What is it ?

Synthetic data is artificial data that is created by using different algorithms that mirror the statistical properties of the original data but does not reveal any information regarding real people.

Use case: privacy, product testing and training machine learning algorithms.

Approaches:

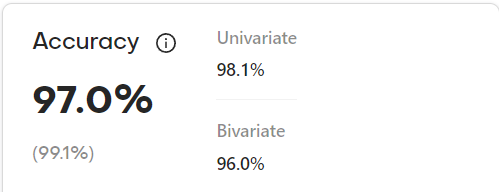
Drawing Numbers From a Distribution: create a distribution of data that follows a curve that is loosely based on real-world data.

Generative Models: Involves automatically discovering and learning the insights and patterns in data in such a way that the model can be used to output new examples that match the same distribution as the real-world data it was trained on. Two common approaches:

1. GAN: treat the training process as a game between two separate networks - a generator network, and a second discriminative network that attempts to classify samples as either coming from the real world (training distribution), or coming from the model (synthetic data). On each training iteration, the generator adjusts its model parameters to generate more convincing examples to (effectively) fool the discriminator, until, ideally, the discriminator is no longer able to differentiate between the real world and synthetic examples.
2. VAE: unsupervised method where encoder compresses the original dataset into a more compact structure and transmits data to the decoder. Then the decoder generates an output which is a representation of the original dataset.

Tools:

1. MOSTLY AI:



1. Gretel:

