DATASHIELD – Implementing (*k, m, t*)-Anonymity for Transactional Datasets PROJECT TUTORIAL

OF

DATABASE AND ONLINE SOCIAL MEDIA SECURITY (CSLM 654)

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

Submitted By

ARWAZ KHAN (242210005), ARYAM SHRIVASTAVA (242210006) & JOSHUA JOY (242211009)

Submitted To

DR. SHELLY SACHDEVA (ASSOCIATE PROFESSOR, DoCSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY DELHI MAY 2025

Project Tutorial

1. Set up IDE and Project Environment

- Download & Install VS Code https://code.visualstudio.com/download
- Download & Install Python https://www.python.org/downloads/
- Download Code Repository https://github.com/arwazkhan189/Datashield
- Install Python libraries in the command prompt
  pip install flask, pandas, numpy, seaborn, faker, tqdm

2. Create the file structure

- app.py
- anonymized_output.py
- anonymizer.py
- compareDatasets.ipynb
- dataset_generator.py
- requirements.txt
- static / styles.css
- static/ script.js
- templates / index.html
- datasets / synthetic_healthcare_dataset.csv

3. Workflow of Project

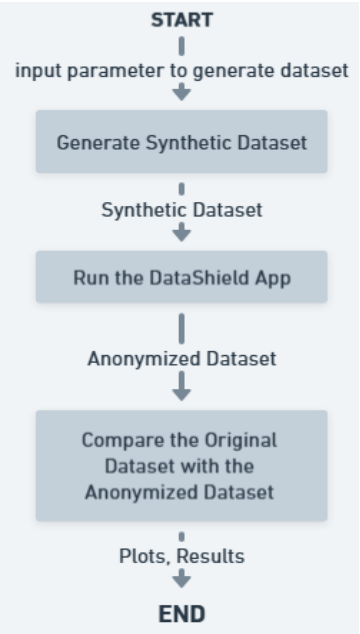

Fig 3.1: Workflow of the project

4. Steps to generate the synthetic dataset using faker library

Step 1: Modify the *dataset_generator.py* code according to your need

```python
import pandas as pd
import random
from faker import Faker
from tqdm import tqdm

# Initialize Faker
fake = Faker()
Faker.seed(42)
random.seed(42)

# Configuration
num_records = 500_000

# Sample data pools
genders = ['Male', 'Female', 'Other']
diseases = [
    'Diabetes', 'Hypertension', 'Asthma', 'Cancer', 'Arthritis',
    'Flu', 'Migraine', 'COVID-19', 'Tuberculosis', 'Heart Disease'
]
medications = [
    'Paracetamol', 'Ibuprofen', 'Metformin', 'Amlodipine', 'Lisinopril',
    'Omeprazole', 'Azithromycin', 'Prednisone', 'Atorvastatin', 'Insulin'
]
hospitals = [
    'City Hospital', 'Green Valley Medical Center', 'Sunrise Clinic',
    'Metro Health Institute', 'Apollo Medicals', 'National Care Center'
]
```

Fig 4.1: dataset_generator.py

Step 2: Run the code to generate the dataset.

	Name	Age	Gender	Pincode	Disease	Medication	Visit_Date	Doctor_Name	Hospital_Name
2	Allison Hill	82	Male	29757	Diabetes	Lisinopril	18-08-2023	Megan Mcclain	Green Valley Medical Center
3	Javier Johnson	29	Male	29158	Hypertension	Atorvastatin	29-10-2023	Alyssa Gonzalez	City Hospital
4	Kimberly Robinson	76	Female	77737	Diabetes	Paracetamol	04-12-2023	Abigail Shaffer	City Hospital
5	Gina Moore	28	Male	44619	Tuberculosis	Insulin	20-07-2023	Brent Abbott	City Hospital
6	Renee Blair	72	Male	70785	Tuberculosis	Azithromycin	06-08-2023	Jamie Arnold	Green Valley Medical Center
7	Lisa Hensley	58	Other	76175	Arthritis	Paracetamol	28-09-2023	Amber Perez	Green Valley Medical Center
8	Bobby Hall	90	Female	13739	Flu	Lisinopril	21-02-2024	Mark Diaz	Green Valley Medical Center
9	Daniel Adams	28	Female	90094	Hypertension	Ibuprofen	16-09-2024	Mark Ferguson	Metro Health Institute
10	Joel Nelson	13	Female	84387	Flu	Insulin	17-10-2024	Melinda Cameron	Sunrise Clinic
11	Crystal Johnson	6	Other	41848	COVID-19	Atorvastatin	08-03-2024	Daniel Hahn	City Hospital
12	Emily Rios	49	Male	52357	Tuberculosis	Lisinopril	19-03-2025	Judy Baker	National Care Center
13	Justin Baker	80	Female	77123	Heart Disease	Amlodipine	29-03-2024	Jennifer Robinson	National Care Center
14	Ms. Ann Williams MD	9	Male	82741	Cancer	Lisinopril	03-09-2023	Jennifer Brown	City Hospital
15	Zachary Rice	30	Male	73013	Migraine	Lisinopril	28-02-2025	Melanie Wilson	Metro Health Institute
16	Nicole Mack	82	Female	45088	Asthma	Omeprazole	27-07-2023	Christopher Smith	Sunrise Clinic
17	Michelle Stanton	27	Other	23917	Arthritis	Ibuprofen	30-05-2024	Sheila Evans	Apollo Medicals
18	Lisa Hernandez	82	Male	21675	Tuberculosis	Amlodipine	23-06-2024	Tammy Sellers	Green Valley Medical Center
19	Katherine Rodriguez	60	Female	48077	Arthritis	Atorvastatin	14-03-2025	Dr. Cynthia Allen	Green Valley Medical Center
20	Angela Dennis	88	Female	9572	Diabetes	Amlodipine	12-12-2024	Beth Keller	City Hospital
21	Carmen Rose	41	Female	80008	Arthritis	Ibuprofen	25-03-2024	Tanya Campos	Green Valley Medical Center
22	Michelle Ross	73	Other	88540	Flu	Amlodipine	17-09-2024	Steven Hayes	National Care Center
23	Austin Smith	64	Female	73104	COVID-19	Metformin	27-10-2023	Adrienne Zimmerm	Sunrise Clinic
24	Austin Johnson	18	Male	43810	Tuberculosis	Atorvastatin	26-06-2023	Diana Washington	Sunrise Clinic
25	Miranda Khan	96	Other	76026	Migraine	Insulin	03-05-2024	John Russell	Metro Health Institute
26	Matthew Gomez	47	Male	54384	Asthma	Atorvastatin	28-04-2024	Amy Valdez	Metro Health Institute
27	Amy Chandler	12	Male	14823	Hypertension	Metformin	11-11-2023	Joshua Taylor	National Care Center
28	Joel Baxter	21	Other	10381	Migraine	Insulin	10-04-2024	Savannah Garcia	City Hospital
29	Kimberly Smith	50	Female	12725	Heart Disease	Prednisone	20-04-2025	Cynthia Russell	Apollo Medicals
30	Dr. Steven Martin	33	Other	783	Diabetes	Ibuprofen	29-02-2024	Richard Gibson	National Care Center

Fig 4.2: Generated synthetic patients healthcare dataset

5. Steps to start the project

Step 1: Download the ZIP file from the following link and extract its contents: https://github.com/arwazkhan189/Datashield

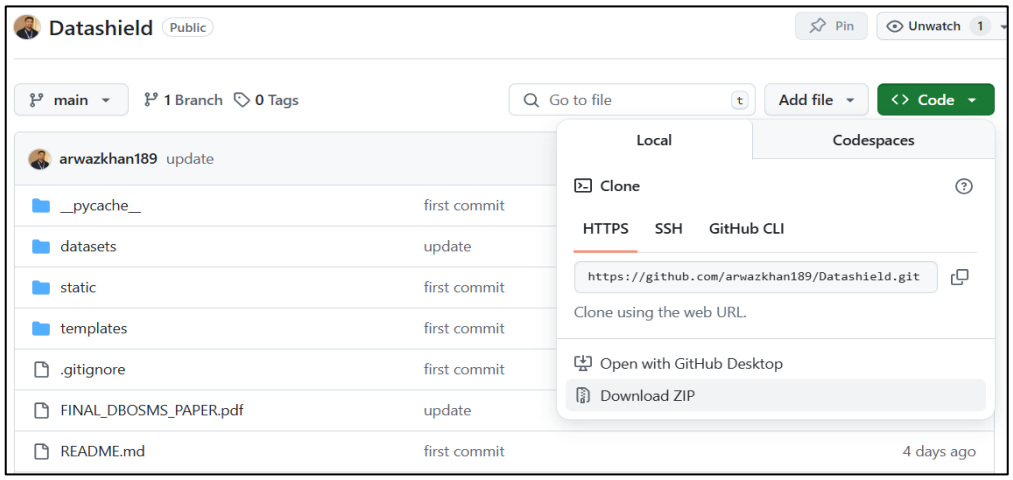

Fig 5.1: GitHub Repository Page

Step 2: Launch VS Code and open the extracted project folder.

Step 3: Open the terminal in VS Code and run the application using the command:

py app.py

Step 4: Once the server starts, open the localhost URL displayed in the terminal in your web browser.

Step 5: The web application will now be displayed in your browser.

Fig 5.2: Web Application Interface

Step 6: Choose a sample dataset on which you want to apply k-anonymity.

1	Name	Age	Gender	Pincode	Disease
2	Alice	29	Female	560001	Flu
3	Bob	35	Male	560002	Cold
4	Carol	42	Female	560003	Diabetes
5	David	33	Male	560004	Asthma
6	Eve	27	Female	560005	Flu
7	Frank	30	Male	560001	Cancer
8	Grace	31	Female	560002	Cold
9	Hank	28	Male	560003	Diabetes
10	Ivy	36	Female	560004	Flu

Fig 5.3: Sample Dataset Used for (k, m, t) –Anonymity

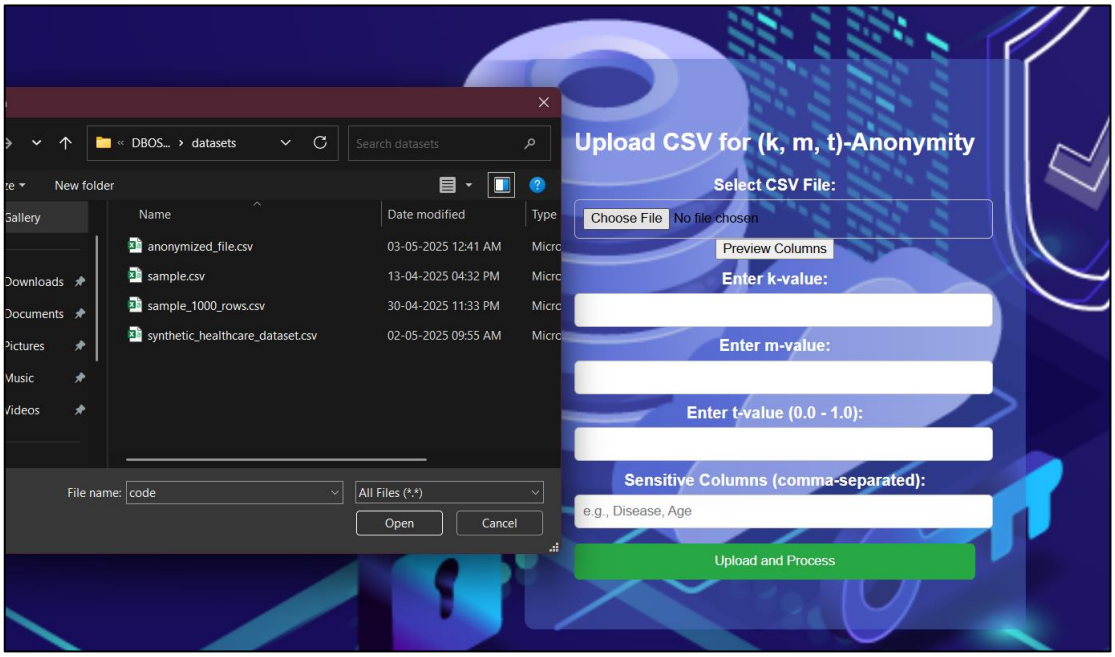

Fig 5.4: Selecting the Sample Dataset for Processing

Step 7: Specify the (k, m, t)-value to define the level of anonymity.

Step 8: Provide the column names, separated by commas, that should be considered for anonymization.

Fig 5.5: Defining the (k, m, t)-Value and Specifying Column Names for Anonymization

Step 9: Click on the "Upload and Process" button to process the dataset and download the anonymized output.

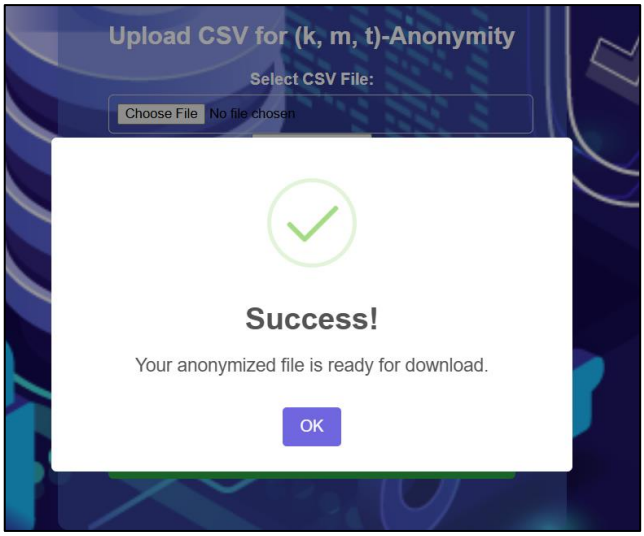

Fig 5.6: Processed Dataset Downloaded Successfully

Step 10: Navigate to the Downloads folder and open the file named anonymized_file.csv to view the anonymized dataset.

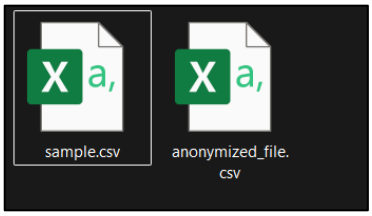

Fig 5.7: Downloaded Anonymized Dataset

1	name	age	gender	pincode	disease
2	***	29	Female	***	Flu
3	***	35	Male	***	Cold
4	***	42	Female	***	Diabetes
5	***	33	Male	***	Asthma
6	***	27	Female	***	Flu
7	***	30	Male	***	Cancer
8	***	31	Female	***	Cold
9	***	28	Male	***	Diabetes
10	***	36	Female	***	Flu
11	***	40	Male	***	Asthma

Fig 5.8: View of the Anonymized Dataset in CSV Format

6. Compare anonymized dataset with the generated synthetic dataset

Step 1: Use *compareDatasets.ipynb* code to compare both the datasets and generate plots and results.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np


# Load datasets
original_df = pd.read_csv("datasets/synthetic_healthcare_dataset.csv")
anonymized_df = pd.read_csv("datasets/anonymized_file.csv")
```

Fig 6.1: compareDatasets.ipynb code

Step 2: Include path of both original and anonymized dataset and run the code to get the results and plots.

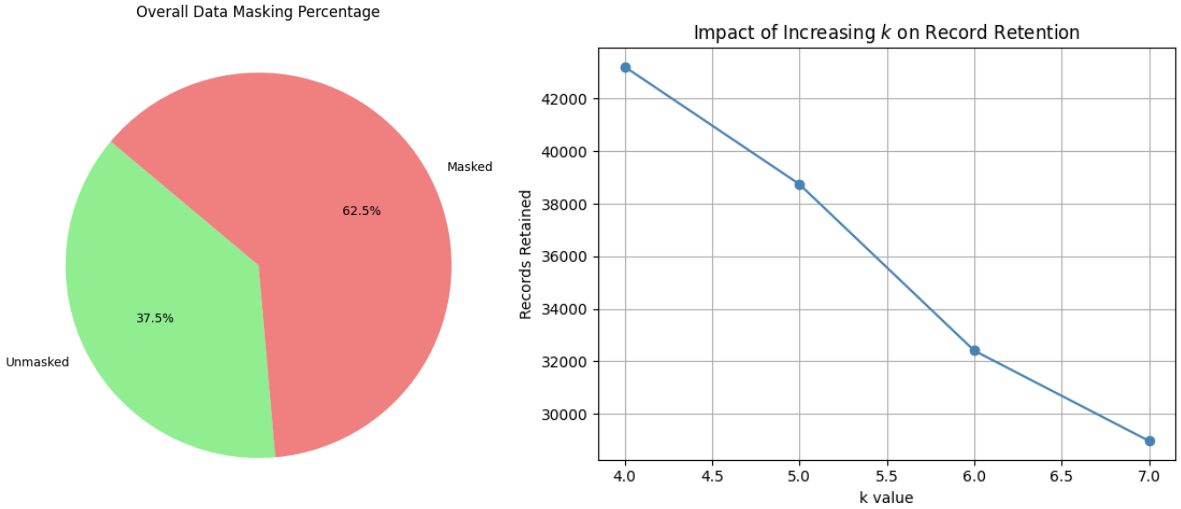

Fig 6.2: Plots of comparison of datasets

References

[1] Puri, Vartika, Parmeet Kaur, and Shelly Sachdeva. “(k, m, t)-anonymity: Enhanced privacy for transactional data.” Concurrency and Computation: Practice and Experience 34.18 (2022): e7020.

[2] L. Sweeney, “k-anonymity: A model for protecting privacy,” Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[3] G´erard, J. (2014). Faker: Python package for generating fake data. https://faker.readthedocs.io/en/master/

[4] Walonoski, Jason, et al. “Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record.” Journal of the American Medical Informatics Association 25.3 (2018): 230-238.

[5] BMS-WebView1 dataset. Available at: http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

[6] INFORMS Data Mining Challenge dataset. Available at: https://sites.google.com/site/informsdataminingcontest/data

[7] Code Reference: https://github.com/arwazkhan189/Datashield