

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. Select the link provided in the **Data Set** section to download the data set.

2. Open the website

Select 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:

💡 Download the complete guide to AI-generated synthetic data!

MOSTLY AI

[Platform](#)

[Synthetic Data](#)

[Resources](#)

[Company](#)

[Pricing](#)

[Doc](#)



Synthetic Data.

Better than real.

Still struggling with real data? Use existing data for synthetic data generation. Synthetic data is more accessible, more flexible, and simply...smarter.

[Request a demo](#)

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.

← → ↻

app.mostly.ai/d/home

MOSTLY AI

Home

Generators

Synthetic datasets

Connectors

Welcome, Abhishek Gagneja 🙌

Data innovation through Generative AI: train your generator to craft synthetic datasets.

Latest generators

Sample Census Data Generator

✓ Ready

2 weeks ago

Sample Baseball Data Generator

✓ Ready

2 weeks ago

Train a generator with your own data

Upload file

>

Connect to source

>

Get API key

>

How to start?

Explore the available generators and **start generating data.**

+ New sy


4. Upload the data set

Upload the CSV file of the data set to the interface by using the upload option available on the console. After you upload the data set, you will see its filename on the console. Then select Proceed as seen in the following screen captures:

< Add data

×

Upload file



Drag a file here or click to browse

CSV, TSV, and Parquet files are supported.

Proceed

<

+

Add data

×

Upload file

Drag a file here or click to browse
 CSV, TSV, and Parquet files are supported

Table name

insurance_dataset

insurance_dataset.csv

🗑️

Proceed

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

MOSTLY AI

Home

Generators

Synthetic datasets

Connectors

insurance_dataset

🔗

Step 1/2

Data configuration

Rel

Table	Primary key ⓘ	Foreign keys ⓘ
<div>▼</div> <div>insurance_dataset</div>	<div>⋮</div> <div>▼</div>	<div>⋮</div> <div>🔗</div>

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

Add data

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.

← → ↺

app.mostly.ai/d/generators/757e583d-1cf2-439c-acb9-ecabcb9c243/model-config

MOSTLY·AI

HomeGeneratorsSynthetic datasetsConnectors

insurance_dataset

Step 2/2

Model configuration

Configuration presets ⓘ

Accuracy ⓘ

ⓘ

1

Models ⓘ

insurance_dataset

Table type ⓘ

Subject

Max sample size ⓘ

1,338 rows

Max training time ⓘ

10 min

Max sequence window ⓘ

-

Max sample size ⓘ

1,338

rows

Max training time ⓘ

10

mins

Max sequence window ⓘ

Not applicable for subject tables

Max training epochs ⓘ

100

Model size ⓘ

Medium

Batch size ⓘ

Auto

Flexible generation ⓘ

OnOff

Value protection ⓘ

OnOff

Rare category replacement ⓘ

Constant

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

7. Model training

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

MOSTLY·AI

HomeGeneratorsSynthetic datasetsConnectors

insurance_dataset

Trained by Abhishek Gagneja • Created on March 20, 2024 at 00:02

Accuracy

92.9%

Description

Edit description

Data insights

Table	Accuracy ⓘ			
	Overall	Univariate	Bivariate	C
insurance_dataset	92.9% (94.6%)	96.8%	89.0%	-

Model samples

Training status

Ready

Configuration

Click the **Model** 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 for `insurance_dataset`

generated on 19 Mar 2024, 19:01

Dataset ⓘ

Original Samples	1,338
Synthetic Samples	1,338
Target Columns	7

Accuracy ⓘ

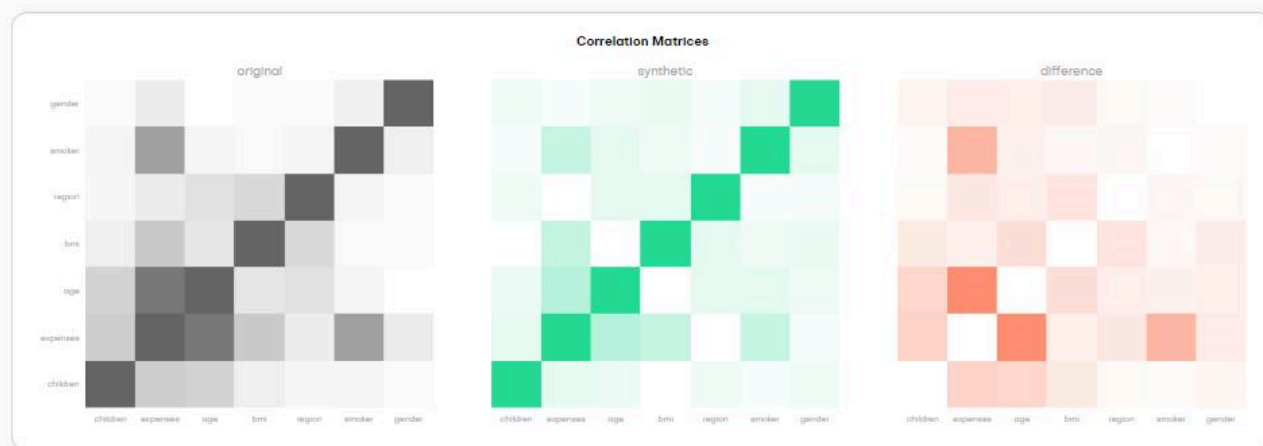
Univariate	96.8%
Bivariate	89.0%

92.9%
(94.6%)

Distances ⓘ

Identical Matches	0.0% (0.1%)
Average Distances	2.04 (1.30)

Correlations



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

On the original page, click [Generate Synthetic Data](#) to use this trained model to generate the required synthetic data.

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[Home](#)[Generators](#)[Synthetic datasets](#)[Connectors](#)

insurance_dataset

Generator used insurance_dataset

Configure Synthetic Dataset

[Dataset destination: Download as CSV/PARQUET/XLSX](#) [Relationships](#)

Table	Sample size ⓘ	Temperature ⓘ	Top P ⓘ	Imputed columns ⓘ	Rebalancing column ⓘ
insurance_dataset	1,338 rows	1	1	-	-

Sample size

1,338 rows

Temperature

1

Top P

1

Imputed columns

Rebalancing column

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.

insurance_dataset

Generator used insurance_dataset

Generated by Abhishek Gagneja • Created on March 20, 2024 at 00:37

Overall accuracy

93.1%

Data points

9,366

Used credits

0.01

Description

Edit description

Data insights

Table	Accuracy ⓘ				Distances ⓘ	Reports
	Overall	Univariate	Bivariate	Coherence		
insurance_dataset	93.1% (94.6%)	96.9%	89.3%	-	2.04 (1.27)	Model Data

Data samples

insurance_dataset

age	gender	bmi	children	smoker	region	expenses
37	male	38.2	3	no	southeast	7144.63
45	male	36	0	no	northeast	1458.47
53	female	28.8	0	no	northwest	2277.83

Click on Download synthetic dataset to download the dataset created.
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

© IBM Corporation. All rights reserved.

