**Lesson 07 Demo 05**

**Generating Synthetic and Augmented Datasets Using   
Generative AI Tools**

**Objective:** To generate synthetic and augmented datasets using generative AI tools for validating data integrity, enhancing test coverage, and ensuring robustness during the testing phase of the Software Development Life Cycle

**Tools required:** Mostly AI and ChatGPT 4

**Prerequisites:** Basic knowledge and difference of synthetic and augmented dataset

Steps to be followed:

1. Generate the synthetic dataset using Mostly AI
2. Generate the synthetic and augmented dataset using ChatGPT

|  |
| --- |
| **Note:** Please note that all the GenAI tools used in this exercise can produce varied outputs even when presented with similar prompts. Thus, you may get different outputs for the same prompt. |

|  |
| --- |
| **Note**: Mostly AI may offer some customization options, primarily focusing on generating realistic synthetic data from scratch rather than augmenting existing datasets.  ChatGPT 4, on the other hand, is a versatile tool that can be used to generate synthetic data and manipulate and augment existing datasets effectively. |

**Step 1:** **Generate the synthetic dataset using Mostly AI**

* 1. Navigate to **https://mostly.ai/**

A screenshot of a computer

Description automatically generated

* 1. Click on the **Log in** button and enter the credentials  
       
     A screenshot of a computer

     Description automatically generated

A screenshot of a computer

Description automatically generated

|  |
| --- |
| **Note**: Create a new account by clicking on the **Sign up here** button if you do not have an account |

* 1. Click on the **Upload file** button  
       
     A screenshot of a computer

     Description automatically generated

A screenshot of a computer

Description automatically generated

|  |
| --- |
| **Note**: You may use any sample data set; if you do not have any data set, use ChatGPT to create a demo dataset of the desired number of rows and columns. Use the following prompt: **Create a dataset of 100 rows for a list of mobile phones, including year of launch, model, price, country, camera configuration, and discount offer** |

* 1. Click on the **Proceed** button  
       
     A screenshot of a computer

     Description automatically generated
  2. Click on the **Configure models** button, as shown in the screenshot below:  
       
     A screenshot of a computer

     Description automatically generated
  3. Click on the **Mobile\_Phone\_Dataset** list and make the desired changes accordingly, as shown in the screenshots below:  
       
     A screenshot of a computer

     Description automatically generated  
       
       
       
       
       
       
       
       
     A screenshot of a computer

     Description automatically generated

|  |
| --- |
| **Note**: Customize the dataset parameters such as column names and data distribution to match specific requirements |

* 1. Click on the **Start training** button  
       
     A screenshot of a computer

     Description automatically generated
  2. Scroll down to the **Training status** section  
       
     A screenshot of a computer

     Description automatically generated
  3. Examine the generated synthetic data in the **Model samples** section to ensure it meets your requirements

A screenshot of a computer

Description automatically generated

|  |
| --- |
| **Note**: You may click **Resample** if the sample data does not meet your requirements. |

* 1. Click on the **Generate synthetic data** button  
       
     A screenshot of a computer

     Description automatically generated
  2. Click on the **Start generation** button  
       
     A screenshot of a computer

     Description automatically generated  
       
     You will see the following interface:  
       
       
       
       
       
       
       
       
       
       
     A screenshot of a computer

     Description automatically generated
  3. Click on the **Download synthetic dataset** dropdown and select the desired option to download or directly export the synthetic dataset, as shown in the screenshots below:  
       
     A screenshot of a computer

     Description automatically generated  
       
     A screenshot of a computer

     Description automatically generated

**Step 2:** **Generate the synthetic and augmented dataset using ChatGPT**

* 1. Navigate to the **https://chat.openai.com** website and log in to your account  
       
     **A screenshot of a computer

     Description automatically generated**

|  |
| --- |
| **Note:** Sign up if you do not have an account |

A screenshot of a computer

Description automatically generated

* 1. Generate the synthetic data using the following prompt:  
     **Create synthetic data including the following columns:**

**Card Number**

**Cardholder name**

**Expiry date**

**CVV**

**Amount**

**Currency**

**Generate it to 1000 rows,**

**Provide a downloadable file**  
A screenshot of a computer

Description automatically generated

* 1. Click on the **Download Synthetic Credit Card Data** link, as shown in the screenshot below:  
       
     A screenshot of a computer

     Description automatically generated  
       
     The generated synthetic data is as shown below:  
       
     A screenshot of a computer

     Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. Upload the synthetic dataset and generate the augmented dataset using the following prompt:  
     **Generate augmented data based on that dataset with 2000 rows**  
       
     A screenshot of a computer

     Description automatically generated
  2. Click on the **Download Augmented Synthetic Credit Card Data** link  
       
     A screenshot of a computer

     Description automatically generated  
       
     The generated augmented data is as shown below:  
       
     A screenshot of a computer

     Description automatically generated

By following these steps, you have successfully generated synthetic and augmented datasets using generative AI tools for validating data integrity, enhancing test coverage, and ensuring robustness during the testing phase of the Software Development Life Cycle.