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, MostlyAI, 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

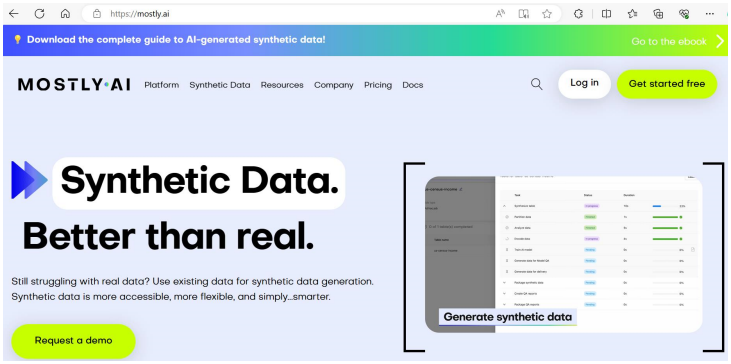
The first step is to download the dataset on your machine. You will need to upload this file to the interface in a subsequent step. Select the link provided in the **Data Set** section to download the data set.

2. Open the website

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

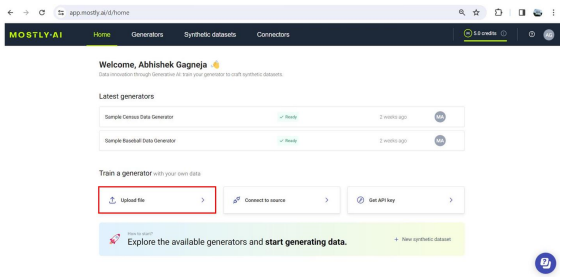
https://mostly.ai/.

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



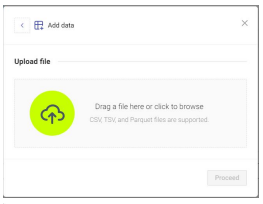
3. Create an account

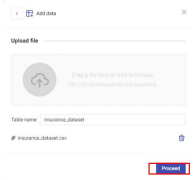
You can create an account on this website free of charge, or you can simply log in using your Gmail ID. After you log in, you'll see the following interface.



4. Upload the data set

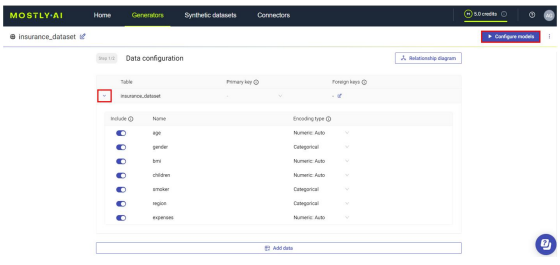
Upload the CSV file of the data set to the interface by using the upload option available on the console. After you upload the data set, you will see its filename on the console. Then select `Proceed` as seen in the following screen capture:





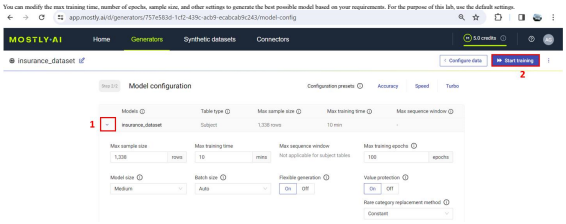
5. Data configuration settings

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.



6. Model configuration settings

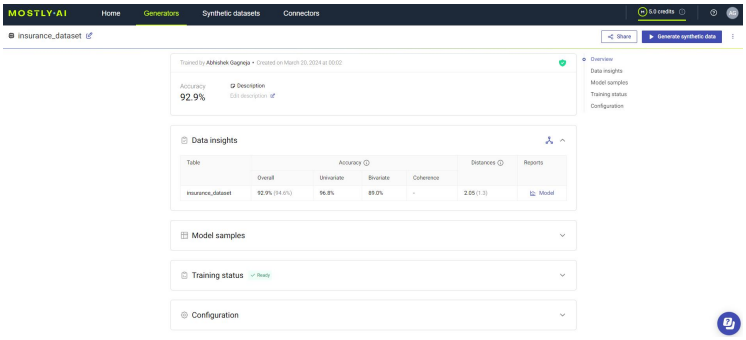
You can modify the max training time, number of epochs, sample size, and other settings to generate the best possible model based on your requirements. For the purpose of this lab, use the default settings.



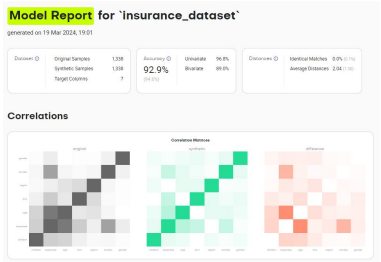
When you complete working with the settings, select `Start training`. You will find this option on the top right corner of the web page.

7. Model training

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



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.

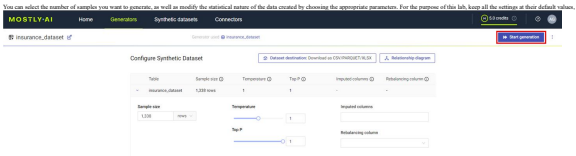


Note that the training accuracy can be different every time the model is trained.

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

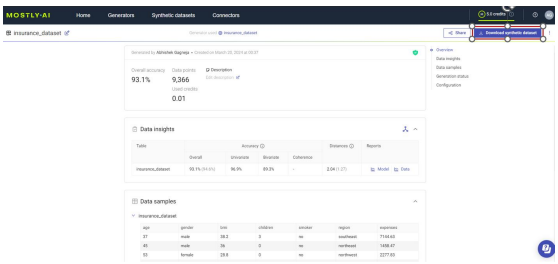
8. Create Synthetic data

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.



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.



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

Skills Network