Information Technology (IT)

Multiplatform Data Acquisition, Collection And Analytics (MDACA)

Synthetic Data Engine User Guide

Version 2.0

BACKGROUND

Powered by Spin Systems Inc. (SpinSys), Multiplatform Data Acquisition, Collection and Analytics (MDACA) is a copyright of SpinSys. All other copyrights, trademarks, and trade names are the property of their respective owners.

Please visit http://mdaca.io/terms/ for our terms of use and http://mdaca.io/privacy/ for our privacy policy.

TABLE OF CONTENTS

LIST OF FIGURES

1 PURPOSE AND OVERVIEW

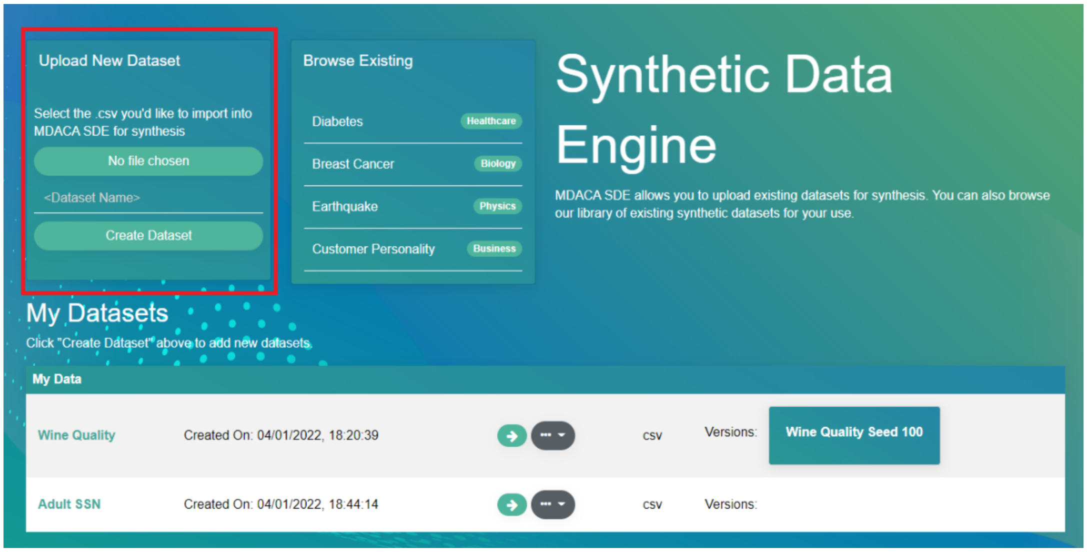
MDACA (Multiplatform Data Acquisition, Collection and Analytics) is a scalable big data suite of applications for data acquisition, storage, and access. The MDACA data de-identification and synthetic data generation tools that are available within the MDACA product suite provide end-to-end capabilities for anonymization of sensitive data and generation of shareable, synthetic data sets.

This document is a technical user guide for analysts, data scientists and data owners looking to generate synthetic data with high level of statistical fidelity and strong privacy guarantees.

2 SYNTHETIC DATA GENERATION

The MDACA Synthetic Data Generator is a web-based tool for taking in a de-identified data set as input and generating a synthetic dataset as output that is structurally and statistically similar to the de-identified data set.

Click the topmost green button to select the CSV file from your local directory for upload. You may optionally assign the file a label so that it can be easily re-used later. See Figure 1.



Select the .csv you'd like to import into MDACA SDE for synthesis

Diabetes	Healthcare
Breast Cancer	Biology
Earthquake	Physics
Customer Personality	Business

Synthetic Data Engine

MDACA SDE allows you to upload existing datasets for synthesis. You can also browse our library of existing synthetic datasets for your use.

My Datasets

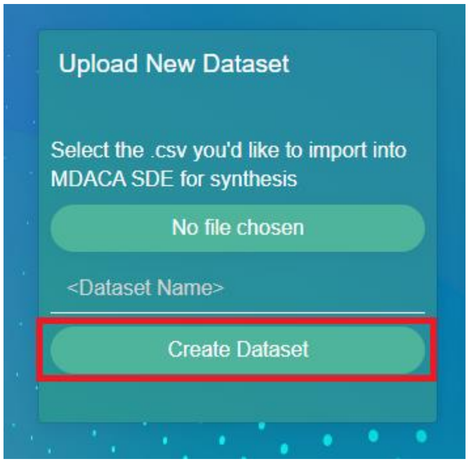
Click "Create Dataset" above to add new datasets.

My Data

Wine Quality	Created On: 04/01/2022, 18:20:39	csv	Versions:	Wine Quality Seed 100
Adult SSN	Created On: 04/01/2022, 18:44:14	csv	Versions:	

Figure 1: Choose File

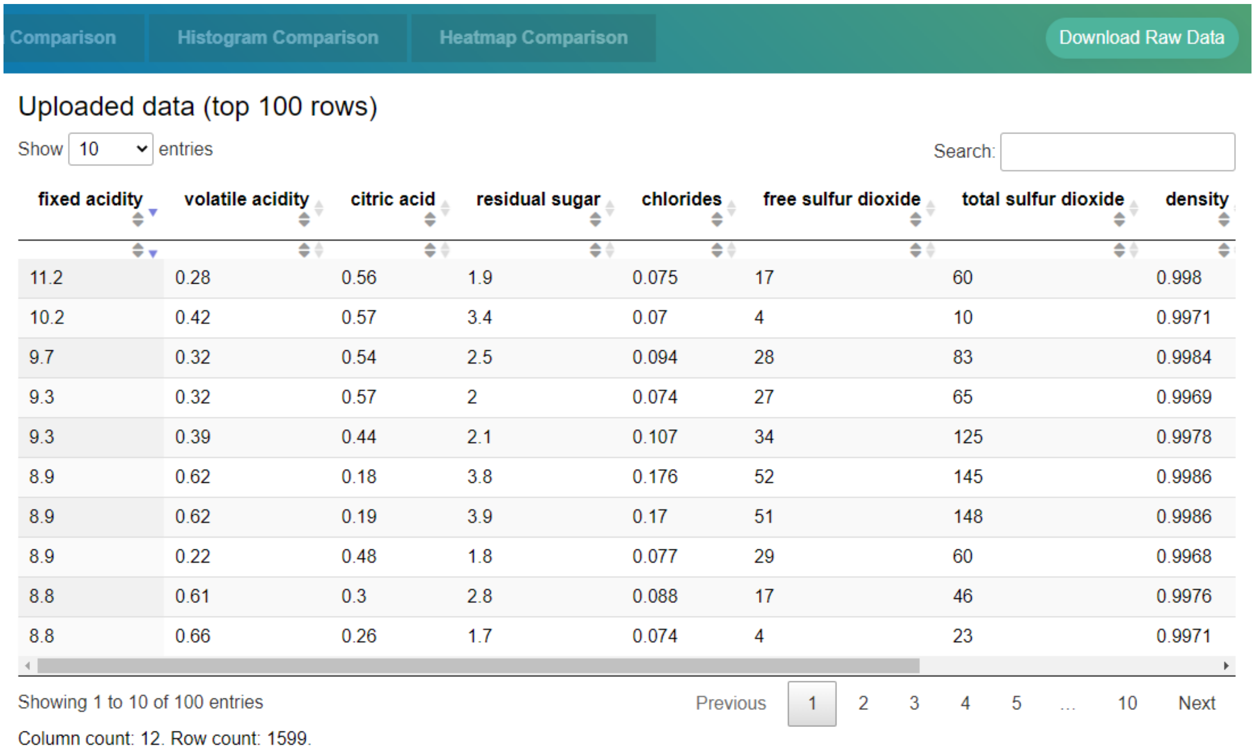
After the file has been selected, click the **Upload Data** button to upload the file to the server. The time to upload the file will vary depending on the size of the file. See Figure 2.



Select the .csv you'd like to import into MDACA SDE for synthesis

Figure 2: Create Dataset

After the CSV file has been uploaded, the **Uploaded data** section of the page displays the top 100 rows of the de-identified CSV file in a paginated format. Check for any missing attribute values and inadvertent exposure of PHI/PII, sensitive information. See Figure 3.



Uploaded data (top 100 rows)

Show 10 entries

Search:

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density
11.2	0.28	0.56	1.9	0.075	17	60	0.998
10.2	0.42	0.57	3.4	0.07	4	10	0.9971
9.7	0.32	0.54	2.5	0.094	28	83	0.9984
9.3	0.32	0.57	2	0.074	27	65	0.9969
9.3	0.39	0.44	2.1	0.107	34	125	0.9978
8.9	0.62	0.18	3.8	0.176	52	145	0.9986
8.9	0.62	0.19	3.9	0.17	51	148	0.9986
8.9	0.22	0.48	1.8	0.077	29	60	0.9968
8.8	0.61	0.3	2.8	0.088	17	46	0.9976
8.8	0.66	0.26	1.7	0.074	4	23	0.9971

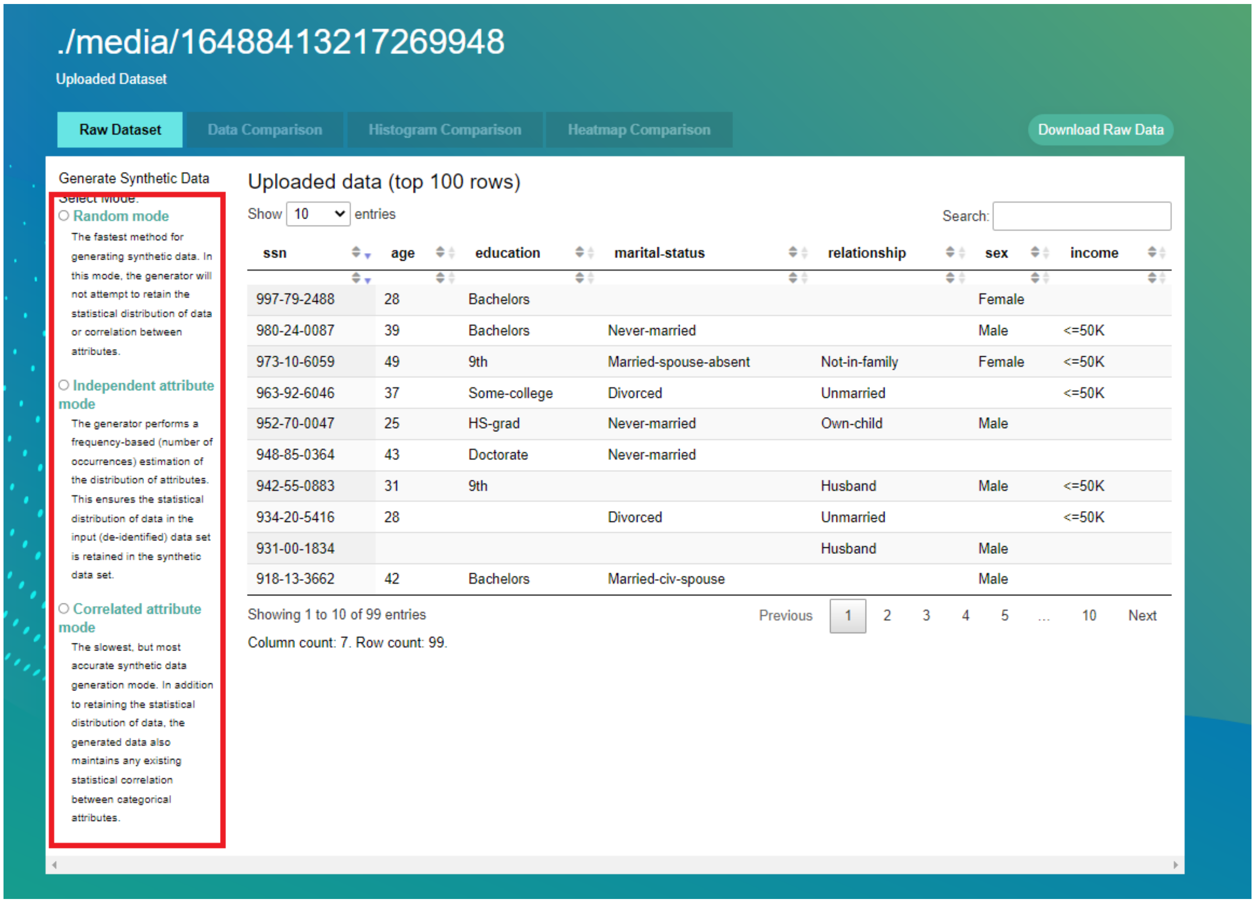
Showing 1 to 10 of 100 entries

Column count: 12. Row count: 1599.

Figure 3: Uploaded Data Section

2.1 Parameters Setup

Once complete, a Generation Mode must be chosen. See Figure 4.



./media/16488413217269948

Uploaded Dataset

Generate Synthetic Data

Select Mode:

○ Random mode

The fastest method for generating synthetic data. In this mode, the generator will not attempt to retain the statistical distribution of data or correlation between attributes.

○ Independent attribute mode

The generator performs a frequency-based (number of occurrences) estimation of the distribution of attributes. This ensures the statistical distribution of data in the input (de-identified) data set is retained in the synthetic data set.

○ Correlated attribute mode

The slowest, but most accurate synthetic data generation mode. In addition to retaining the statistical distribution of data, the generated data also maintains any existing statistical correlation between categorical attributes.

Uploaded data (top 100 rows)

Show 10 entries

Search:

ssn	age	education	marital-status	relationship	sex	income
997-79-2488	28	Bachelors			Female	
980-24-0087	39	Bachelors	Never-married		Male	<=50K
973-10-6059	49	9th	Married-spouse-absent	Not-in-family	Female	<=50K
963-92-6046	37	Some-college	Divorced	Unmarried		<=50K
952-70-0047	25	HS-grad	Never-married	Own-child	Male	
948-85-0364	43	Doctorate	Never-married			
942-55-0883	31	9th		Husband	Male	<=50K
934-20-5416	28		Divorced	Unmarried		<=50K
931-00-1834				Husband	Male	
918-13-3662	42	Bachelors	Married-civ-spouse		Male	

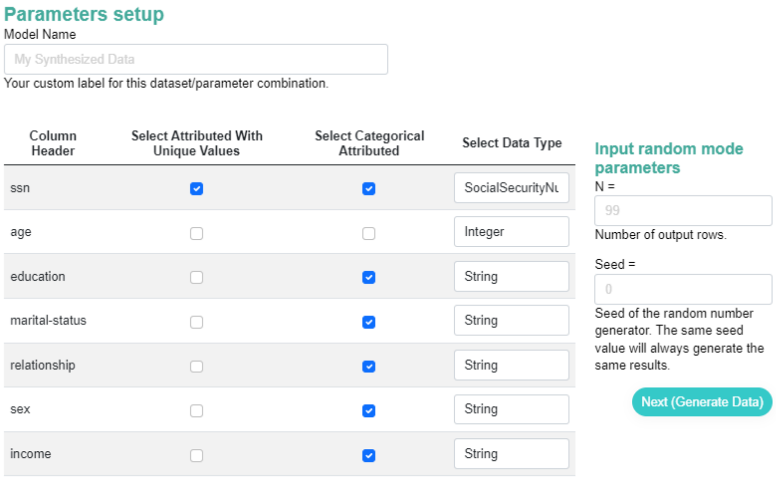
Showing 1 to 10 of 99 entries

Column count: 7. Row count: 99.

Figure 4: Choose Generation Mode

2.1.1 Random Mode

Random mode is the fastest method for generating synthetic data. In this mode, the generator will not attempt to retain the statistical distribution of data or correlation between attributes. The data generator will generate random, type-consistent values for each attribute. For string attributes, the random string generated will always fall within the observed range of string lengths within the dataset. Attributes with less than 20 distinct or unique values are identified as **Categorical attributes**. When a particular attribute is repeated in the dataset, it will not be marked as a categorical attribute. However, the analyst can override this behavior by selecting or deselecting the checkbox for that attribute. See Figure 5.



Parameters setup

Model Name

My Synthesized Data

Your custom label for this dataset/parameter combination.

Column Header	Select Attributed With Unique Values	Select Categorical Attributed	Select Data Type
ssn	☑	☑	SocialSecurityNu
age	□	□	Integer
education	□	☑	String
marital-status	□	☑	String
relationship	□	☑	String
sex	□	☑	String
income	□	☑	String

Input random mode parameters

N =

99

Number of output rows.

Seed =

0

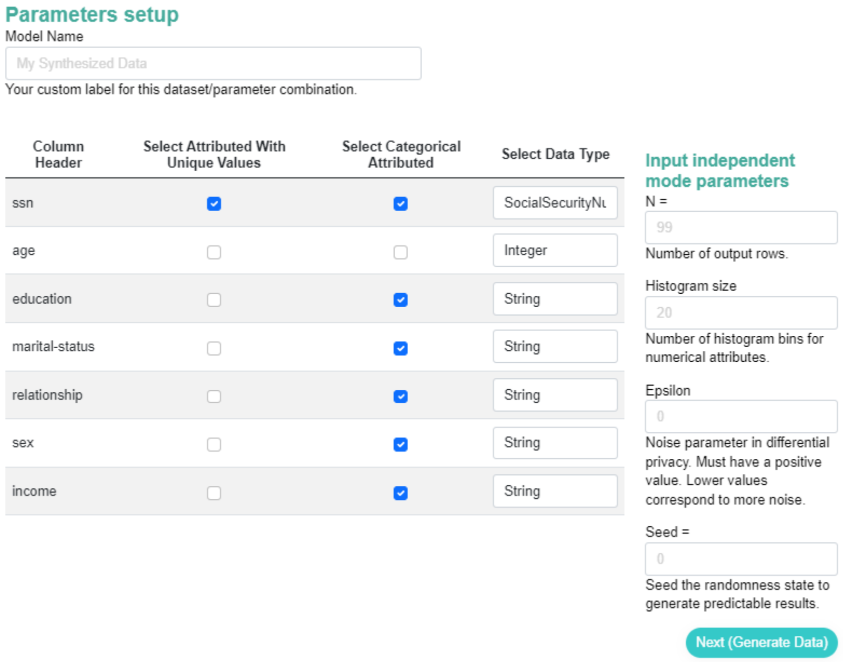
Seed of the random number generator. The same seed value will always generate the same results.

Figure 5: Random Mode

2.1.2 Independent Attribute Mode

In the **Independent Attribute mode**, the generator performs a frequency-based (number of occurrences) estimation of the distribution of attributes. This ensures the statistical distribution of data in the input (de-identified) data set is retained in the synthetic data set. The precision of how well the distribution of numerical attributes is captured in the generated data set can be refined with the **Histogram Size** parameter that is set to 20 by default. As each attribute is assumed to be independent of others, any correlation between attributes is ignored in this mode of data generation.

In the Independent Attribute mode, the **Epsilon** parameter can be used for generating privacy-preserving synthetic data sets. The Epsilon parameter value can vary between 0 and 1, with lower values (closer to zero), enforcing higher levels of privacy. Set the Epsilon value to 0 for turning off differential privacy. See Figure 6.



Parameters setup

Model Name

My Synthesized Data

Your custom label for this dataset/parameter combination.

Column Header	Select Attributed With Unique Values	Select Categorical Attributed	Select Data Type
ssn	☑	☑	SocialSecurityNu
age	☐	☐	Integer
education	☐	☑	String
marital-status	☐	☑	String
relationship	☐	☑	String
sex	☐	☑	String
income	☐	☑	String

Input independent mode parameters

N =

99

Number of output rows.

Histogram size

20

Number of histogram bins for numerical attributes.

Epsilon

0

Noise parameter in differential privacy. Must have a positive value. Lower values correspond to more noise.

Seed =

0

Seed the randomness state to generate predictable results.

Figure 6: Epsilon Parameter

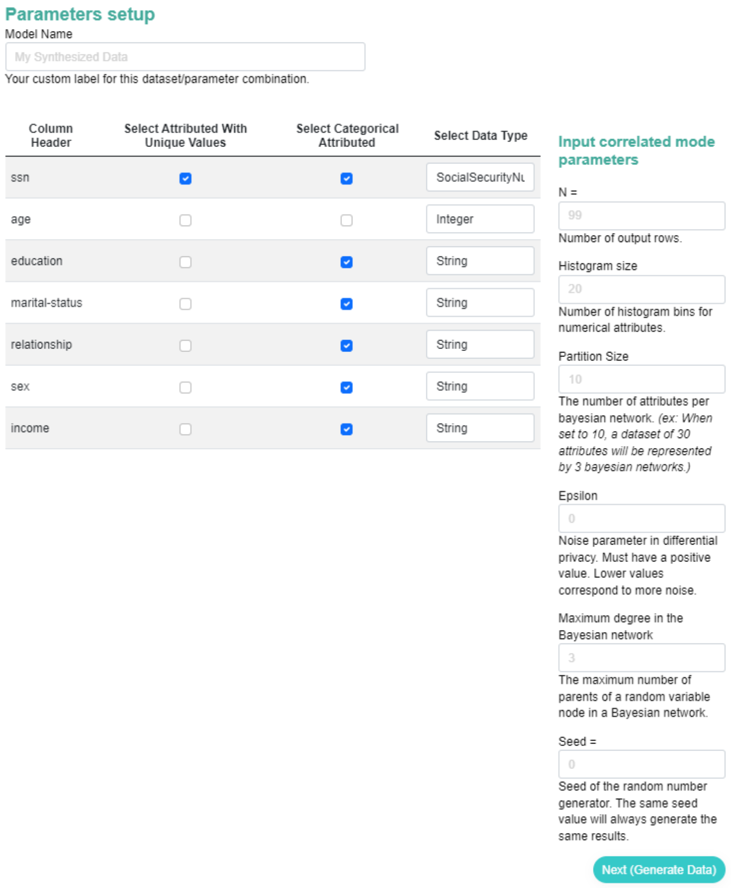
2.1.3 Correlated Attribute Mode

Correlated Attribute mode is the slowest but most accurate synthetic data generation mode. In addition to retaining the statistical distribution of data, the generated data also maintains any existing statistical correlation between categorical attributes. The generator uses a Bayesian Network (BN) to model the relationship between correlated attributes in the input dataset. Depending on the extent and number of correlated attributes, and total number of categorical attributes in the input data set, data generation times may exceed a reasonable time limit.

The **Maximum Degree** parameter can be lowered to 2 or 1 from the default value of 3 to reduce compute time.

For data with many columns, computation of the network is divided into partitions. The size of these partitions can be controlled with the **Partition Size** parameter, serving as an optional accuracy-speed tradeoff.

The differential privacy parameter settings for the Independent Attribute mode are also applicable and relevant to the Correlated Attribute mode. See Figure 7.



Parameters setup

Model Name

My Synthesized Data

Your custom label for this dataset/parameter combination.

Column Header	Select Attributed With Unique Values	Select Categorical Attributed	Select Data Type
ssn	☑	☑	SocialSecurityNu
age	□	□	Integer
education	□	☑	String
marital-status	□	☑	String
relationship	□	☑	String
sex	□	☑	String
income	□	☑	String

Input correlated mode parameters

N =

99

Number of output rows.

Histogram size

20

Number of histogram bins for numerical attributes.

Partition Size

10

The number of attributes per bayesian network. *(ex: When set to 10, a dataset of 30 attributes will be represented by 3 bayesian networks.)*

Epsilon

0

Noise parameter in differential privacy. Must have a positive value. Lower values correspond to more noise.

Maximum degree in the Bayesian network

3

The maximum number of parents of a random variable node in a Bayesian network.

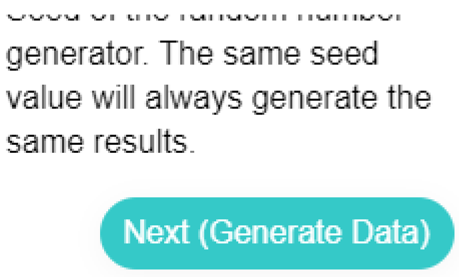
Seed =

0

Seed of the random number generator. The same seed value will always generate the same results.

Figure 7: Correlated Attribute Mode

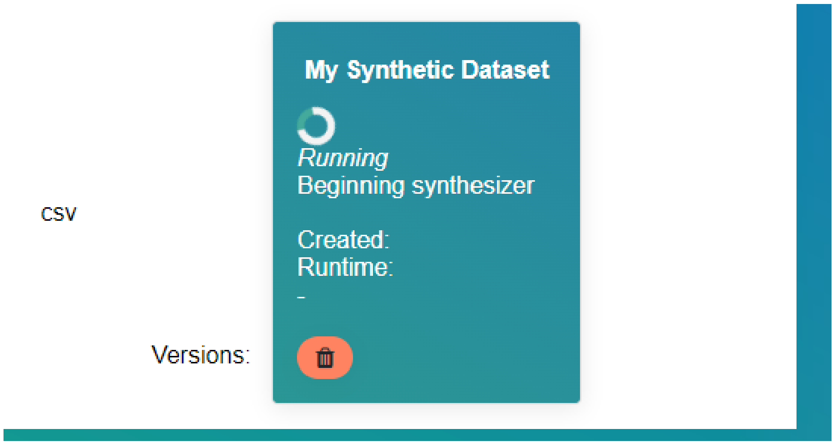
After all the parameters have been set for the chosen data generation mode (Correlated Attribute Mode chosen below), click the **Next** (Generate Data) button. See Figure 8.



Seed of the random number generator. The same seed value will always generate the same results.

Figure 8: Next Button to Generate Data

The system will redirect you to the Home page, where your uploaded dataset, as well as any prior uploads, are displayed. Shown as a horizontal row is the file you uploaded, as well as the label you assigned it. Below this, are various runs of the file, each with their own configurations as selected on the **Parameter Settings** page navigated to earlier within these steps. See Figure 9.



csv

My Synthetic Dataset

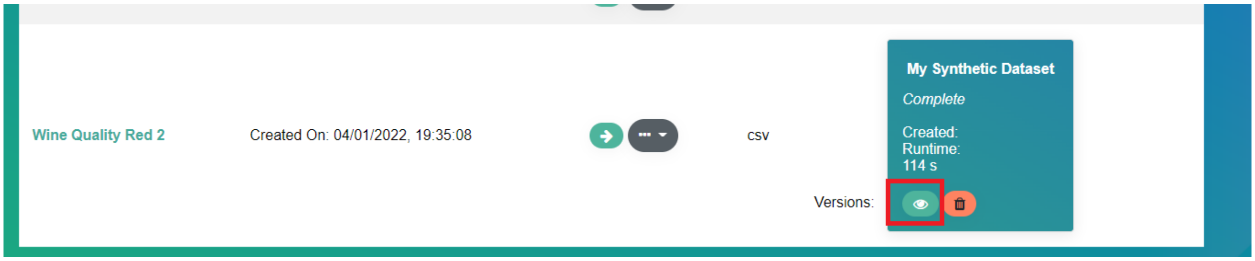
Running
Beginning synthesizer

Created:
Runtime:
-

Versions:

Figure 9: My Synthetic Dataset

Additionally, running datasets will provide information on duration of processing and details on the process currently running. A moderately sized dataset should take a few seconds to complete in the random and independent modes, and around one minute to complete in correlated attribute mode. Once complete, an “eyeball” icon will appear - the synthetic data generation is now complete and a full input (de-identified) to output (synthetic) data set comparison report is displayed on the screen. See Figure 10.



Wine Quality Red 2

Created On: 04/01/2022, 19:35:08

csv

My Synthetic Dataset

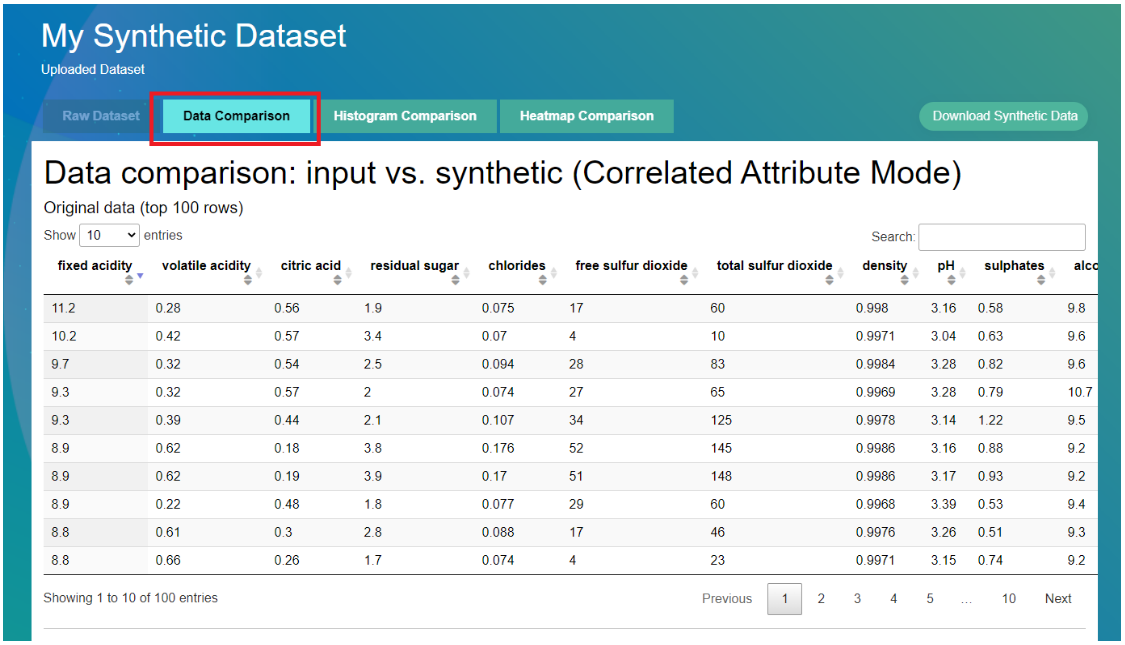
Complete

Created:
Runtime:
114 s

Versions:

Figure 10: Indicator Synthetic Data Generation Completed

The **Data Comparison** option, selected by default, displays the first 100 rows from the original data set and the generated synthetic data set in a paginated format. Inspect the synthetic data set to see if the any rows from the original, sensitive data set can be identified. See Figure 11.



Data comparison: input vs. synthetic (Correlated Attribute Mode)

Original data (top 100 rows)

Show 10 entries Search:

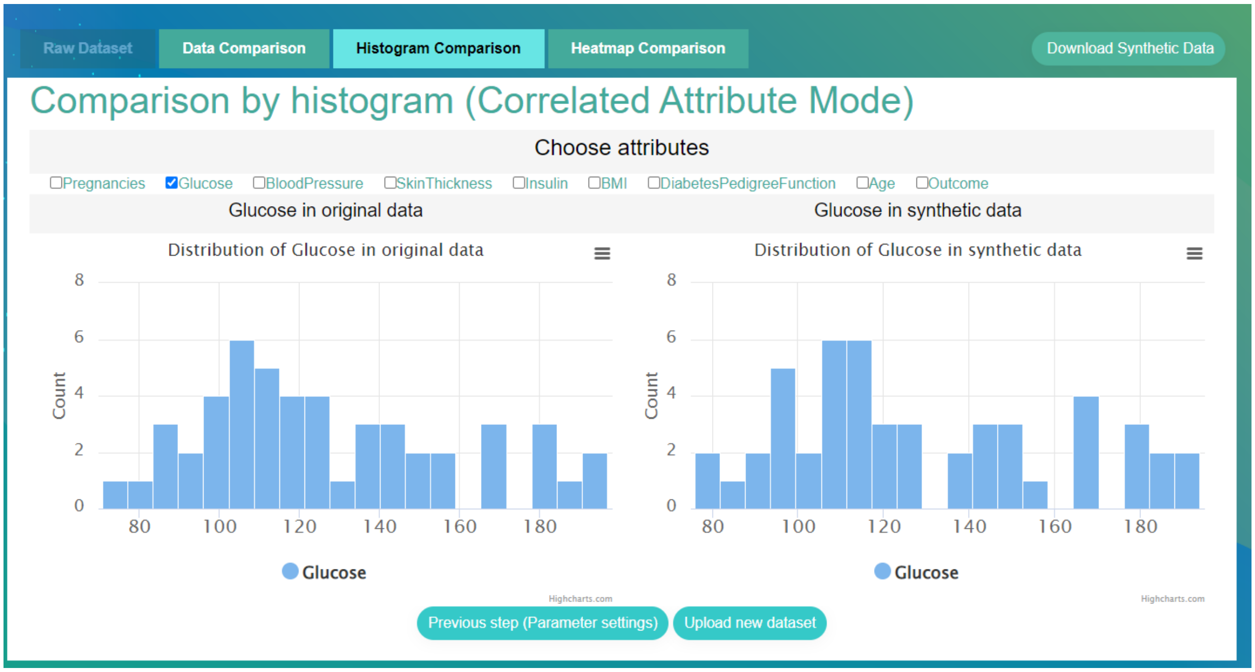
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alco
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8
10.2	0.42	0.57	3.4	0.07	4	10	0.9971	3.04	0.63	9.6
9.7	0.32	0.54	2.5	0.094	28	83	0.9984	3.28	0.82	9.6
9.3	0.32	0.57	2	0.074	27	65	0.9969	3.28	0.79	10.7
9.3	0.39	0.44	2.1	0.107	34	125	0.9978	3.14	1.22	9.5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2
8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4
8.8	0.61	0.3	2.8	0.088	17	46	0.9976	3.26	0.51	9.3
8.8	0.66	0.26	1.7	0.074	4	23	0.9971	3.15	0.74	9.2

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 … 10 Next

Figure 11: Data Comparison Option

Click the **Histogram Comparison** menu option to compare the estimated per-attribute distribution of the de-identified data set with those from the synthetic data set. See Figure 12.



Comparison by histogram (Correlated Attribute Mode)

Choose attributes

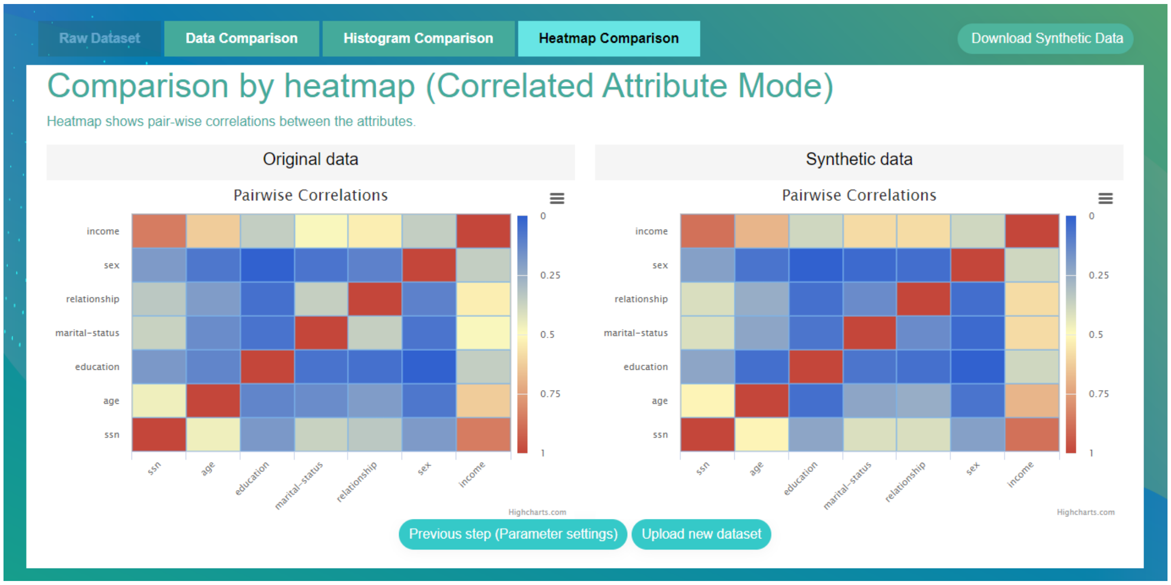
□Pregnancies ☑Glucose □BloodPressure □SkinThickness □Insulin □BMI □DiabetesPedigreeFunction □Age □Outcome

Glucose in original data

Glucose in synthetic data

Figure 12: Histogram Comparison

Click the **Heatmap Comparison** menu option to view a pair-wise attribute correlation comparison for the de-identified data set and the synthetic data set. The color-coded, heatmap index value is a numerical measure of the strength of correlation between attributes. See Figure 13.



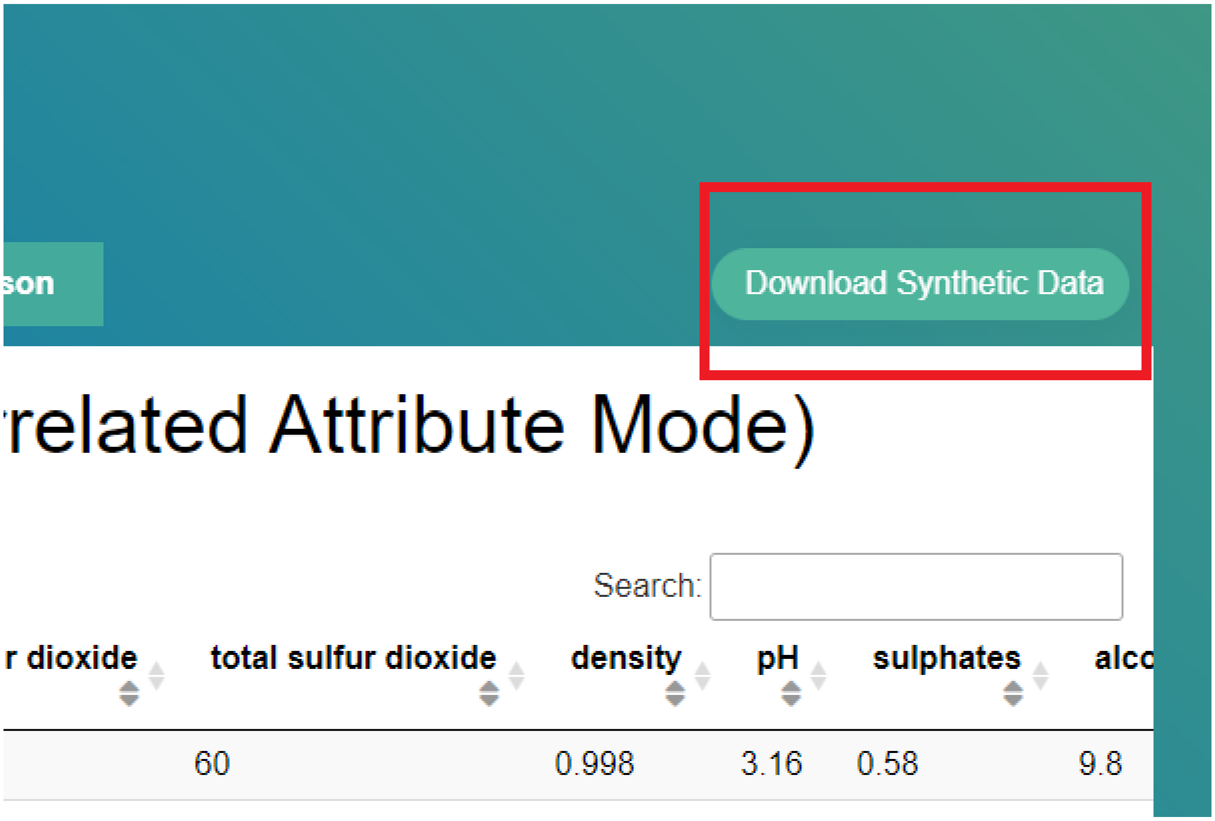
Comparison by heatmap (Correlated Attribute Mode)

Heatmap shows pair-wise correlations between the attributes.

Figure 13: Heatmap Comparison

Following an inspection of the comparison reports, download the synthetic data set (.csv file) to a local directory location by clicking the **Download Synthetic Data** link in the left menu bar.

Optionally, clicking the **Download Dataset Data** link in the left menu bar will download the dataset description file describing the statistical model used for synthetic data generation. See Figure 14.



rrelated Attribute Mode)

r dioxide	total sulfur dioxide	density	pH	sulphates	alco
	60	0.998	3.16	0.58	9.8

Figure 14: Download Synthetic Data

NOTE: At every step of the data generation process, the user has the option to return to the previous step or completely exit the process and start over with a new dataset.

3 REPORTING ISSUES AND TECHNICAL SUPPORT

In the event you encounter a technical issue, please submit a ticket to our service desk at mdaca_support@spinsys.com.