Mini Project: Loan Prediction using Data Engineering & Machine Learning

Problem Statement

This project aims to develop a **Loan Approval Prediction System** by combining **Data Engineering** and **Machine Learning** techniques.
The workflow involves:

1. Loading multiple `.json` files into a **MySQL database**.
2. Retrieving and preprocessing the data in **Python**.
3. Training a classification model to predict **loan approval status**.
4. Saving the model as a `.pickle` file.
5. Deploying the trained model as an interactive **Streamlit Web Application**.

The goal is to build a complete end-to-end pipeline, from **data storage** → **preprocessing** → **ML model** → **deployment**.

Dataset Description

The dataset is divided into **three JSON files**, representing different aspects of loan applications:

- `applicant_info.json` → Contains demographic details (age, gender, education).
- `financial_info.json` → Contains financial details (income, credit history, etc.).
- `loan_info.json` → Contains loan details (loan amount, term, status).

👉Dataset link : Data Engineering project Datasets

These datasets are combined in MySQL to form a **unified dataset** for training.

Project Requirements

1. Python Libraries

- **pandas, numpy** → Data handling & manipulation
- **scikit-learn** → Model training & evaluation
- **pymysql / sqlalchemy** → MySQL database connection
- **pickle** → Save and load trained models
- **streamlit** → Web application for deployment

2. Database

- **MySQL** for structured data storage.

3. Model Architecture

- Use **classification algorithms** (RandomForest, Logistic Regression, or Decision Trees).
- Include **preprocessing steps**: Encoding categorical variables, scaling numeric features.
- Save the trained model as `model.pkl` using pickle.

Project Steps

1. Data Engineering: Loading JSON into MySQL

Objective: Store semi-structured JSON data into a relational database.

Tasks:

- Load the JSON files (`applicant_info.json`, `financial_info.json`, `loan_info.json`) into Python.
- Create MySQL tables for each file.
- Insert JSON data into MySQL database (`loan_db`).

2. Data Retrieval & Preprocessing

Objective: Prepare data for training.

Tasks:

- Retrieve data from MySQL into Pandas DataFrames.
- Merge datasets into a single table using a common key (e.g., Applicant_ID).
- Handle missing values.
- Encode categorical features (Gender, Education, Marital Status, etc.).
- Scale numerical features if necessary.
- Split the dataset into **train/test sets**.

3. Model Training

Objective: Build a classification model for loan approval prediction.

Tasks:

- Train a **RandomForestClassifier** (or other suitable ML algorithm).
- Evaluate using accuracy, precision, recall, and F1-score.
- Save trained model as **`model.pkl`** using pickle.

4. Streamlit Application

Objective: Deploy a web app for interactive predictions.

Tasks:

- Build a **Streamlit app (`app.py`)** that allows users to input loan application details.
- Preprocess input data in the same way as training.
- Use `model.pkl` to predict loan approval status (`Approved` / `Rejected`).
- Display prediction result in a user-friendly interface.

5. Deployment

Objective: Host the app for real-world use.

Tasks:

- Push project to **GitHub**.
- Deploy Streamlit app on **Streamlit Cloud**.
- Provide project documentation in **README.md** with workflow diagram.

Expected Outcomes

1. **Data Engineering Skills**
    - Learn how to handle **JSON data** and load it into **MySQL**.
    - Understand database integration with Python.
2. **Machine Learning Development**
    - Preprocess structured data for ML.

- Train and evaluate a **classification model** for loan prediction.

3. **Model Deployment**
    - Save trained models with **pickle**.
    - Build and deploy a **Streamlit app** for real-world use.
4. **End-to-End Project Understanding**
    - Complete pipeline: **Data Engineering → ML Model → Deployment**.