

# Laboratoire pratique : IA générative pour la génération et l'augmentation des données

Temps estimé nécessaire : **30** minutes

L'un des principaux avantages de l'IA générative est sa capacité à générer des données synthétiques réalistes. Les données synthétiques sont générées lorsqu'un modèle génératif pré-entraîné répond à une invite, crée de nouveaux échantillons de données ou transfère des apprentissages sur un ensemble de données donné. En outre, il crée des échantillons qui peuvent augmenter l'ensemble de données existant tout en conservant la distribution statistique et l'interprétabilité de l'ensemble de données.

Dans ce laboratoire, vous apprendrez à utiliser l'IA générative pour générer des échantillons de données synthétiques et transférer des apprentissages sur un ensemble de données donné.

## Objectif d'apprentissage

Dans ce laboratoire, vous apprendrez à utiliser un outil populaire, [Mostly.ai](#), pour créer des échantillons de données synthétiques afin d'augmenter un ensemble de données CSV.

## Ensemble de données

Vous utiliserez un ensemble de données comprenant des dossiers d'assurance.

L'ensemble de données est disponible sur le lien suivant :

[Ensemble de données sur les assurances](#)

Cet ensemble de données est une version nettoyée de l'ensemble de données [de prévision des prix de l'assurance médicale](#), disponible sous la [licence universelle CC0 1.0](#) sur le site Web [Kaggle](#).

## Mesures

### 1. Téléchargez l'ensemble de données

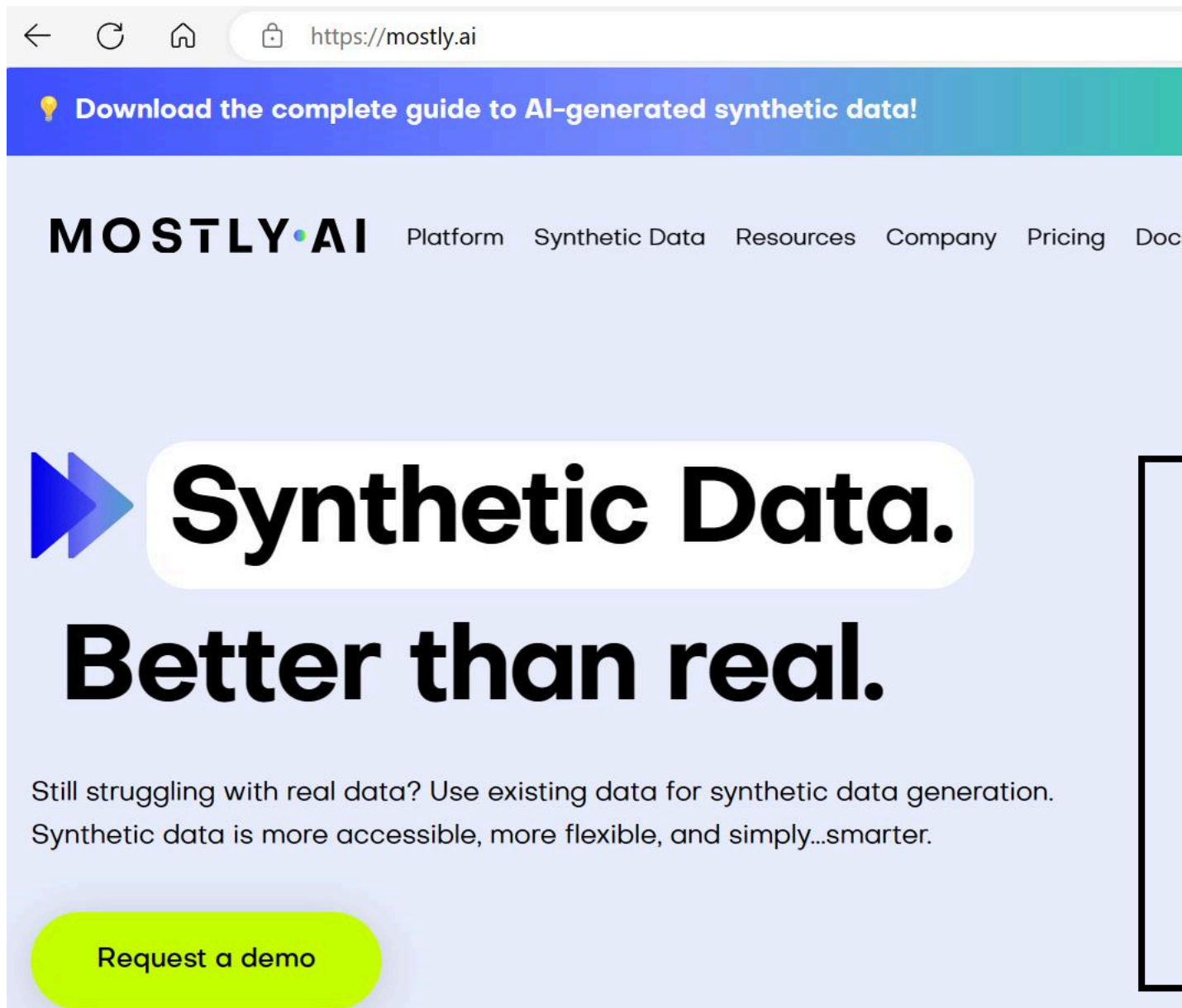
La première étape consiste à télécharger l'ensemble de données sur votre machine. Vous devrez télécharger ce fichier sur l'interface lors d'une étape ultérieure. Sélectionnez le lien fourni dans la section **Ensemble de données** pour télécharger l'ensemble de données.

### 2. Ouvrez le site Web

Sélectionnez le lien suivant pour ouvrir le site Web et l'interface mostly.ai.

<https://mostly.ai/>

Ce lien s'ouvre dans un nouvel onglet de navigateur et vous devriez voir une page Web qui ressemble à la capture d'écran suivante :



← ↻ 🏠 <https://mostly.ai>

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

**MOSTLY AI** Platform Synthetic Data Resources Company Pricing Docs

# 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. Créer un compte

Vous pouvez créer un compte sur ce site gratuitement ou vous connecter simplement à l'aide de votre identifiant Gmail. Une fois connecté, vous verrez l'interface suivante.

← → ↻ 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 syn](#)

#### 4. Téléchargez l'ensemble de données

Téléchargez le fichier CSV de l'ensemble de données sur l'interface en utilisant l'option de téléchargement disponible sur la console. Après avoir téléchargé l'ensemble de données, vous verrez son nom de fichier sur la console. Sélectionnez ensuite Proceed comme indiqué dans les captures d'écran suivantes :

< 📁 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.csv

Proceed

5. Paramètres de configuration des données

Vous pouvez choisir de modifier la catégorie d'un attribut ou d'inclure un paramètre dans le processus d'augmentation sans ces paramètres. Pour les besoins de ce laboratoire, ne modifiez pas ces paramètres. Sélectionnez simplement `Configure model` pour accéder aux paramètres de configuration du modèle.

MOSTLY AI

Home

Generators

Synthetic datasets

Connectors

insurance\_dataset

Step 1/2

Data configuration

Relat

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

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. Paramètres de configuration du modèle

Vous pouvez modifier le temps d'entraînement maximal, le nombre d'époques, la taille de l'échantillon et d'autres paramètres pour générer le meilleur modèle possible en fonction de vos besoins. Pour les besoins de ce laboratoire, utilisez les paramètres par défaut.

← → ↺

app.mostly.ai/d/generators/757e583d-1cf2-439c-acb9-ecab9c243/model-config

MOSTLY.AI

HomeGeneratorsSynthetic datasetsConnectors

insurance\_dataset

Step 2/2

Model configuration

Configuration presets ⓘ

Accuracy

Sp

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 ⓘ

On

Off

Value protection ⓘ

On

Off

Rare category replacement ⓘ

Constant

Une fois les paramètres définis, sélectionnez Start training. Vous trouverez cette option dans le coin supérieur droit de la page Web.

7. Entraînement du modèle

Une fois la formation du modèle terminée, vous verrez un résultat à l'écran similaire à celui que vous voyez sur la capture d'écran suivante.

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	Col
insurance_dataset	92.9% (94.6%)	96.8%	89.0%	-

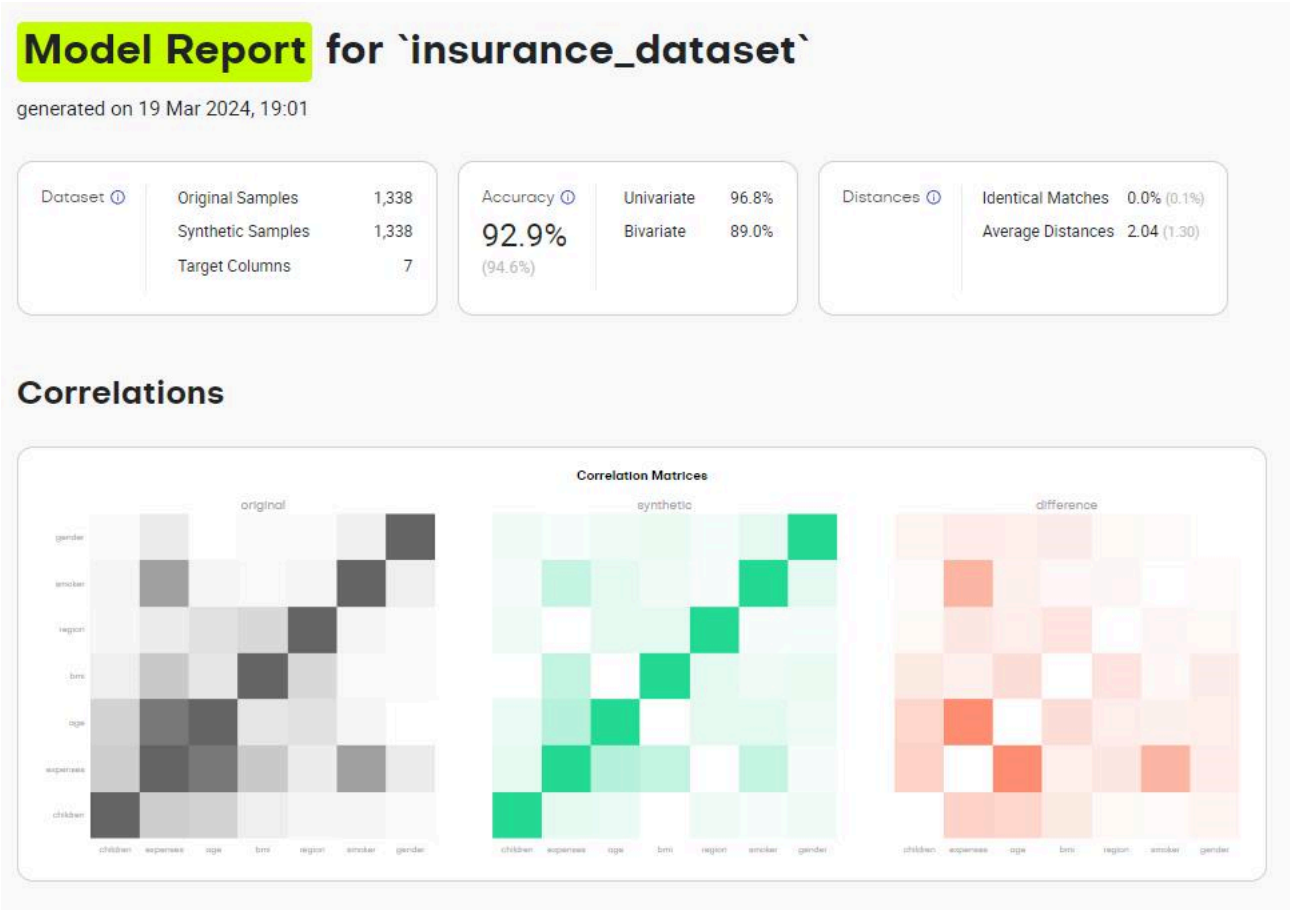
Model samples

Training status

Ready

Configuration

Cliquez sur le [Modèle](#) lien hypertexte pour ouvrir le rapport d'assurance qualité dans un onglet séparé. La page s'affiche de manière similaire à ce que vous voyez dans la capture d'écran suivante.



MOSTLY AI

HomeGeneratorsSynthetic datasetsConnectors

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)	<a href="#">Model</a> <a href="#">Data</a>

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

Cliquez sur [Download synthetic dataset](#) pour télécharger l'ensemble de données créé. Vous pouvez désormais utiliser cet ensemble de données synthétiques pour des opérations de science des données ; ou, vous pouvez également compléter l'ensemble de données d'origine avec ces échantillons.

Conclusion

Félicitations ! Vous avez terminé le laboratoire sur l'augmentation des données à l'aide de l'outil Mostly.ai.

Auteur(s)

[Abhishek Gagneja](#)



Skills Network