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

Estimated Effort: 30 mins

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 are going to use a popular tool, Mostly.ai, for creating synthetic data samples to augment a CSV data set.

Data Set

You are going to use a data set on Insurance records.

The data set is available at the link below.
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

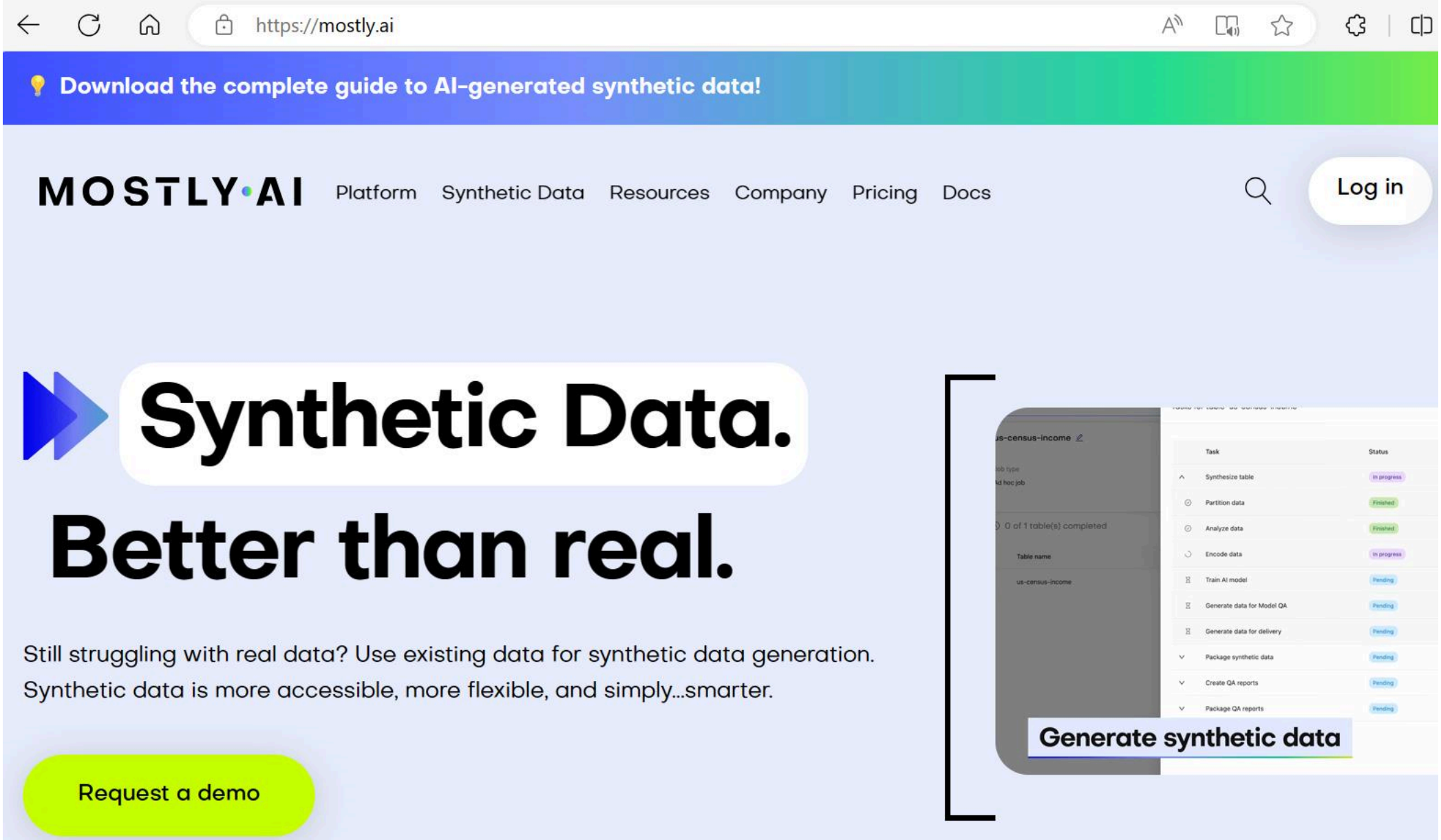
The first step is to download the dataset on your machine. You will need to upload this file to the interface in a subsequent step. Click the link provided in the **Data Set** section to download it.

2. Open the website

Click the link below to open the interface.

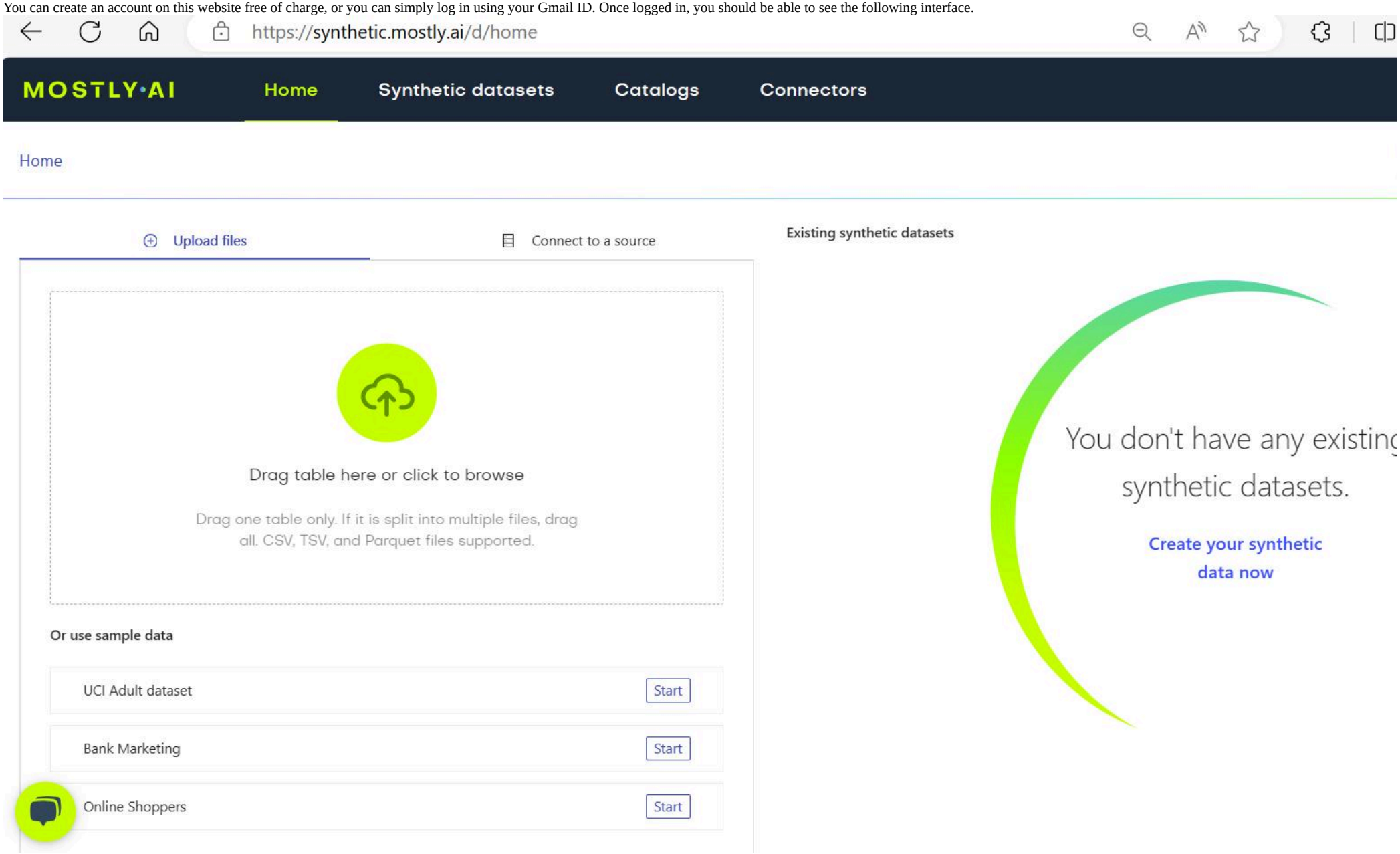
https://mostly.ai/

This link will open in a new browser tab, and you should see an interface that looks as shown below.



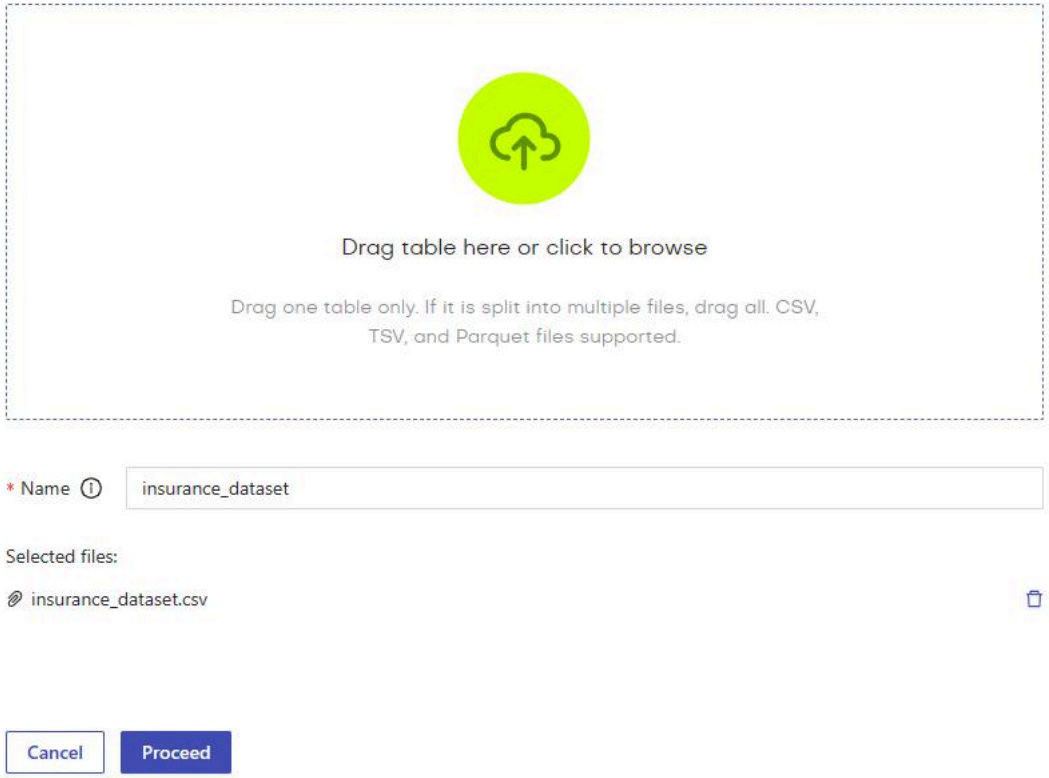
3. Create an account

You can create an account on this website free of charge, or you can simply log in using your Gmail ID. Once logged in, you should be able to see the following interface.



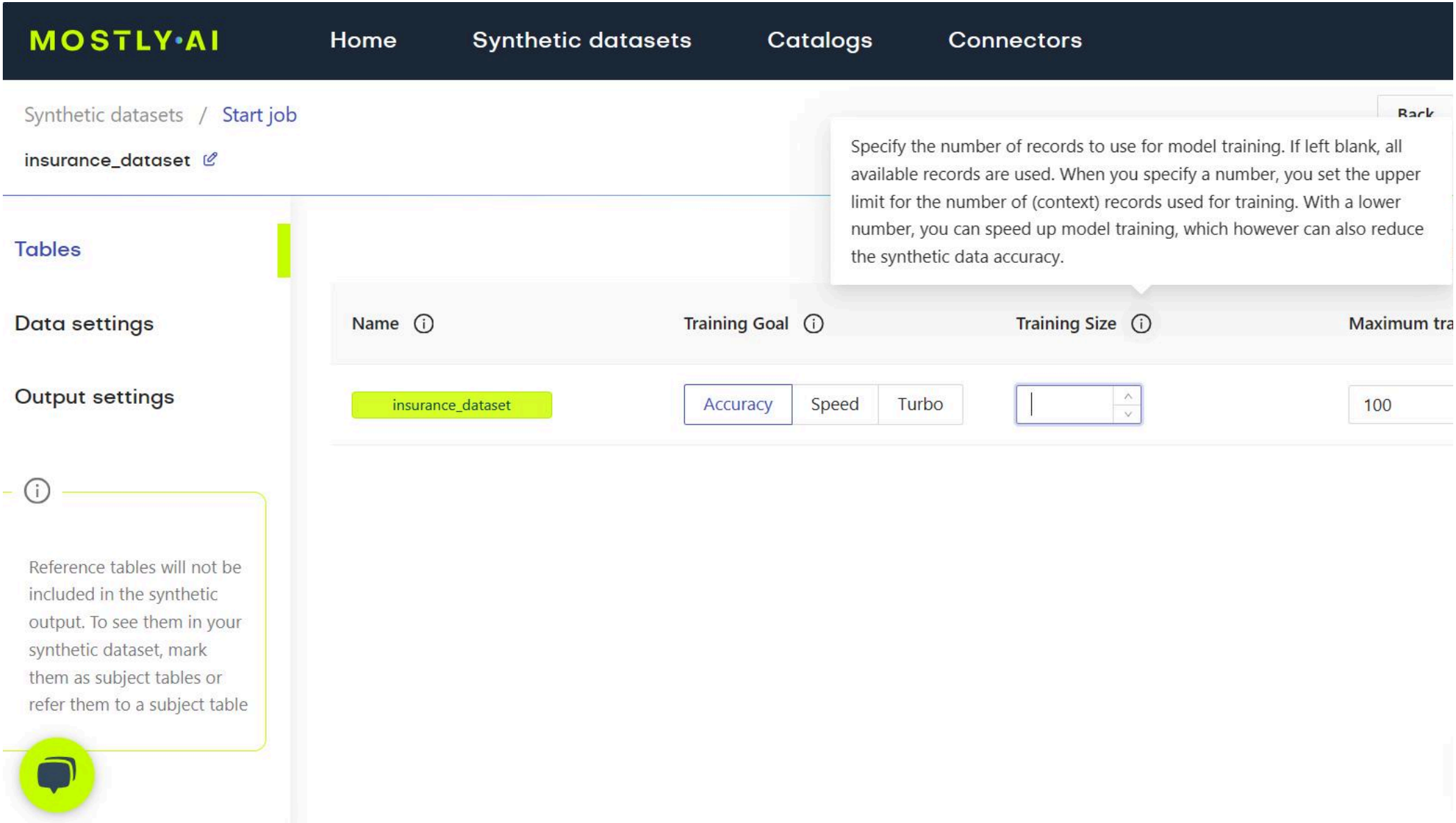
4. Upload the data set

Upload the CSV file of the data set to the interface by using the upload option available on the console. Once uploaded, the filename will be visible on the console. Then you can click `Proceed` as shown below.



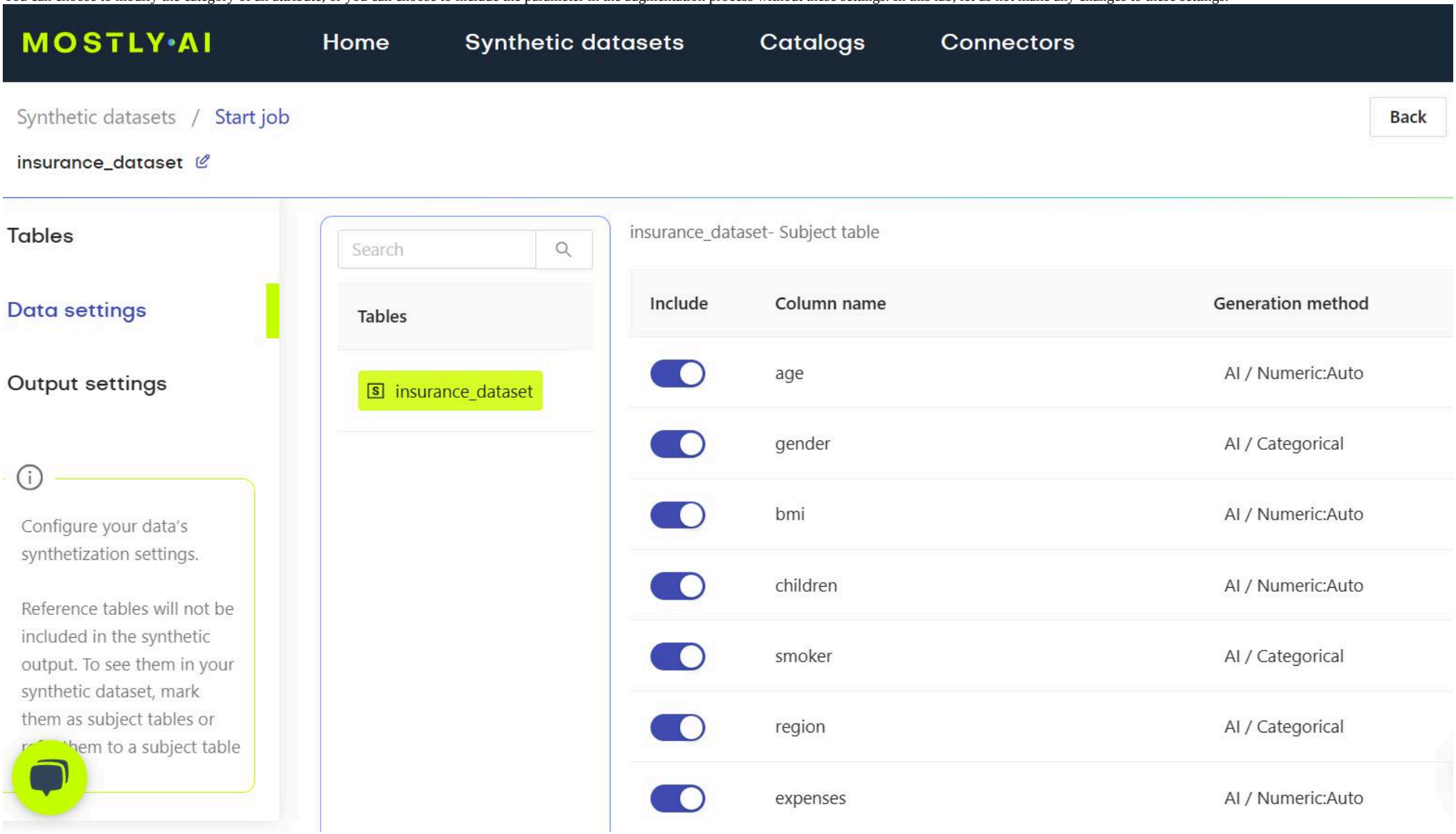
5. Training parameter selection

On the interface, you will see a space provided to choose the number of training samples that can be used from your data set. If you leave this entry blank, it is interpreted as allowing the use of all entries in the data set to train the generative model. You may leave this entry blank, since the more data that is used for training the model, the more accurate the synthetic data created will be.



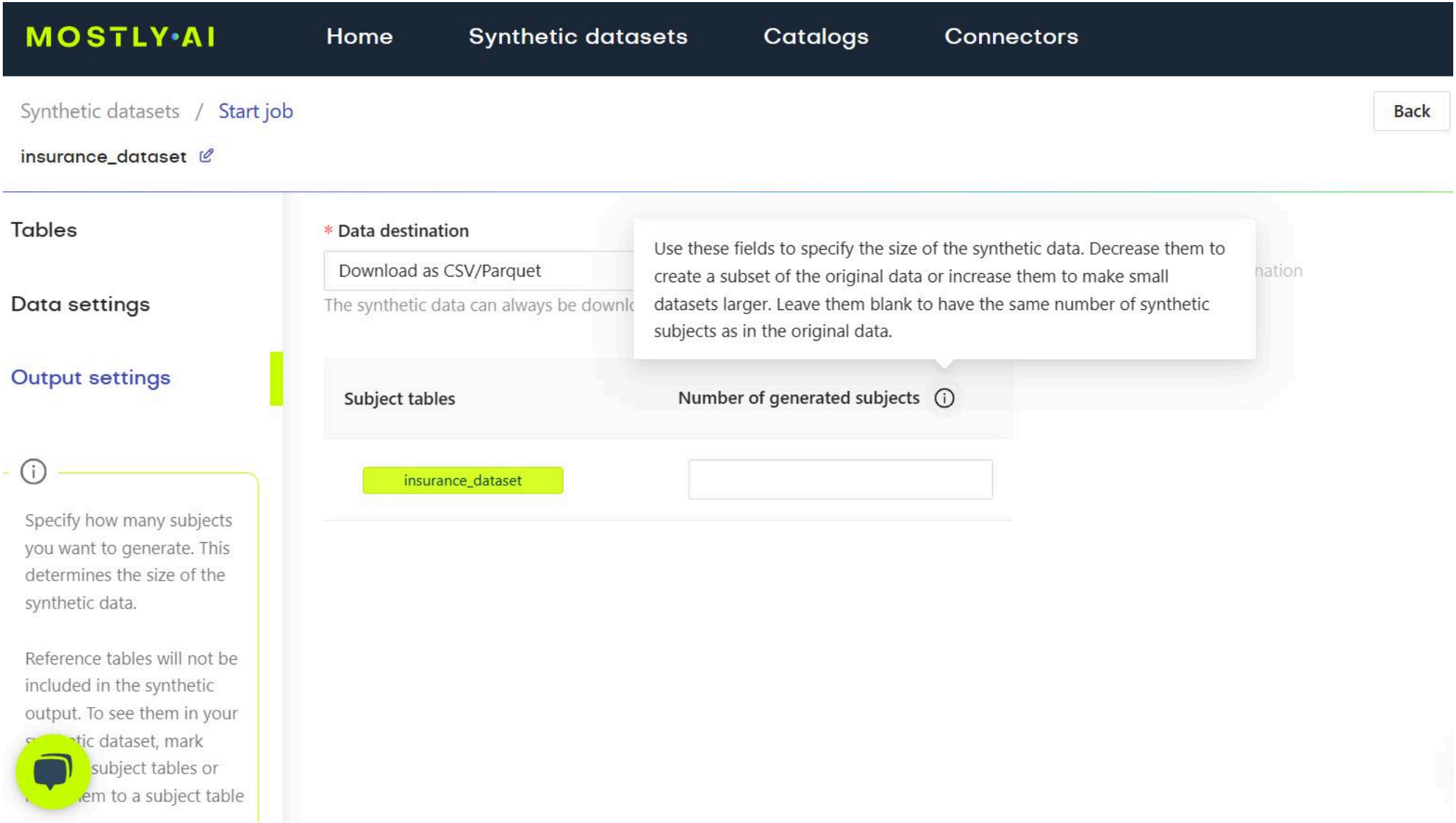
6. Data settings

You can choose to modify the category of an attribute, or you can choose to include the parameter in the augmentation process without these settings. In this lab, let us not make any changes to these settings.



7. Output settings

You can specify the number of samples you want to create synthetically using this setting. If you leave it blank, it generates the exact same number of data samples as there are in the original file. Let us leave this setting blank, choosing to generate the same number of samples as in the original data set.



Once these settings are done, click `Create a synthetic dataset`, as visible on the top right corner of the webpage.

8. Data set creation process

You will see an interface as shown below.

MOSTLY·AI Home Synthetic datasets Catalogs Connectors

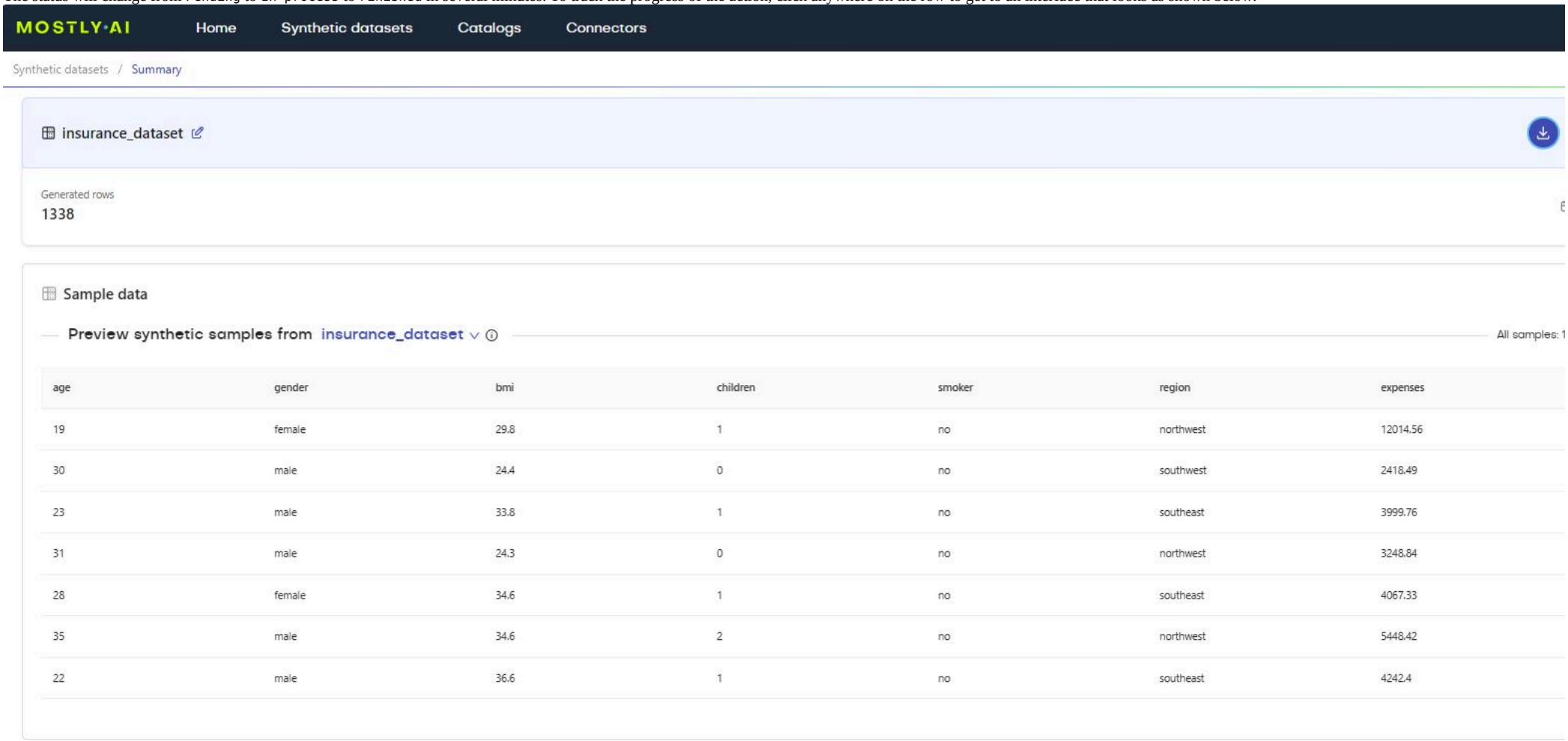
Synthetic datasets

Daily consumption ⓘ

0%

Name	Type	Status	Creation date
insurance_dataset	Upload a file	In progress	Dec 12, 2023 6:51 PM

The status will change from `Pending` to `In process` to `Finished` in several minutes. To track the progress of the action, click anywhere on the row to get to an interface that looks as shown below.



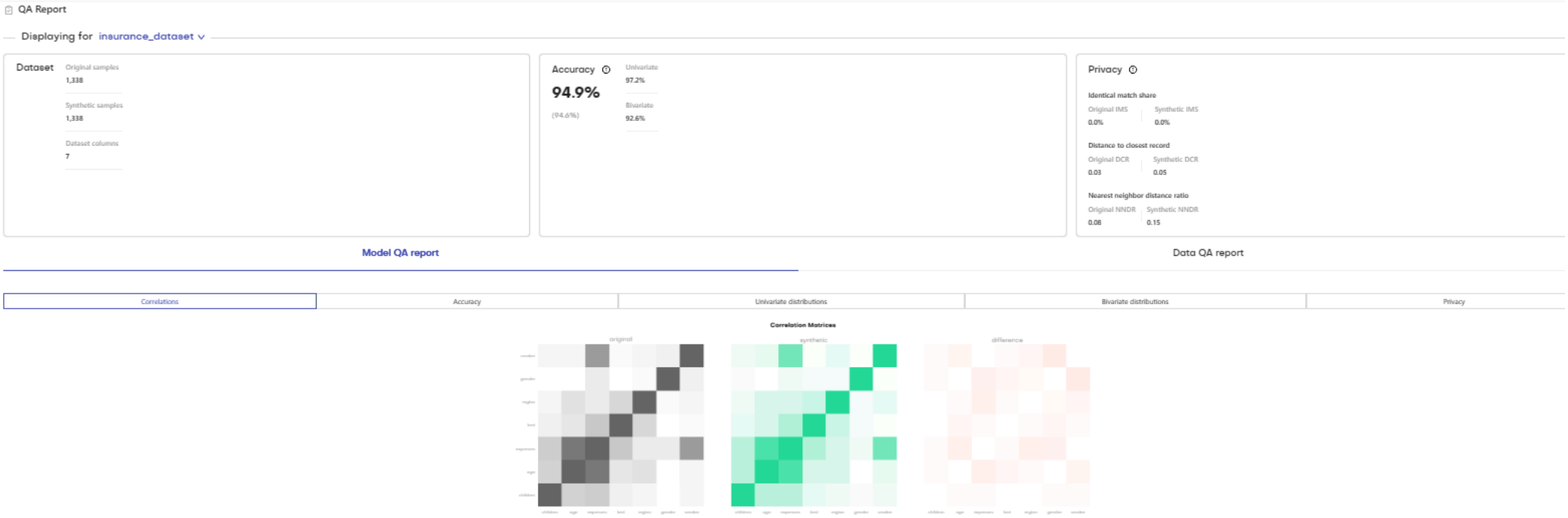
age	gender	bmi	children	smoker	region	expenses
19	female	29.8	1	no	northwest	12014.56
30	male	24.4	0	no	southwest	2418.49
23	male	33.8	1	no	southeast	3999.76
31	male	24.3	0	no	northwest	3248.84
28	female	34.6	1	no	southeast	4067.33
35	male	34.6	2	no	northwest	5448.42
22	male	36.6	1	no	southeast	4242.4

In the top right corner, you can access the log section to note each of the tasks happening in the process of this data creation.

Logs

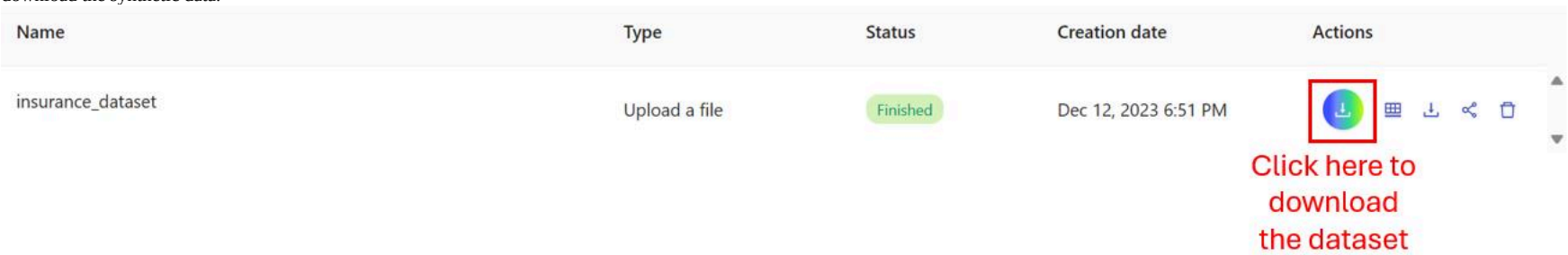
Model	Model Type	Step	Progress	Durati
insurance_dataset	Flat	Fetch data		
insurance_dataset	Flat	Analyze data		
insurance_dataset	Flat	Encode data		
insurance_dataset	Flat	Train AI model	Trained 36 epochs	
insurance_dataset	Flat	Generate data for Model QA	1.3K out of 1.3K done	
insurance_dataset	Flat	Analyze data for Model QA		
insurance_dataset	Flat	Generate data for delivery	1.3K out of 1.3K done	
insurance_dataset	Flat	Analyze data for Data QA		
insurance_dataset	Flat	Post-process data		
insurance_dataset	Flat	Export data		
↳		Package synthetic dataset		
↳		Package QA reports		

Once completed, you can also check the statistical similarity of the synthetic data samples with the original dataset in the QA report section. It will also show the accuracy score of the synthetic data set.



9. Download the synthetic data

Once the process is complete, you can click `Synthetic datasets` on the web page to see that the status will show as `Finished`. At this point, your synthetic data set is now ready. Click the download link, as shown in the image below, to download the synthetic data.



You may now use this synthetic data set for data science operations, or you can also augment the original data set with these samples to be used together.

Conclusion

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

Author(s)

[Abhishek Gagneja](#)

Skills Network