Data Profiling Project

# Introduction

This project is a data profiling application built with Streamlit. It allows users to upload a CSV file, visualize data distributions, train a machine learning model, query customer details, generate reports, and provide feedback.

# Features

* **Data Summary**: Provides a summary of the uploaded data. It includes basic statistics such as mean, median, standard deviation, and counts for each column in the dataset. The summary includes information like the number of missing values, unique values, and data types for each column.
* **Data Visualizations**: Visualizes distributions for all numeric columns in the data. It generates histograms and other relevant plots to help users understand the data distribution. The visualizations include histograms, box plots, and scatter plots for numeric columns, and bar charts for categorical columns.
* **Train Model**: Allows users to train a machine learning model on the uploaded data. Users can select a target column and train a model to predict that target. The feature includes data preprocessing steps, model training, and evaluation metrics such as accuracy, precision, recall, and F1 score.
* **Query Customer Details**: Allows users to query customer details based on specific criteria. Users can select a column, an operator (e.g., ==, !=, >, <), and a value to filter the data. The feature supports various query operators and allows users to filter data based on multiple conditions.
* **Generate Report**: Generates a report with visualizations and model accuracy. It includes charts, summaries, and other relevant information about the data and the trained model. The report includes data summaries, visualizations, model performance metrics, and any insights derived from the data.
* **Feedback**: Allows users to provide feedback on their experience using the application. Users can rate their experience using smiley faces and submit their feedback. The feedback feature includes options for users to rate their experience and provide additional comments.

# Setup

1. Clone the repository:

git clone https://github.com/your-username/data-profiling-project.git  
cd data-profiling-project

2. Create a virtual environment and activate it:

python -m venv venv  
# On Windows  
venv\Scripts\activate  
# On macOS/Linux  
source venv/bin/activate

3. Install the required dependencies:

pip install -r requirements.txt

4. Run the Streamlit application:

streamlit run src/app.py

# Usage

1. Open the Streamlit application in your web browser.
2. Upload a CSV file using the file uploader.
3. Navigate through the tabs to explore the features of the application.
4. Use the 'Query Customer Details' tab to query customer details based on specific criteria.
5. Generate a report with visualizations and model accuracy using the 'Generate Report' tab.
6. Provide feedback on your experience using the 'Feedback' tab.

# Example CSV File

Here is an example of the CSV file format:  
  
Customer\_ID,Account\_Balance,Transaction\_Amount,Reported\_Amount,Currency,Country,Transaction\_Date,Risk\_Score  
1001,15000,500,500,USD,US,2/25/2025,3  
1002,32000,1200,1200,EUR,DE,2/20/2025,2  
1003,-5000,10300,300,GBP,UK,2/18/2025,6  
1004,70000,2000,1800,USD,US,2/28/2025,5

# License

This project is licensed under the MIT License. See the LICENSE file for details.