** Highlights sensitive features and their influence on predictions.

5. **Cybersecurity:** User login system with registration to demonstrate secure access. Used SQLite3 for the database.

4. Results & Observations

Example Predictions:

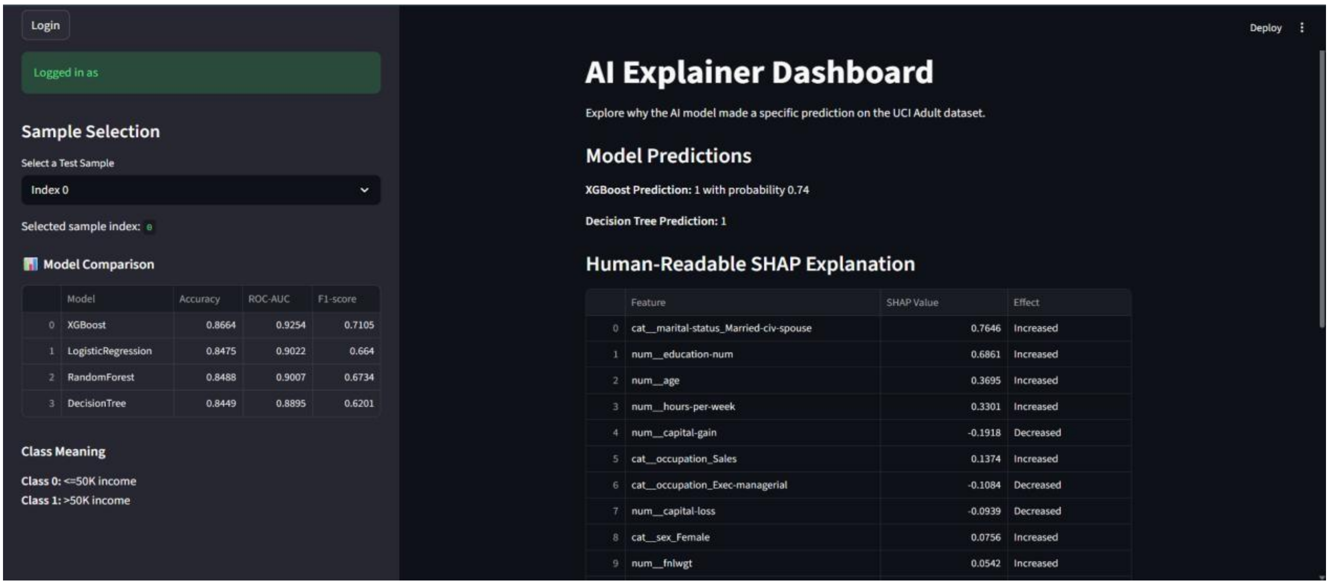
- **Sample Index 16:**
  - XGBoost Prediction: 0 (Income ≤50K, probability 0.99)
  - Key SHAP Influences:

Feature	SHAP Value	Impact
marital-status_Married-civ-spouse	-1.623	Decreased
age	-1.023	Decreased
education-num	0.585	Increased
sex_Female	-0.395	Decreased
hours-per-week	-0.236	Decreased
capital-gain	-0.180	Decreased
occupation_Sales	0.123	Increased

- **Interpretation:** Negative SHAP values decrease the probability of class 0, positive values increase the probability.

General Observations:

- Age and marital status significantly impact income predictions.
- Sensitive features are monitored to ensure fairness.
- Local explanations help users understand why the model made a certain decision.



5. Security, Ethical, & Governance Considerations

- **Sensitive Data:** Features like sex, race, and age are highlighted to detect potential bias.
- **User Authentication:** Only authorized users can access the dashboard (login system).
- **Explainability:** Using SHAP increases trust in AI decisions.
- **Audit Trails:** User actions and predictions are logged for accountability.

Ethical Implications:

- Transparent AI can prevent discriminatory outcomes.
- Users can validate model predictions before taking automated actions.

5. References

1. https://archive.ics.uci.edu/ml/datasets/adult
2. https://www.datacamp.com/tutorial/introduction-to-shap-values-machinelearninginterpretability
3. https://joblib.readthedocs.io/en/stable/