

Hands-on Lab: Generative AI for Data Generation and Augmentation

Estimated time needed: **30** minutes

One of the principle advantages of generative AI is its ability to generate realistic synthetic data. The synthetic data is generated when a pretrained generative model responds to either a prompt, create new data samples, or transfers learns on a given data set. In addition, it creates samples that can augment the existing data set while maintaining the statistical distribution and interpretability of the data set.

In this lab, you will learn how to use generative AI to generate synthetic data samples and transfer learns on a given data set.

Learning Objective

In this lab, you will learn how to use a popular tool, [Mostly.ai](#), to create synthetic data samples to augment a CSV data set.

Data Set

You will use a data set that includes insurance records.

The data set is available at the following link:

[Insurance Dataset](#)

This data set is a cleaned-up version of the [Medical Insurance Price Prediction](#) data set, available under the [CC0 1.0 Universal License](#) on the [Kaggle](#) website.

Steps

1. Download the data set

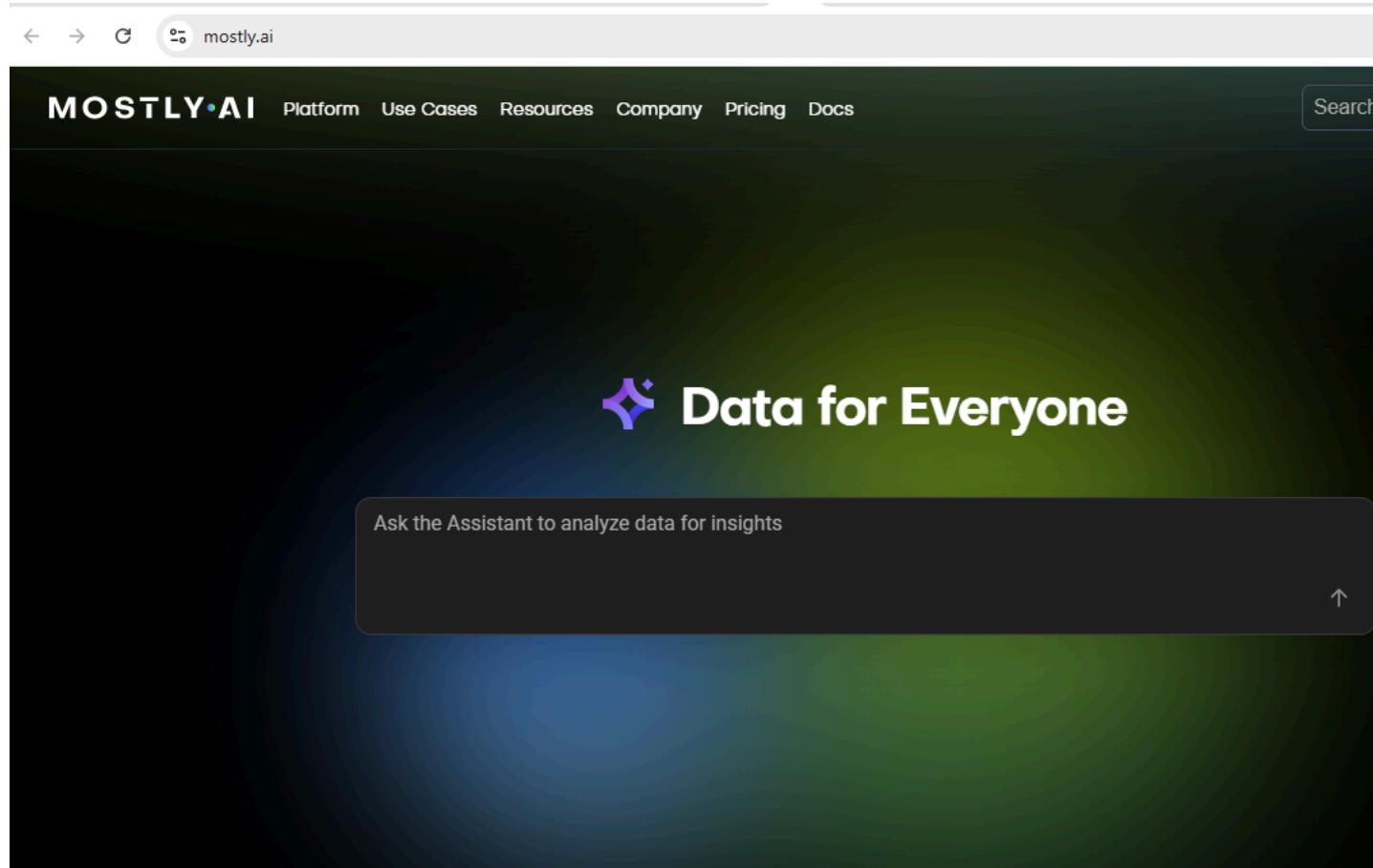
The first step is to download the dataset on your machine. You will need to upload this file to the interface in a subsequent step. Click the link provided in the **Data Set** section to download the data set.

2. Open the website

Click the following link to open the mostly.ai website and interface.

<https://mostly.ai/>

This link opens in a new browser tab, and you should see an web page that looks similar to the following screen capture:



3. Create an account

You can create an account on this website free of charge, or you can simply log in using your Gmail ID. After you log in, you'll see the following interface.

The screenshot shows the Mostly AI web interface. On the left sidebar, there are several options: '+ New chat', 'Search chats', 'Datasets', 'Connectors', 'Generators' (which is highlighted with a red box), and 'Synthetic datasets'. The main area features a large green gradient background with the text 'Data for Everyone' and a purple star icon. Below this is a search bar containing 'Ask the Assistant to ana' and a button labeled 'Ask'. At the bottom of the sidebar, there are four buttons: 'Start with a file', 'Connect your data', 'Explore a dataset', and 'Analyze data'. A URL 'https://app.mostly.ai/d/home' is visible at the top of the browser window.

4. Upload the data set

- Click on the **Generators** given on the left hand side of the page.

The screenshot shows the 'Generators' page from the Mostly AI interface. The left sidebar includes '+ New chat', 'Search chats', 'Datasets', 'Connectors', 'Generators' (highlighted with a red box), and 'Synthetic datasets'. The main area has a green gradient background with the text 'Data for Everyone' and a purple star icon. Below it is a search bar with 'Ask the Assistant to g' and a 'Ask' button. At the bottom of the sidebar are 'Start with a file', 'Connect your data', 'Explore a dataset', and 'Analyze data' buttons. A URL 'https://app.mostly.ai/d/generators' is at the bottom left. A section titled 'Popular datasets' is visible at the bottom of the main content area.

- And upload the CSV file of the data set to the interface by using the `Upload your data` option available on the console.

- + New chat
- Q Search chats
-
- ↳ Datasets
- ↳ Connectors
- ↳ Generators**
-
- ↳ Synthetic datasets
-
- Organizations
- + New organization
-
- Chats
- New chat
- New chat

Generators

① Generators are models that learn from original data. Once trained, they allow to create any number of synthetic samples as well as new data points.

Train a generator

On platform

Train a generator with your data on platform.

Start from a connector
Upload your data

Locally

Train a generator on your environment and then upload it to the platform.

Use the SDK
Import a generator

Available generators

Name	Visibility	Status	Activity
drxtn93742/insurance_dataset (1)	Private	Ready	0 1

- After you upload the data set, you will see its filename on the console. Then select Proceed as seen in the following screen captures:

The screenshot shows the Mostly AI web interface. On the left, there's a sidebar with navigation links like 'New chat', 'Search chats', 'Datasets', 'Connectors', 'Generators' (which is selected), 'Synthetic datasets', 'Organizations', 'Chats', 'New chat', and 'New chat'. The main area has a title 'Generators' and a sub-section 'Train a generator'. It shows two options: 'On platform' (selected) and 'Locally'. Under 'On platform', there are buttons for 'Start from a connector' and 'Upload your data' (the latter is highlighted with a red box). Below this is a section titled 'Available generators' with a table showing one entry: 'drxtn93742/insurance_dataset (1)' with status 'Ready'. A modal window is open in the foreground, titled 'Add table' (Step 1/2). It has a 'Upload file' section with a cloud icon and the text 'Drag a file here or click to browse. CSV, TSV, and Parquet files are supported.' Below this is a 'Table name' input field containing 'insurance_dataset'. Underneath the table name is a file preview for 'insurance_dataset.csv' with a delete icon. At the bottom right of the modal is a large blue 'Proceed' button, which is also highlighted with a red box.

5. Data configuration settings

- You can choose to modify the category of an attribute, or you can choose to include a parameter in the augmentation process without these settings. For the purposes of this lab, do not change these settings.

The screenshot shows the Mostly AI interface. On the left sidebar, under the 'Generators' section, there is a red box around the 'insurance_dataset' entry. The main area is titled 'Step 1/2 Data configuration'. It displays the 'insurance_dataset' with a table type of 'Subject' and 1338 rows. A blue 'Add table' button is located on the right side of the dataset card.

- Simply select `Configure models` to go to the model configuration settings.

The screenshot shows the Mostly AI interface in the 'Data configuration' step. The 'insurance_dataset' is selected. The 'Table columns' section is expanded, showing seven columns: 'age', 'gender', 'bmi', 'children', 'smoker', 'region', and 'expenses'. Each column has an 'Include' toggle switch (all are turned on) and an 'Encoding type' dropdown set to 'Tabular/Numeric: Auto'. A 'Primary key' icon is also present in this section.

6. Model configuration settings

You can modify the max training time, number of epochs, sample size, and other settings to generate the best possible model based on your requirements. For the purpose of this lab, use the default settings.

The screenshot shows the MOSTLY.AI web interface. On the left sidebar, under the 'Generators' section, the 'insurance_dataset' item is highlighted with a red box. The main workspace is titled 'Step 2/2 Model configuration'. At the top, it shows 'drxtn93742 / insurance_dataset' and a 'Private' button. Below that, the dataset name 'insurance_dataset' is listed as 'tabular'. To the right, there are buttons for 'Subject table', '1,338 rows', and '10 min'. The overall interface has a clean, modern design with light blue and white colors.

When you complete working with the settings, select `Start training`. You will find this option on the top right corner of the web page.

This screenshot continues from the previous one, showing the 'Model configuration' step. The 'insurance_dataset' is still selected as 'tabular'. The configuration section includes:

- Model**: A dropdown menu set to 'MOSTLY_AI/Medium'.
- Compute**: A dropdown menu set to 'CPU Intel Xeon Spot: 14 CPUs, 26GB'.
- Training parameters**: A section with a note: 'Adjust the training parameters to prioritize speed over accuracy, or vice versa.' It includes 'Max training time' (set to 10 mins) and 'Max sample size' (set to 1,338).

7. Model training

After the model training completes, you will see an onscreen result similar to what you see on the following screen capture.

+ New chat Q Search chats

Datasets Connectors Generators Synthetic datasets

Organizations + New organization

Chats New chat New chat

insurance_dataset

0 0 Created by dixtn93742 6 minutes ago

Description Add description...

Accuracy 88.3%

Number of tables 1

Data insights

insurance_dataset Model report

Sample size 1,338 | 1,338 Accuracy 88.3% | 91.1% Cosine similarity 0.98208 | 0.99518 Discriminator AUC 58.8% | 45.2% Distances 0.206 | 0.2

Model samples

insurance_dataset

age	gender	bmi	children	smoker	region	expenses
19	male	34.9	2	yes	northwest	40302.4
29	female	27.6	0	no	southwest	1356.2

Click the Model report hyperlink to open the Quality Assurance Report in a separate tab. The page displays similar to what you see in the following screen capture.

Model Report Correlations Univariate Bivariate Accuracy Similarity Distances

Model Report for insurance_dataset:tabular

Generated on 01 Sep 2025, 07:07 • 1,338 original samples, 1,338 synthetic samples

Accuracy 88.3%
(91.1%)

Univariate 96.7%
(96.8%)

Bivariate 90.0%
(92.3%)

Trivariate 78.2%
(84.3%)

Similarity 0.98208
(0.99518)

Cosine Similarity 0.98208
(0.99518)

Discriminator AUC 58.8%
(50.0%)

Distances 0.0%
(0.0%)

Identical Matches 0.0%
(0.0%)

Average Distances 0.206
(0.217)

DCR Share 54.2%
(50.0%)

NNDR Ratio 0.811
(1.000)

Correlations

Correlation Matrices

Note that the training accuracy can be different every time the model is trained.

On the original page, click **Generate data** to use this trained model to generate the required synthetic data.

The screenshot shows the MOSTLY AI interface. On the left sidebar, under the 'Generators' section, 'Synthetic datasets' is highlighted. The main content area displays the 'insurance_dataset' page. At the top, there's a search bar and a 'Private' button. Below the title 'insurance_dataset' (with 0 likes and 0 views, created 6 minutes ago), there's a 'Description' section with a placeholder 'Add description...'. A summary card shows 'Accuracy 88.3%' and 'Number of tables 1'. The 'Data insights' section includes a 'Model report' card with metrics: Sample size 1,338 | 1,338, Accuracy 88.3% | 91.1%, Cosine similarity 0.98208 | 0.99518, and Discriminator AUC 58.8% | 45.2%.

8. Create Synthetic data

You can select the number of samples you want to generate, as well as modify the statistical nature of the data created by choosing the appropriate parameters. For the purpose of this lab, keep all the settings at their default values, and select **Start generation** to create the required synthetic data.

The screenshot shows the 'insurance_dataset' configuration page. The sidebar has 'Synthetic datasets' selected. The main area shows 'Synthetic dataset configuration' using the 'insurance_dataset' generator (Table type: Subject, 1,338 rows). A red box highlights the 'insurance_dataset' entry. Configuration options include 'Sample size' (1,338 rows), 'Conditional simulation', 'Sampling controls', and 'Fairness'.

9. Download the synthetic data

After the synthetic data generation is complete, you will see a web page as shown within the following screen capture.

+ New chat Q Search chats

Datasets Connectors Generators Synthetic datasets

Organizations + New organization

Chats New chat New chat

insurance_dataset

Created by drxtn93742 • 1 minute ago

Description Add description...

Generator used insurance_dataset

Number of tables 1

Total generated rows 1,338

Data insights

insurance_dataset	tabular	Model report	Data report
Generated rows 1,338 1,338	Temperature 1.0	TopP 1.0	Rebalancing Not applied
	Fairness Not applied	Generate with seed Not applied	Imput Not a

Data samples

- insurance_dataset

Click on Download synthetic data to download the dataset created. The dataset can be downloaded in any of the available formats.

+ New chat Q Search chats

Datasets Connectors Generators Synthetic datasets

Organizations + New organization

Chats New chat New chat

insurance_dataset

Created by drxtn93742 • 1 minute ago

Description Add description...

Generator used insurance_dataset

Number of tables 1

Total generated rows 1,338

Data insights

insurance_dataset	tabular	Model report	Data report
Generated rows 1,338 1,338	Temperature 1.0	TopP 1.0	Rebalancing Not applied
	Fairness Not applied	Generate with seed Not applied	Imput Not a

Data samples

- insurance_dataset

You can now use this synthetic data set for data science operations; or, you can also augment the original data set with these samples.

Conclusion

Congratulations! You have completed the lab on data augmentation using the Mostly.ai tool.

Author(s)

[Abhishek Gagneja](#)



Skills Network