User Guide for Dynamic De-Identification Application

Your Company Name

October 13, 2024

Contents

1 Introduction

Welcome to the **Dynamic De-Identification** application user guide. This application is designed to assist users in processing, anonymizing, and generating synthetic data to ensure data privacy and compliance with various privacy models. The application leverages powerful Python libraries and Streamlit for an interactive user experience.

2 Prerequisites

Before using the application, ensure that you have the following installed on your system:

- **Python 3.7 or higher**
- **pip** (Python package installer)
- **Virtual Environment (optional but recommended)**

2.1 Installing Python

Download and install Python from the official website: https://www.python.org/downloads/.

2.2 Setting Up a Virtual Environment (Optional)

Creating a virtual environment helps manage dependencies and avoid conflicts.

```
python -m venv venv
source venv/bin/activate  # On Windows: venv\Scripts\activate
```

3 Installation

Follow these steps to install and set up the Dynamic De-Identification application.

3.1 Clone the Repository

Clone the application repository from GitHub (replace the URL with your repository):

```
git clone https://github.com/yourusername/dynamic-deidentification.git
cd dynamic-deidentification
```

3.2 Install Dependencies

Install the required Python packages using pip:

```
pip install -r requirements.txt
```

Note: Ensure that the ‘requirements.txt‘ file includes all necessary packages such as Streamlit, pandas, matplotlib, etc.

4 Running the Application

To launch the Streamlit application, navigate to the project directory and execute:

```
streamlit run application.py
```

This command will start the application and open it in your default web browser.

5 Application Overview

The Dynamic De-Identification application is organized into several tabs, each serving a specific function in the data processing and anonymization workflow. Below is an overview of each tab and its functionalities.

5.1 Main Components

- **Sidebar:** Upload datasets, configure settings, and access application information.
- **Tabs:** Navigate through different functionalities such as Binning, Location Granularizer, Unique Identification Analysis, Data Anonymization, Synthetic Data Generation, and Help.
- **Logs:** Optionally display application logs for monitoring processes.

6 Using the Application

Follow the steps below to effectively use each feature of the application.

6.1 1. Uploading Data

1. Navigate to the **Sidebar** on the left.
2. Under **Upload & Settings**, click on **Upload your dataset**.
3. Select a CSV (`.csv`) or Pickle (`.pkl`) file from your local machine.
4. Choose the desired output file type (`csv` or `pkl`) from the dropdown menu.

6.2 2. Configuring Data Processing Settings

1. In the **Sidebar**, locate the **Data Processing Settings** section.
2. Adjust the following parameters as needed:
    - **Date Detection Threshold**: Sets the sensitivity for detecting date columns.
    - **Numeric Detection Threshold**: Determines the threshold for identifying numeric columns.
    - **Factor Threshold Ratio**: Ratio used in factor detection.

- **Factor Threshold Unique**: Unique value threshold for factor columns.
- **Day First in Dates**: Checkbox to specify date format.
- **Convert Factors to Integers**: Checkbox to convert factor columns.
- **Date Format**: Specify the date format if known (e.g., `%Y-%m-%d`).

6.3 3. Manual Binning

1. Click on the **Manual Binning** tab.
2. Under **Select Columns to Bin**, choose the columns you wish to bin from the available options.
3. The application will display binning configurations based on your selection.
4. Toggle the **Start Dynamic Binning** checkbox to initiate the binning process.
5. Upon completion, you can:
   - Run an integrity report to assess data integrity post-binning.
   - Perform association rule mining with configurable support and confidence thresholds.
   - Generate density plots for visual analysis.
   - Download the binned data for external use.

6.4 4. Location Data Geocoding Granularizer

1. Navigate to the **Location Data Geocoding Granulariser** tab.
2. The application will automatically detect geographical columns.
3. Select the column(s) you want to geocode.
4. Click on **Start Geocoding** to initiate the geocoding process.
5. After geocoding:
   - Choose the desired granularity level (e.g., address, city, state).
   - Generate granular location columns based on the selected granularity.
   - Optionally, visualize the geocoded data on a map.

6.5 5. Unique Identification Analysis

1. Select the **Unique Identification Analysis** tab.
2. The application will display selected columns from previous steps (Binning and Location Granularizer).
3. Configure the analysis by specifying:

- **Minimum Combination Size**: The smallest number of columns to consider in combinations.
- **Maximum Combination Size**: The largest number of columns to consider.

4. Click on **Perform Unique Identification Analysis** to execute.
5. Review the results, which include:
   - Unique identification metrics.
   - Integrity loss reports.
   - Density distribution plots.
   - Option to download the analysis results.

6.6 6. Data Anonymization

1. Access the **Data Anonymization** tab.
2. Configure anonymization settings:
   - Select the desired privacy model (**k-anonymity**, **l-diversity**, or **t-closeness**).
   - Specify parameters such as:
     - **k**: The anonymity level.
     - **l**: The diversity level (for l-diversity).
     - **t**: The closeness threshold (for t-closeness).
   - Select sensitive attributes if applicable.
3. Configure binning settings:
   - Choose columns to bin.
   - Define minimum and maximum bins per column.
   - Select the binning method (**quantile** or **equal width**).
   - Choose the optimization method (**genetic algorithm** or **simulated annealing**).
   - Set optimizer-specific hyperparameters.
4. Click on **Optimize Binning** to start the anonymization process.
5. Upon completion, review:
   - Best binning configuration.
   - Binned data samples.
   - Optimization summaries and plots.
   - Privacy compliance visualizations.
   - Options to download the anonymized data and configurations.

6.7 7. Synthetic Data Generation

1. Select the **Synthetic Data Generation** tab.
2. Choose the columns to include in the synthetic data generation process.
3. The application will automatically detect and display data types (Datetime, Categorical, Numerical).
4. Optionally, adjust column data types as needed.
5. Handle missing values by selecting an appropriate strategy:
   - Drop rows with missing values.
   - Mean, median, or mode imputation.
   - Fill with a specific value.
6. Select the synthetic data generation method (**CTGAN** or **Gaussian Copula**).
7. Configure model parameters based on the chosen method.
8. Specify the number of synthetic samples to generate.
9. Click on **Generate Synthetic Data** to start the process.
10. After generation, review:
    - Synthetic data samples.
    - Option to download the synthetic dataset.
    - Comparative distribution plots between original and synthetic data.

7 Help & Documentation

Access the **Help & Documentation** tab for additional guidance and best practices.

7.1 Using the Help Tab

- **How to Use This Application:** Step-by-step instructions similar to this guide.
- **Understanding the Settings:** Detailed explanations of binning methods and anonymization models.
- **Best Practices:** Recommendations for effective data anonymization and synthetic data generation.
- **Troubleshooting:** Common issues and their resolutions.

8 Advanced Features

8.1 Session State Information

- The application maintains a **Session State** to track user interactions and data processing steps.
- Access detailed session information through the **Session State Info** expander in the sidebar.
- Monitor session state logs and variable types for debugging and transparency.

8.2 Logging

- Application logs are stored in the `logs/app.log` file.
- Optionally display logs within the application interface by enabling the **Show Logs in Interface** checkbox in the sidebar.
- Logs provide insights into the application's operations and help in troubleshooting errors.

9 Data Management

9.1 Data Loading and Saving

- Uploaded datasets are saved in the `data/` directory.
- Processed data is stored in the `processed_data/` directory.
- Reports and analysis results are saved in the `reports/` directory.
- Logs are maintained in the `logs/` directory.

9.2 Downloading Results

- After each processing step, use the provided download buttons to export results in CSV or Pickle formats.
- Download options are available for:
  - Binned data.
  - Binning configurations.
  - Integrity reports.
  - Unique identification analysis results.
  - Anonymized datasets.
  - Synthetic datasets.

10 Best Practices

- **Start with Default Settings:** Familiarize yourself with the application's workflow before customizing settings.
- **Validate Anonymization:** Always review integrity loss reports and privacy compliance visualizations to ensure data privacy.
- **Backup Original Data:** Maintain a backup of your original datasets before performing any processing or anonymization.
- **Monitor Logs:** Regularly check application logs for any errors or warnings during data processing.
- **Optimize Parameters:** Experiment with different binning and anonymization parameters to achieve the best balance between data utility and privacy.

11 Troubleshooting

11.1 Common Issues

- **Unsupported File Type:** Ensure that you upload data in CSV (`.csv`) or Pickle (`.pkl`) formats.
- **Missing Columns:** Verify that selected columns exist in both original and processed datasets.
- **Privacy Not Achieved:** Adjust anonymization parameters or select different sensitive attributes to meet privacy requirements.
- **Model Training Errors:** Check data types and handle missing values appropriately before generating synthetic data.

11.2 Accessing Logs

- Enable the **Show Logs in Interface** option to view real-time logs.
- Alternatively, access the `logs/app.log` file for detailed logs.

12 Conclusion

The Dynamic De-Identification application provides a robust framework for data anonymization and synthetic data generation, ensuring data privacy while maintaining data utility. By following this guide, users can effectively leverage the application's features to process and protect their datasets.