Model Report for `insurance_dataset:tabular`

Generated on 09 Nov 2024, 17:40 • 1,338 original samples, 1,338 synthetic samples • Generator: ef5e2c4c-1538-4c1c-8be7-8ececd55baf9

Accuracy ⓘ

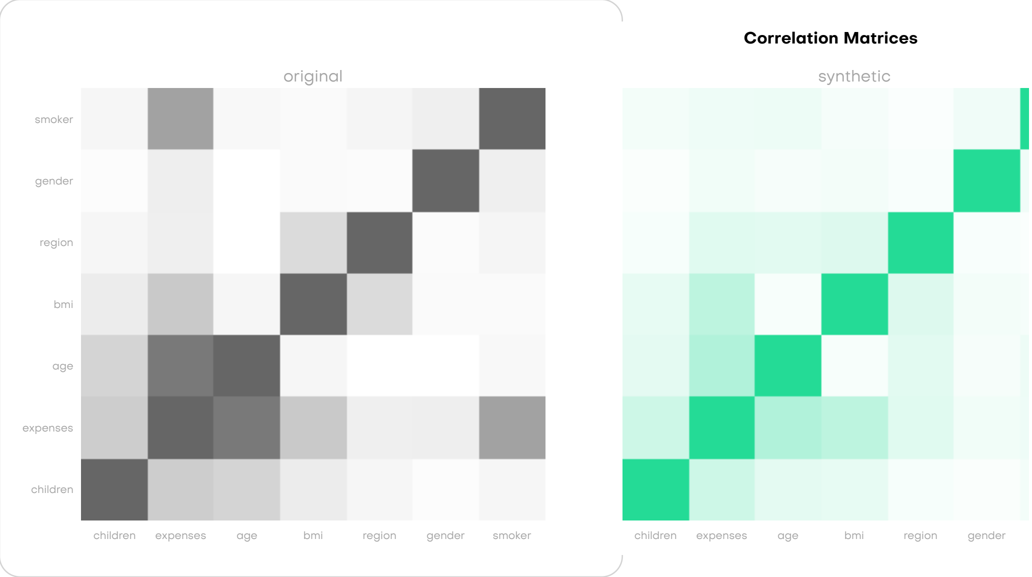
92.6%
(94.4%)

Univariate	96.3%
Bivariate	88.8%

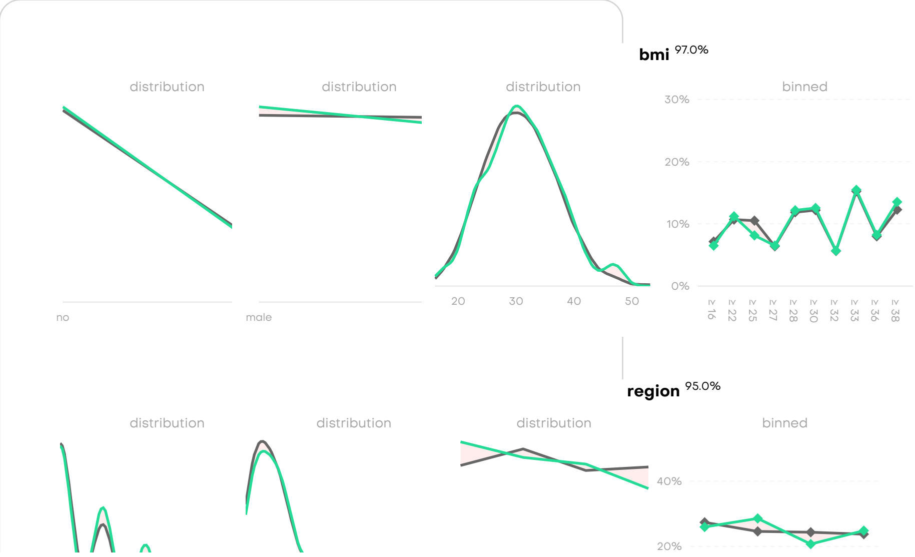
Similarity ⓘ

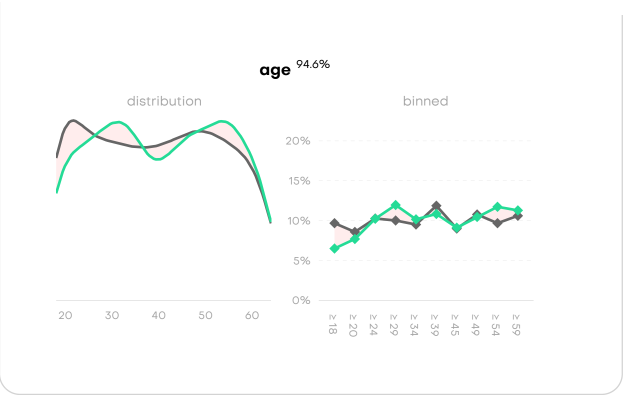
Cosine Similarity	0.99991 (0.99989)
Discriminator AUC	50.4% (52.7%)

Correlations

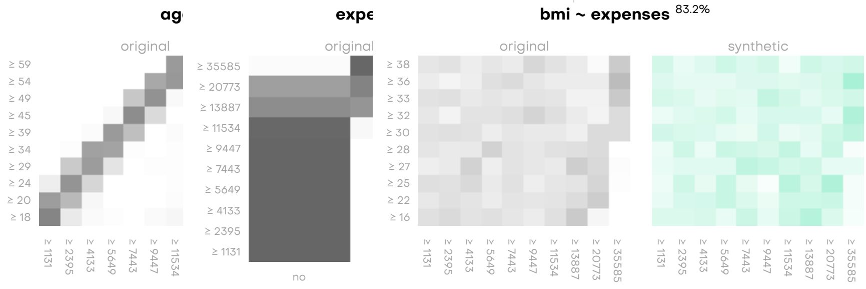


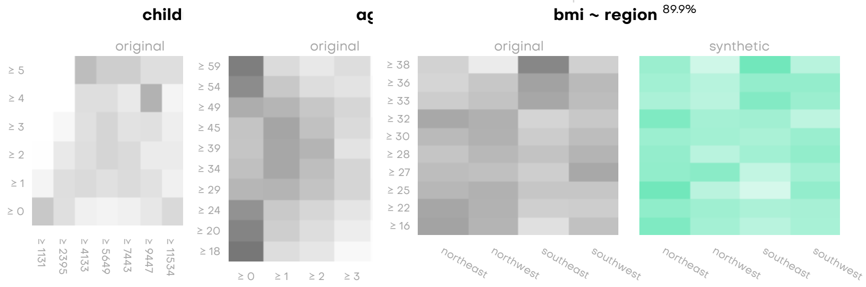
Univariate Distributions

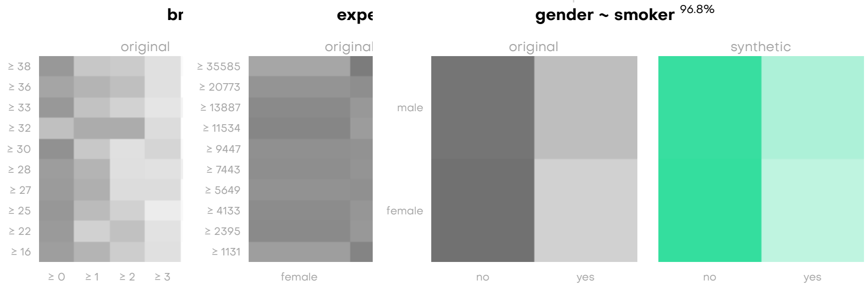


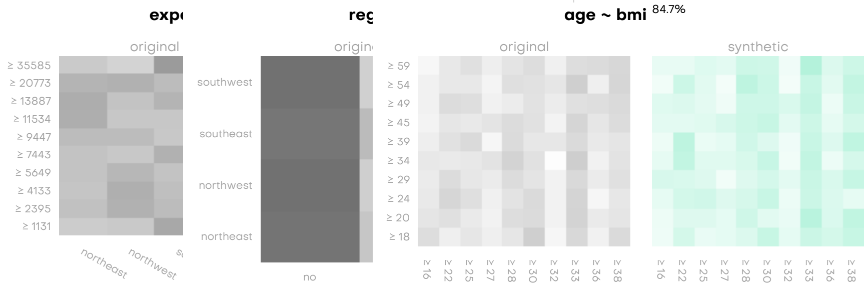


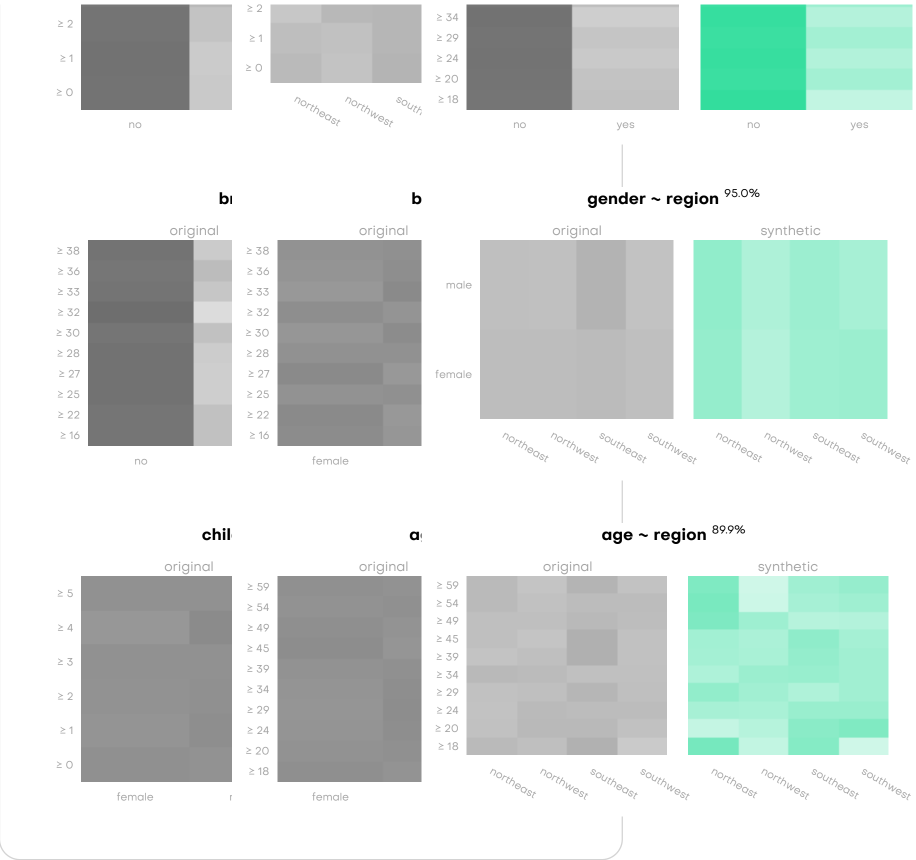
Bivariate Distributions





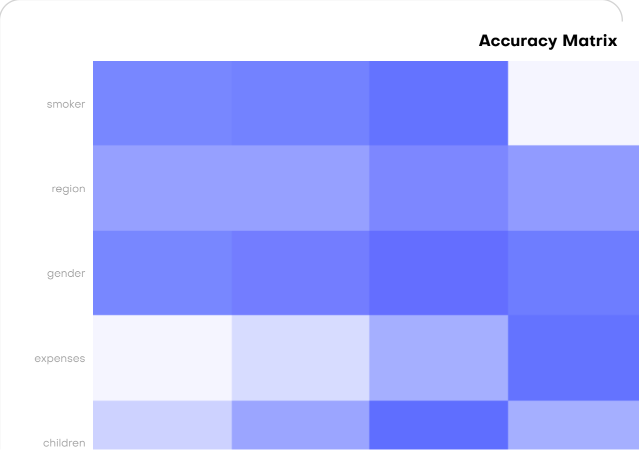


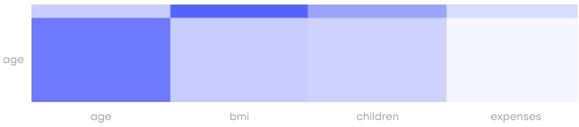




Accuracy

Column	Univariate	Bivariate
smoker	98.5%	90.5%
gender	97.7%	94.9%
bmi	97.0%	89.2%
children	96.1%	91.1%
expenses	95.5%	80.1%
region	95.0%	92.1%
age	94.6%	83.4%
Total	**96.3%**	**88.8%**

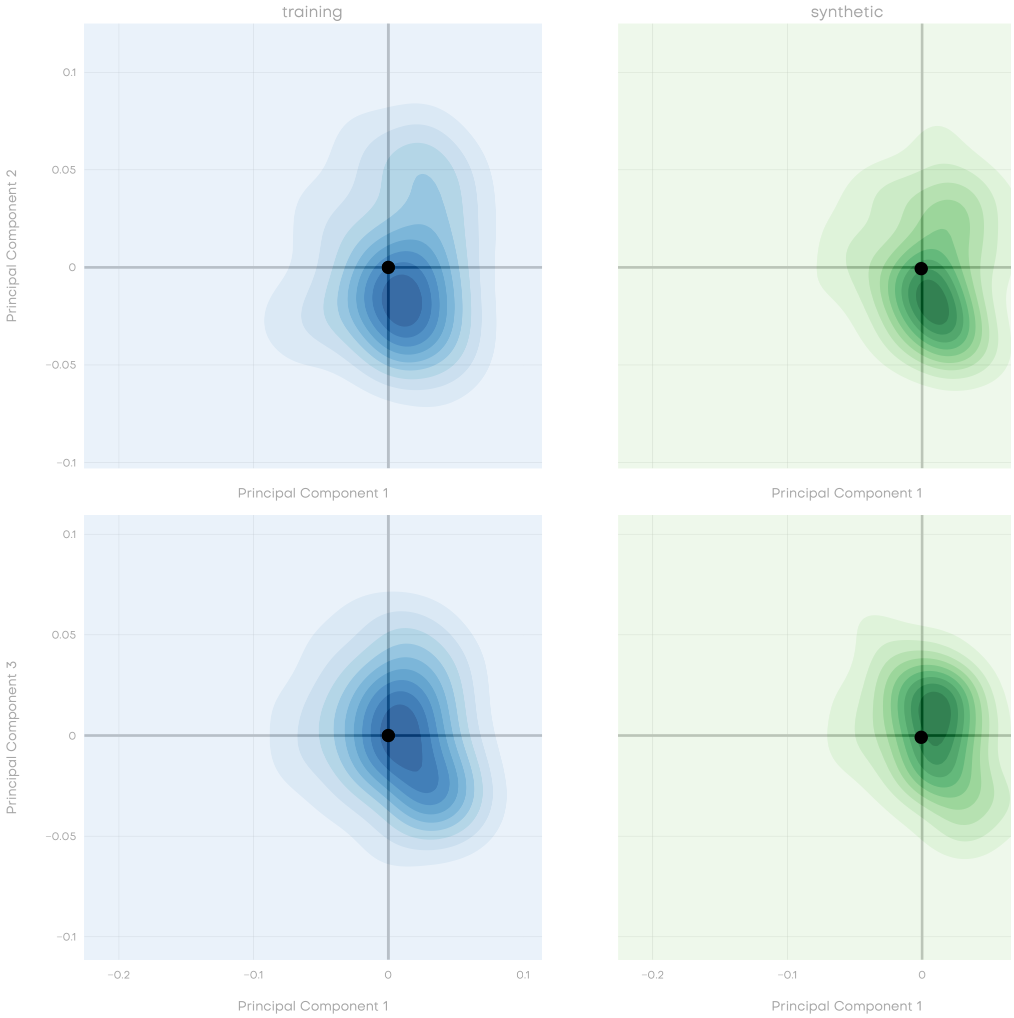




Explainer

Accuracy of synthetic data is assessed by comparing the distributions of the synthetic (shown in green) and the original data (shown in gray). For each distribution plot we sum up the deviations across all categories, to get the so-called total variation distance (TVD). The reported accuracy is then simply reported as 100% - TVD. These accuracies are calculated for all univariate and bivariate distributions. A final accuracy score is then calculated as the average across all of these.

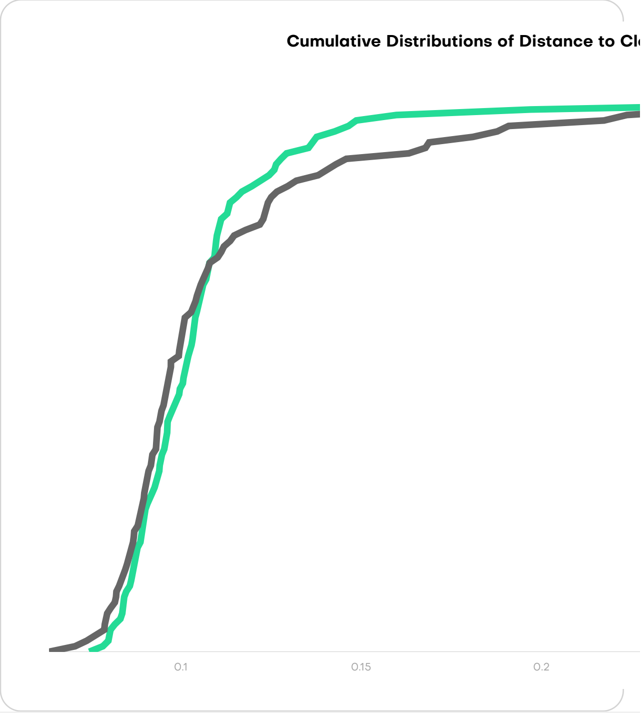
Similarity



These plots show the first 3 principal components of training samples, synthetic samples, and holdout samples within the embedding space. The black dots visualize the centroids of the respective samples. The similarity metric then measures the cosine similarity between these centroids. We expect the cosine similarity to be close to 1, indicating that the synthetic samples are as similar to the training samples as the holdout samples are.

Distances

	Synthetic vs. Training Data	(Synthetic vs. Holdout Data)
Identical Matches	0.0%	(0.0%)
Average Distances	0.105	(0.109)



Explainer

Synthetic data shall be as close to the original training samples, as it is close to original holdout samples, which serve us as a reference. This can be asserted empirically by measuring distances between synthetic samples to their closest original samples, whereas training and holdout sets are sampled to be of equal size. For the visualization above, the distances of synthetic samples to the training samples are displayed in green, and the distances of synthetic found in training just as well as in holdout samples.