End-to-End ML Pipeline: Tabular Data Classification with Explainability

This project implements a complete machine learning pipeline for classifying structured/tabular data. It includes data cleaning, feature selection, model building, evaluation, and model interpretability using LIME and SHAP. The goal is to demonstrate best practices for building explainable and production-ready ML models.

Key Features

- EDA & Preprocessing: Handles missing values, encodes categorical data, and scales features.
- Feature Selection: Uses correlation and mutual information to identify key predictors.
- Modeling & Tuning: Trains multiple models (Random Forest, SVM, Logistic Regression) with cross-validation and hyperparameter tuning via GridSearchCV.
- Explainability: Applies LIME and SHAP for interpreting predictions and feature impact.
- Modular Design: Built using clean functions and Scikit-learn Pipelines for reusability.

Structure

- 1_EDA+preprocess.ipynb -> Exploratory analysis & data prep
- 2_FeatureSelection.ipynb -> Feature selection methods
- 3_model_eval+lime.ipynb -> Model evaluation and LIME visualizations
- Final_Combined_ML_Project.ipynb -> Final polished notebook with all enhancements

Technologies Used

- Python, Pandas, Scikit-learn, Matplotlib, Seaborn, LIME, SHAP

Results

- Achieved an F1-score of ~89% using Random Forest with tuned parameters
- LIME and SHAP visualizations provided clear explanations of individual predictions and global

feature importance

Future Improvements

- Streamlit/Gradio UI for live demo

- Deployment as a REST API

- Experimenting with AutoML or deep learning models