

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

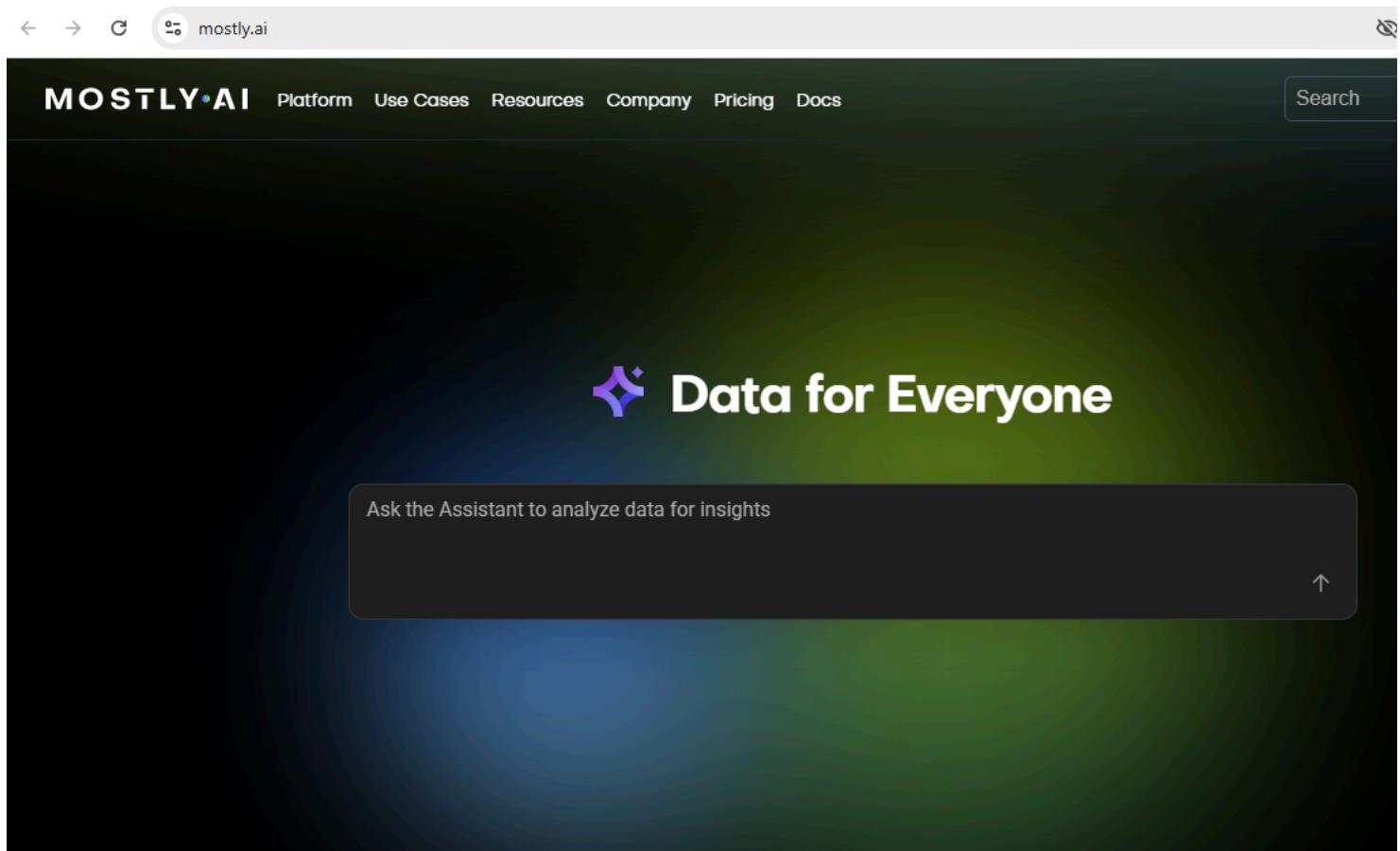
The first step is to download the dataset on your machine. You will need to upload this file to the interface in a subsequent step. Click the link provided in the **Data Set** section to download the data set.

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 application interface. At the top, there is a navigation bar with back, forward, and refresh buttons, followed by the URL 'app.mostly.ai/d/home'. Below the URL is the 'MOSTLY.AI' logo and a search bar with the placeholder 'Search'. On the left side, there is a sidebar with various options: '+ New chat', 'Search chats', 'Datasets', 'Connectors', 'Generators', 'Synthetic datasets', 'Organizations', '+ New organization', 'Chats', 'New chat', and 'New chat'. The main content area features a large green and blue gradient background with the text 'Data for Everyone' and a purple star icon. In the center, there is a text input field with the placeholder 'Ask the Assistant to ana'. Below the input field are several buttons: 'Start with a file', 'Connect your data', 'Explore a dataset', 'Analyze data', and 'Generators'. The overall theme is data management and AI integration.

4. Upload the data set

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

MOSTLY.AI

- + New chat
- Q Search chats
-
- ↳ Datasets
- ↳ Connectors
- AI Generators**
-
- ↳ Synthetic datasets

- Organizations
- + New organization

Chats

- New chat
- New chat

Data for Everyone

Ask the Assistant to g

Start with a file

Connect your data

Explore a dataset

Analyze data

Genera

Popular datasets<https://app.mostly.ai/d/generators>

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

MOSTLY.AI

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

Generators

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

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 imp platform.

Use the SDK

Import a genera

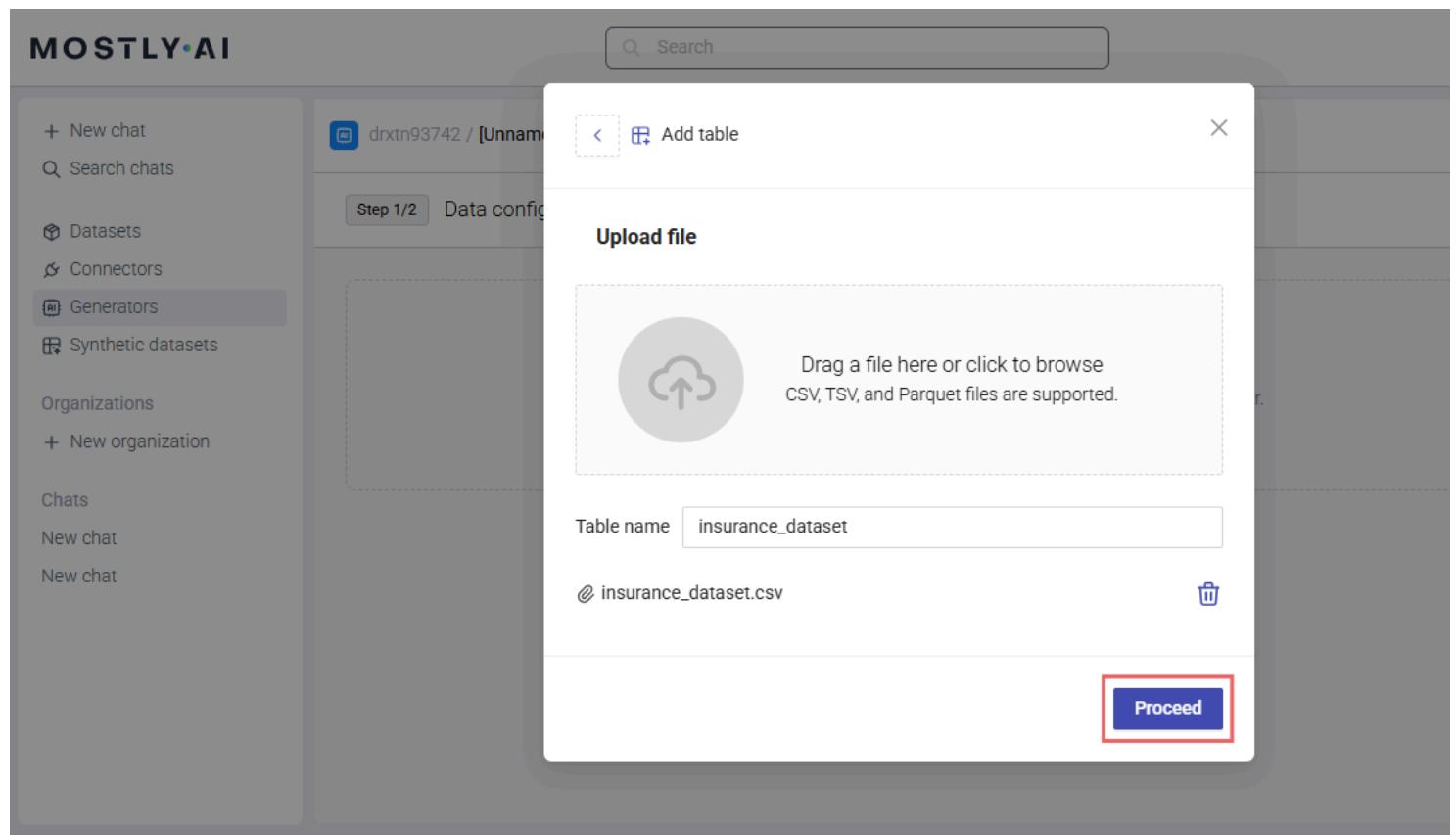
Available generators

Search



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:



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 'Data configuration' page for the 'insurance_dataset'. The dataset name is highlighted with a red box. To the right, the table type is listed as 'Subject' and the number of rows is 1338. At the bottom right of the page, there is a blue 'Add table' button.

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

MOSTLY.AI
 Search

[+ New chat](#)

[Search chats](#)

[Datasets](#)

[Connectors](#)

Generators

[Synthetic datasets](#)

[Organizations](#)

[+ New organization](#)

[Chats](#)

[New chat](#)

[New chat](#)

drxtn93742 / insurance_dataset [Private](#)

Step 1/2 Data configuration

insurance_dataset Table type: Subject 1338 rows

Table relationships

Table columns

Include	Name	Encoding type
<input checked="" type="checkbox"/>	age	Tabular/Numeric: Auto
<input checked="" type="checkbox"/>	gender	Tabular/Categorical
<input checked="" type="checkbox"/>	bmi	Tabular/Numeric: Auto
<input checked="" type="checkbox"/>	children	Tabular/Numeric: Auto
<input checked="" type="checkbox"/>	smoker	Tabular/Categorical
<input checked="" type="checkbox"/>	region	Tabular/Categorical
<input checked="" type="checkbox"/>	expenses	Tabular/Numeric: Auto

Primary key

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.

MOSTLY.AI
 Search

[+ New chat](#)

[Search chats](#)

[Datasets](#)

[Connectors](#)

Generators

[Synthetic datasets](#)

[Organizations](#)

[+ New organization](#)

[Chats](#)

[New chat](#)

[New chat](#)

drxtn93742 / insurance_dataset [Private](#)

Step 2/2 Model configuration

Presets [Accuracy](#)

insurance_dataset tabular Subject table 1,338 rows 10 min

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

Step 2/2 Model configuration

insurance_dataset tabular

Subject table 1,338 rows 10 min

Model Select the tabular model size. MOSTLY_AI/Medium

Compute Select the compute type to train this model. CPU Intel Xeon Spot: 14 CPUs, 26GB

Training parameters Adjust the training parameters to prioritize speed over accuracy, or vice versa. Max training time 10 mins Max sample size 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.

insurance_dataset

Created by dixtn93742 • 6 minutes ago

Description Add description...

Accuracy 88.3% **Number of tables** 1

Data insights

Model report (highlighted)

Sample size	Accuracy	Cosine similarity	Discriminator AUC	Distances
1,338 1,338	88.3% 91.1%	0.98208 0.99518	58.8% 45.2%	0.206 0.217

Model samples

age	gender	bmi	children	smoker	region	expenses
19	male	34.9	2	yes	northwest	40302.47
29	female	27.6	0	no	southwest	1356.25

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.

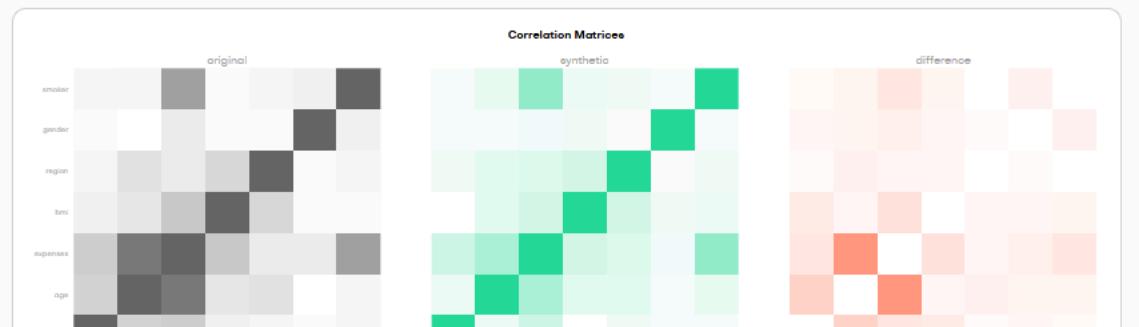
[Model Report](#) [Correlations](#) [Univariates](#) [Bivariate](#) [Accuracy](#) [Similarity](#) [Distances](#)

Model Report for insurance_dataset:tabular

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

Accuracy ⓘ	Univariate	96.7% (96.8%)	Similarity ⓘ	Cosine Similarity 0.98208 (0.99518)	Distances ⓘ	Identical Matches 0.0% (0.0%)
	Bivariate	90.0% (92.3%)		Discriminator AUC 58.8% (50.0%)		Average Distances 0.206 (0.217)
	Trivariate	78.2% (84.3%)				DCR Share 54.2% (50.0%)
						NNDR Ratio 0.811 (1.000)

Correlations



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.

MOSTLY·AI

Search

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

drxtn93742 / insurance_dataset ⓘ Private

insurance_dataset

0 hearts • 0 comments • Created by drxtn93742 • 6 minutes ago

Description

Add description...

Accuracy ⓘ

88.3%

Number of tables

1

Data insights

insurance_dataset

tabular

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%

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.

MOSTLY•AI
 Search

[+ New chat](#)

[Search chats](#)

[Datasets](#)

[Connectors](#)

[Generators](#)

[Synthetic datasets](#) insurance_dataset

[Organizations](#)

[+ New organization](#)

[Chats](#)

[New chat](#)

[New chat](#)

drxtn93742 / insurance_dataset Private

Synthetic dataset configuration Generator used [insurance_dataset](#)

insurance_dataset Table type: Subject 1,338 rows

Sample size Define the number of rows to generate for your synthetic data. rows

Conditional simulation

Sampling controls

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.

MOSTLY•AI

[+ New chat](#)

[Search chats](#)

[Datasets](#)

[Connectors](#)

[Generators](#)

[Synthetic datasets](#) insurance_dataset

[Organizations](#)

[+ New organization](#)

[Chats](#)

[New chat](#)

[New chat](#)

insurance_dataset Created by drxtn93742 • 1 minute ago [Explore](#) [Download](#)

Description Add description...

Generator used	Number of tables	Total generated rows
insurance_dataset	1	1,338

Data insights

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

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.

MOSTLY•AI

Search

- + 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 (0) 1,338
---	------------------------------	--

Data insights

insurance_dataset	tabular	Model report	Data report
Generated rows (0) 1,338 1,338	Temperature (0) 1.0	TopP (0) 1.0	Rebalancing (0) Not applied
Fairness (0) Not applied	Generate with seed (0) Not applied	Imputation (0) Not applied	

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