SUPPLEMENTARY MATERIAL: MASKED LANGUAGE MODELING BECOMES CONDITIONAL DENSITY ESTIMATION FOR TABULAR DATA SYNTHESIS

A Appendix / supplemental material

A.1 Proof of Proposition 1

Proof. Without loss of generality, let σ be an identity permutation. We equally space grid points on the interval $[0, 1]$ such that bins are represented as $[b_{l-1}, b_l), l = 1, \cdots, L$, where h denotes the bin width such that $Lh = 1$. We consider arbitrary $j \in I_C$ and $\mathbf{x} \in \mathbb{R}^p$, where $\mathbf{x}$ is given and fixed. For notational simplicity, let $\mathbf{x}_{-j} \coloneqq (\mathbf{x}_1, \cdots, \mathbf{x}_{j-1})$.

Suppose that $\hat{F}_j(\mathbf{x}_j) \in [b_{l-1}, b_l)$. By the mean value theorem, there exists $u^* \in (b_{l-1}, b_l)$ such that

$$h \cdot c_j^*(u^*|\mathbf{x}_{-j}) = \int_{b_{l-1}}^{b_l} c_j^*(v|\mathbf{x}_{-j})dv = \pi_{jl}^*(\mathbf{x}_{-j}).$$

And we have

$$\begin{aligned}
& \left| c_j^*(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - \hat{c}_j(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}; \theta) \right| \\
= & \left| c_j^*(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - \frac{\pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta)}{h} \right| \\
\leq & \left| c_j^*(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - c_j^*(u^*|\mathbf{x}_{-j}) \right| + \left| c_j^*(u^*|\mathbf{x}_{-j}) - \frac{\pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta)}{h} \right| \\
\leq & |\hat{F}(\mathbf{x}_j) - u^*| \left| \frac{c_j^*(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - c_j^*(u^*|\mathbf{x}_{-j})}{\hat{F}(\mathbf{x}_j) - u^*} \right| + \left| \frac{\pi_{jl}^*(\mathbf{x}_{-j})}{h} - \frac{\pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta)}{h} \right| \\
\leq & hK_j + \frac{1}{h} \left| \pi_{jl}^*(\mathbf{x}_{-j}) - \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right| \quad \text{(by Assumption 2)} \\
\leq & hK_j + \frac{1}{h} \sum_{l=1}^{L} \left| \pi_{jl}^*(\mathbf{x}_{-j}) - \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right|.
\end{aligned}$$

where the first equality follows from the definition of $\hat{c}_j$.

By the Pinsker's inequality,

$$\begin{aligned}
& \frac{1}{2} \sum_{l=1}^{L} \left| \pi_{jl}^*(\mathbf{x}_{-j}) - \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right| \\
\leq & \frac{1}{\sqrt{2}} \left(\sum_{l=1}^{L} \pi_{jl}^*(\mathbf{x}_{-j}) \log \pi_{jl}^*(\mathbf{x}_{-j}) - \sum_{l=1}^{L} \pi_{jl}^*(\mathbf{x}_{-j}) \log \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right)^{1/2} \\
\leq & \frac{1}{\sqrt{2}} \left(\sum_{l=1}^{L} \pi_{jl}^*(\mathbf{x}_{-j}) \log \pi_{jl}^*(\mathbf{x}_{-j}) - \mathbb{E}_{\mathbf{y}_j|\mathbf{x}_{-j}} \left[\sum_{l=1}^{L} \mathbb{I}(\mathbf{y}_j = l) \log \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right] \right)^{1/2},
\end{aligned}$$

where $\mathbf{y}_j|\mathbf{x}_{-j}$ is a random variable having a categorical distribution such that $\Pr(\mathbf{y}_j = l|\mathbf{x}_{-j}) = \Pr(g(\mathbf{x}; \hat{F})_j = l|\mathbf{x}_{-j}) = \pi^*_{jl}(\mathbf{x}_{-j})$ for all $l \in [L]$.

Then,

$$\begin{aligned} & \left| c^*_j(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - \hat{c}_j(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}; \theta) \right| \\ \leq \quad & hK_j + \frac{\sqrt{2}}{h} \left(\sum_{l=1}^{L} \pi^*_{jl}(\mathbf{x}_{-j}) \log \pi^*_{jl}(\mathbf{x}_{-j}) - \mathbb{E}_{\mathbf{y}_j|\mathbf{x}_{-j}} \left[\sum_{l=1}^{L} \mathbb{I}(\mathbf{y}_j = l) \log \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right] \right)^{1/2}. \quad (1) \end{aligned}$$

The total variation distance between $p^*_j(\cdot|\mathbf{x}_{-j})$ and $\hat{p}_j(\cdot|\mathbf{x}_{-j}; \theta)$ is written as

$$\begin{aligned} & \mathrm{TV}\Big(p^*_j(\cdot|\mathbf{x}_{-j}), \hat{p}_j(\cdot|\mathbf{x}_{-j}; \theta) \Big) \\ = \quad & \frac{1}{2} \int_{\mathbb{R}} \Big| p^*_j(\mathbf{x}_j|\mathbf{x}_{-j}) - \hat{p}_j(\mathbf{x}_j|\mathbf{x}_{-j}; \theta) \Big| d\mathbf{x}_j \\ = \quad & \frac{1}{2} \int_{\mathbb{R}} \left| c^*_j(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}) - \hat{c}_j(\hat{F}(\mathbf{x}_j)|\mathbf{x}_{-j}; \theta) \right| \cdot \hat{p}_j(\mathbf{x}_j) d\mathbf{x}_j \quad \text{(by Assumption 1)} \\ \leq \quad & \frac{hK_j}{2} + \frac{1}{\sqrt{2}h} \left(\sum_{l=1}^{L} \pi^*_{jl}(\mathbf{x}_{-j}) \log \pi^*_{jl}(\mathbf{x}_{-j}) - \mathbb{E}_{\mathbf{y}_j|\mathbf{x}_{-j}} \left[\sum_{l=1}^{L} \mathbb{I}(\mathbf{y}_j = l) \log \pi_{jl}(g(\mathbf{x}; F) \odot \mathbf{m}^{(j)}; \theta) \right] \right)^{1/2} \\ = \quad & \frac{hK_j}{2} + \frac{\sqrt{Bias(\theta)}}{\sqrt{2}h} \\ = \quad & \frac{K_j}{2L} + \frac{\sqrt{Bias(\theta)}}{\sqrt{2}/L}, \end{aligned}$$

where the inequality holds for (1).

The proof is complete. □

A.2 Proof of Proposition 2

Proof. Since the data is MAR, $p(\mathbf{r}|\mathbf{x}) = p(\mathbf{r}|\mathbf{x}_{obs})$ and

$$\begin{aligned} p(\mathbf{m}, \mathbf{r}|\mathbf{x}) &= p(\mathbf{r}|\mathbf{x}) \cdot p(\mathbf{m}|\mathbf{x}) = p(\mathbf{r}|\mathbf{x}) \cdot p(\mathbf{m}) \\ &= p(\mathbf{r}|\mathbf{x}_{obs}) \cdot p(\mathbf{m}) = p(\mathbf{m}, \mathbf{r}|\mathbf{x}_{obs}). \end{aligned}$$

The proof is complete. □

A.3 Dataset Descriptions

Download links.

- `abalone` [23]: `https://archive.ics.uci.edu/dataset/1/abalone`
- `banknote` [18]: `https://archive.ics.uci.edu/dataset/267/banknote+authentication`
- `breast` [32]: `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`
- `concrete` [34]: `https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength`
- `covertype` [2]: `https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset`
- `kings` (CC0: Public Domain): `https://www.kaggle.com/datasets/harlfoxem/housesalesprediction`
- `letter` [28]: `https://archive.ics.uci.edu/dataset/59/letter+recognition`
- `loan` (CC0: Public Domain): `https://www.kaggle.com/datasets/teertha/personal-loan-modeling`
- `redwine` [6]: `https://archive.ics.uci.edu/dataset/186/wine+quality`
- `whitewine` [6]: `https://archive.ics.uci.edu/dataset/186/wine+quality`

Dataset	Train/Test Split	#continuous	#categorical	Classification Target
`abalone`	3.3K/0.8K	7	2	`Rings`
`banknote`	1.1K/0.3K	4	1	`class`
`breast`	0.5K/0.1K	30	1	`Diagnosis`
`concrete`	0.8K/0.2K	8	1	`Age`
`covtype`	0.46M/0.12M	10	1	`Cover_Type`
`kings`	17.3K/4.3K	11	7	`grade`
`letter`	16K/4K	16	1	`lettr`
`loan`	4K/1K	5	6	`Personal Loan`
`redwine`	1.3K/0.3K	11	1	`quality`
`whitewine`	3.9K/1K	11	1	`quality`

Table 1: **Description of datasets.** #continuous represents the number of continuous and ordinal variables. #categorical denotes the number of categorical variables. The 'Classification Target' refers to the variable used as the response variable in a classification task to evaluate machine learning utility.

A.4 Experimental Settings for Reproduction

- We run experiments using NVIDIA A10 GPU, and our experimental codes are available with `pytorch`.
- In practice, for all $j \in I_C$, we estimate $\hat{F}_j$ using empirical measure. In the presence of missing data, $\hat{F}_j$ is estimated solely using the observed dataset (refer to [4]).

Hyper-parameters of MaCoDE: As shown in Section A.9, MaCoDE consistently generates high-quality synthetic data without requiring an extensive hyperparameter tuning process, unlike methods such as [14, 15, 13]. This demonstrates the generalizability of our proposed model to various tabular datasets. **Our implementation codes for the proposed model, MaCoDE, are provided in the supplementary material. For all tabular datasets, we applied the following hyperparameters uniformly without any additional tuning**:

- epochs: 500
- batch size: 1024
- learning rate: 0.001 (with AdamW optimizer [19] with 0.001 weight decay parameter)
- the number of bins: $L = 50$
- Transformer encoder dimension: 128
- Transformer encoder #heads: 4
- Transformer encoder #layer: 2
- Transformer encoder dropout ratio: 0.0

A.4.1 Details of Implementing Baseline Models

To compare our proposed method with baseline models, we performed experiments in a manner similar to [36]. We adjusted the hidden or latent dimensions across various methods, ensuring that the number of trainable parameters is comparable. Since each model employs distinct neural network architectures, it would be misleading to compare their performances using the same network configuration. It's worth noting that model size typically reflects the performance of the baseline models. Under these conditions, we reproduced the baseline methods using their official codes and default configurations (except for the hidden or latent dimensions). **Our reproduced codes for the baseline models are provided in the supplementary material.** Below are the detailed implementations of the baseline methods:

- CTGAN and TVAE [33]: We follow the implementations provided in the official repository[*] for CTGAN and TVAE. We adopt the default hyperparameters specified in the module. To ensure fairness in comparison by aligning model sizes, we adjust the latent dimension of CTGAN and TVAE to 100.
- CTAB-GAN and CTAB-GAN+ [40, 41]: We follow the implementations provided in the official repository[†]. The latent dimension is set to 100, and the maximum number of clusters is set to 10. For the specification of feature types, we use three categories: continuous, integer, and categorical variables.
- DistVAE [1]: We follow the implementations provided in the official repository[‡]. We changed the latent dimension to 100 to increase the model size.
- TabDDPM [14]: We utilized the TabDDPM module in `synthcity.Plugins` [§] for synthetic data generation. This module follows the implementations provided in the official repository.
- TabMT [8]: Rather than using K-means clustering, we utilize the Gaussian Mixture Model (GMM) to discretize continuous columns to preserve the original continuous domain. We determine the optimal number of clusters for the GMM, ranging from 2 to 10, based on the Bayesian Information Criterion (BIC). During the generation of continuous columns, we initially predict the component label and then sample from the selected Gaussian component.
- MICE [31]: We employed the `IterativeImputer` package from `Scikit-learn` for multiple imputation using chained equations. Following the authors' experiments, a `max_iter` range of 10 to 20 was considered sufficient for convergence, and we adopted this setting for our experiments. Additionally, to introduce randomness, we set `imputation_order` to `random`, which randomly selects a variable for imputation in each iteration. The remaining parameters were left at their default values to maintain the integrity of the MICE implementation.
- GAIN [35]: We follow the implementations provided in the official repository[¶]. As the paper does not explicitly discuss the separate handling of categorical and continuous variables, the code treats them simultaneously. Consequently, a rounding process is employed to handle categorical variables afterward.
- missMDA [12]: We utilized the `missMDA` package[‖] in `R` for multiple imputation.
- VAEAC [10]: We follow the implementations provided in the official repository[**]. The authors provided hyperparameters that adequately address both continuous and categorical variables, so we used these without further modification during model fitting.
- MIWAE [20]: The implemented MIWAE code in the official repository[††] was designed for continuous variables only. To accommodate heterogeneous tabular datasets, we treated the conditional distribution of categorical columns as categorical distributions and employed cross-entropy loss for reconstruction. For comparison with not-MIWAE, we set the latent dimension to $p - 1$.
- not-MIWAE [9]: The implemented not-MIWAE code in the official repository[‡‡] also focused on continuous variables exclusively. To handle categorical variables, we made the same modifications as in MIWAE. For comparison with MIWAE, we set the latent dimension to $p - 1$. Training was conducted for 100K steps, consistent with the official implementations.

[*] `https://github.com/sdv-dev/CTGAN`
[†] `https://github.com/Team-TUD/CTAB-GAN`, `https://github.com/Team-TUD/CTAB-GAN-Plus`
[‡] `https://github.com/an-seunghwan/DistVAE`
[§] `https://github.com/vanderschaarlab/synthcity`
[¶] `https://github.com/jsyoon0823/GAIN/tree/master`
[‖] `https://cran.r-project.org/web/packages/missMDA/index.html`
[**] `https://github.com/tigvarts/vaeac`
[††] `https://github.com/pamattei/miwae`
[‡‡] `https://github.com/nbip/notMIWAE`

- EGC [39]: We utilized the `gcimpute` package[§§] to implement EGC. The model is free of hyperparameters.

Tasks	Model	Description
`Regression`	Random Forest	Package: `sklearn.ensemble.RandomForestRegressor`, setting: random_state=0, defaulted values
`Classification`	Logistic Regression	Package: `sklearn.linear_model.LogisticRegression`, setting: random_state=0, max_iter=1000, defaulted values
	Gaussian Navie Bayes	Package: `sklearn.naive_bayes.GaussianNB`, setting: defaulted values
	K-Nearest Neighbors	Package: `sklearn.neighbors.KNeighborsClassifier`, setting: defaulted values
	Decision Tree	Package: `sklearn.tree.DecisionTreeClassifier`, setting: random_state=0, defaulted values
	Random Forest	Package: `sklearn.ensemble.RandomForestClassifier`, setting: random_state=0, defaulted values

Table 2: **Regressor and classifier used to evaluate synthetic data quality in machine learning utility.** The names of all parameters used in the description are consistent with those defined in corresponding packages.

A.5 Evaluation Procedure of Q1

Regression performance (SMAPE).

1. Train a synthesizer using the real training dataset.
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train a machine learning model (Random Forest regressor) using the synthetic dataset, where each continuous column serves as the regression target variable.
4. Assess regression prediction performance by averaging the SMAPE values from the test dataset for each Random Forest regressor trained on the continuous columns.

Classification performance (F_1).

1. Train a synthesizer using the real training dataset.
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train machine learning models (Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors classifier, Decision Tree classifier, and Random Forest classifier) using the synthetic dataset (see Table 1 for the classification target variable and refer to Table 2 for detailed configuration).
4. Assess classification prediction performance by averaging the F_1 values from the test dataset from five different classifiers.

Model selection performance (Model).

1. Train a synthesizer using the real training dataset.
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train machine learning models (Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors classifier, Decision Tree classifier, and Random Forest classifier) using both the real training dataset and the synthetic dataset (see Table 1 for the classification target variable and refer to Table 2 for detailed configuration).
4. Evaluate the classification performance (AUROC) of all trained classifiers on the test dataset.
5. Assess model selection performance by comparing the AUROC rank orderings of classifiers trained on the real training dataset and those trained on the synthetic dataset using Spearman's Rank Correlation.

[§§] `https://github.com/udellgroup/gcimpute/tree/master`

Feature selection performance (Feature).

1. Train a synthesizer using the real training dataset.
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train a Random Forest classifier using both the real training dataset and the synthetic dataset (see Table 1 for the classification target variable and refer to Table 2 for detailed configuration).
4. Determine the rank-ordering of important features for both classifiers.
5. Assess feature selection performance by comparing the feature importance rank orderings of classifiers trained on the real training dataset and those trained on the synthetic dataset using Spearman's Rank Correlation.

A.6 Evaluation Procedure of Q2

Regression performance (SMAPE).

1. For each random seed, we generate the mask and train MaCoDE using a masked training dataset (i.e., incomplete dataset).
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train a machine learning model (Random Forest regressor) using the synthetic dataset, where each continuous column serves as the regression target variable.
4. Assess regression prediction performance by averaging the SMAPE values from the test dataset for each Random Forest regressor trained on the continuous columns.

Classification performance (F_1).

1. For each random seed, we generate the mask and train MaCoDE using a masked training dataset (i.e., incomplete dataset).
2. Generate a synthetic dataset with the same size as the real training dataset.
3. Train machine learning models (Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors classifier, Decision Tree classifier, and Random Forest classifier) using the synthetic dataset (see Table 1 for the classification target variable and refer to Table 2 for detailed configuration).
4. Assess classification prediction performance by averaging the F_1 values from the test dataset from five different classifiers.

A.7 Evaluation of Q3

A.7.1 Related Works

Data missingness is a common challenge in research and practical analysis, categorized into three primary missing mechanisms: (1) Missing Completely at Random (MCAR), (2) Missing at Random (MAR), and (3) Missing Not at Random (MNAR).

Under the **MCAR** mechanism, the reason for missingness has no relationship with any data, neither observed nor unobserved. In other words, the likelihood of data being missing is equal across all observations. The primary advantage of MCAR is that it does not introduce bias into the data analysis. However, despite this advantage, data missingness can still reduce the statistical power of the study because of the reduced sample size.

MAR occurs when the probability of missingness is related to the observed data but not the unobserved missing data. Essentially, even though the data is missing, the mechanism assumes that the missingness is explainable by other variables in the dataset. That is, the missingness can be modeled and imputed using the information available in the data, allowing for more accurate analyses despite the missingness.

Lastly, if the missingness is not specified by either MCAR or MAR, it becomes **MNAR**. The MNAR is the most challenging mechanism, as it implies that the missingness is related to the unobserved data itself. In this case, the missing data is systematically different from the observed data, which introduces bias if not properly accounted for. For example, patients with severe symptoms may be less likely to report their health status, making their data missing. MNAR requires sophisticated statistical methods to address, as ignoring or improperly handling it can lead to biased and unreliable results.

Recent methods employ the deep generative model to estimate a joint distribution and generate samples for imputations. GAN-based imputers such as GAIN [35] and MisGAN [17] adopt an adversarial learning approach to generate both missing entries and masking vectors. Other approaches like VAEAC [10], HI-VAE [24], and ReMasker [7] employ strategies that allow them to learn conditional distributions on arbitrary conditioning sets using the uniform masking strategy. MIWAE [20], based on the Importance Weighted Autoencoder [3], demonstrates that under mild conditions and the MAR assumption, the target likelihood can be approximated regardless of the imputation function. Additionally, not-MIWAE [9] extends MIWAE to handle cases where the data is under MNAR assumption by modeling missing entries as a latent variable. In parallel, the Optimal Transport (OT)-based method employs distributional matching, utilizing the 2-Wasserstein distance to compare distributions in both data and latent spaces [22, 38]. To handle mixed-type tabular datasets, [39] introduced EGC (extended Gaussian copula), which relies on a latent Gaussian distribution to support single and multiple imputations.

A.7.2 Evaluation Metrics

Algorithm 1 Evaluation procedure for multiple imputation [30, 27]

Input: Complete dataset $D = \{x_i\}_{i=1}^n$
Output: Bias, Coverage, and Confidence interval length

Target estimand: $Q^* = 1/n \sum_{i=1}^n \mathbb{I}(x_i > \bar{x})$, where $\bar{x} = 1/n \sum_{i=1}^n x_i$
for (random seeds) $s = 1, \cdots, S$ **do**
    Generate missing dataset $D^{(s)}$
    Training the imputation model using the dataset $D^{(s)}$
    Perform multiple imputation: $\hat{D}_m^{(s)}, m = 1, \cdots, M$
    for $m = 1, 2, \cdots, M$ **do**
        $\hat{Q}_m^{(s)} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i > \bar{x})$, where $\hat{D}_m^{(s)} = \{x_i\}_{i=1}^n$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
        $\hat{U}_m^{(s)} = \hat{Q}_m^{(s)}(1 - \hat{Q}_m^{(s)})/n$
    $\hat{Q}^{(s)} = \frac{1}{M} \sum_{m=1}^M \hat{Q}_m^{(s)}$
    $\hat{U}^{(s)} = \frac{1}{M} \sum_{m=1}^M \hat{U}_m^{(s)} + (M+1)/M \cdot \frac{1}{M-1} \sum_{m=1}^M (\hat{U}_m^{(s)} - \bar{U}^{(s)})^2$,
where $\bar{U}^{(s)} = \frac{1}{M} \sum_{m=1}^M \hat{U}_m^{(s)}$
Bias: $\frac{1}{S} \sum_{s=1}^S |\hat{Q}^{(s)} - Q^*|$
Coverage: $\frac{1}{S} \sum_{s=1}^S \mathbb{I}(Q^* \in (\hat{Q}^{(s)} \pm 1.96 \cdot \sqrt{\hat{U}^{(s)}}))$
Confidence interval length: $\frac{1}{S} \sum_{s=1}^S 2 \cdot 1.96 \cdot \sqrt{\hat{U}^{(s)}}$

For Q3, we selected the following multiple imputation models that can handle heterogeneous tabular datasets: MICE [31], GAIN [35], missMDA [12], VAEAC [10], MIWAE [20], not-MIWAE [9], and EGC [39].

We assess the effectiveness of multiple imputations by employing interval inference for the population mean, which was proposed by Rubin [27]. However, since obtaining the population mean is not feasible, we utilize the sample column-wise mean of the complete dataset as a parameter of interest [16, 37]. The evaluation procedure for multiple imputations is outlined in Algorithm 1, and we utilize `torch.random.manual_seed()` to set seeds in performing multiple imputations. We report the mean and standard error of bias, coverage, and confidence interval length across 10 different random seeds, all continuous columns, and datasets [30]. Note that we do not split the data into training and test sets [9].

Remark 1. *We acknowledge that existing imputation methods have been evaluated using RMSE (root mean square error), a metric for assessing single imputation methods [9, 20, 24, 22, 38, 11]. However, since our objective focuses on distributional learning rather than recovering missing values, we adopt the evaluation procedure proposed in [27, 30], better suited for assessing multiple imputation methods.*

Following [22, 11, 38], we generate the missing value mask for each dataset with three mechanisms in four settings. (MCAR) In the MCAR setting, each value is masked according to the realization of a Bernoulli random variable with a fixed parameter. (MAR) In the MAR setting, for each experiment, a fixed subset of variables that cannot have missing values is sampled. Then, the remaining variables have missing values according to a logistic model with random weights, which takes the non-missing variables as inputs. A bias term is fitted using line search to attain the desired proportion of missing values. (MNAR) Finally, two different mechanisms are implemented in the MNAR setting. The first, MNARL, is identical to the previously described MAR mechanism, but the inputs of the logistic model are then masked by an MCAR mechanism. Hence, the logistic model's outcome depends on missing values. The second

mechanism, MNARQ, samples a subset of variables whose values in the lower and upper pth percentiles are masked according to a Bernoulli random variable, and the values in-between are left not missing.

Model	Bias ↓	Coverage	Width ↓
MICE	$.010_{\pm.001}$	$.845_{\pm.019}$	$\mathbf{.040}_{\pm.002}$
GAIN	$.019_{\pm.002}$	$.633_{\pm.033}$	$\mathbf{.040}_{\pm.002}$
missMDA	$.015_{\pm.001}$	$.700_{\pm.022}$	$.043_{\pm.002}$
VAEAC	$\underline{.008}_{\pm.001}$	$.905_{\pm.016}$	$\mathbf{.040}_{\pm.002}$
MIWAE	$\mathbf{.006}_{\pm.000}$	$\underline{.952}_{\pm.012}$	$.043_{\pm.002}$
not-MIWAE	$\mathbf{.006}_{\pm.000}$	$\mathbf{.949}_{\pm.012}$	$\underline{.042}_{\pm.002}$
EGC	$\mathbf{.006}_{\pm.000}$	$.996_{\pm.004}$	$.058_{\pm.002}$
MaCoDE(MAR)	$\mathbf{.006}_{\pm.000}$	$.963_{\pm.009}$	$.051_{\pm.003}$

Table 3: **Q3**: Multiple imputation under MAR at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes lower is better. Coverage close to 0.95 indicates better performance. The best value is bolded, and the second best is underlined.

A.7.3 Results

The missing data mechanism within the parentheses refers to the mechanism applied to the dataset on which MaCoDE was trained. Table 3 indicates that MaCoDE consistently exhibits competitive performance against all baseline models across metrics assessing multiple imputation performances, including bias, coverage, and confidence interval length. This suggests our proposed approach can support multiple imputations for deriving statistically valid inferences from missing data with the MAR mechanism.

A.8 Additional Experiments

A.8.1 Privacy Preservability

- **Evaluation metrics:** *The k-anonymity property* [29] is a measure used to assess the level of privacy protection in synthetic data. It ensures that each individual's information in the dataset cannot be distinguished from that of at least $k-1$ other individuals. In other words, each record in the dataset has at least $k-1$ similar records in terms of quasi-identifiers, attributes that could potentially identify a subject. A higher k value implies a higher level of anonymity and better privacy preservation.

  DCR (Distance to Closest Record) [25, 40] is defined as the distances between all real training and synthetic samples. A higher DCR value indicates more effective privacy preservation, indicating a lack of overlap between the real training data and the synthetic samples. Conversely, an excessively large DCR score suggests a lower quality of the generated synthetic dataset. Therefore, the DCR metric provides insights into both the privacy-preserving capability and the quality of the synthetic dataset.

  Attribute disclosure [5, 21] refers to the situation where attackers can uncover additional covariates of a record by leveraging a subset of covariates they already possess, along with similar records from the synthetic dataset. To quantify the extent to which attackers can accurately identify these additional covariates, we employ classification metrics. Higher attribute disclosure metrics indicate an increased risk of privacy leakage, implying that attackers can precisely infer unknown variables. In terms of privacy concerns, attribute disclosure can be considered a more significant issue than membership inference attacks, as attackers are assumed to have access to only a subset of covariates for a given record.
- **Evaluation procedure:** We evaluate the k-anonymity following the approach described in [26] [¶¶]. Additionally, similar to [40], we define DCR as the 5^{th} percentile of the L_2 distances between all real training samples and synthetic samples. Since DCR relies on L_2 distance and continuous variables, it is computed solely using continuous variables. We assess attribute disclosure using the methodology outlined in [5].
- **Result:** The right panel of Figure 1 demonstrates MaCoDE's capability to regulate the privacy level by adjusting the temperature parameter τ. Simultaneously, the left panel illustrates that the quality of synthetic data, measured by feature selection performance, remains notable even with increasing privacy levels from $\tau = 1$ to $\tau = 3$. However, increasing τ beyond 3 leads to declining feature selection performance compared to other models despite DCR remaining competitive. For additional results on other metrics related to the trade-off between privacy level and synthetic data quality as τ varies, please refer to Table 4 and Table 5.

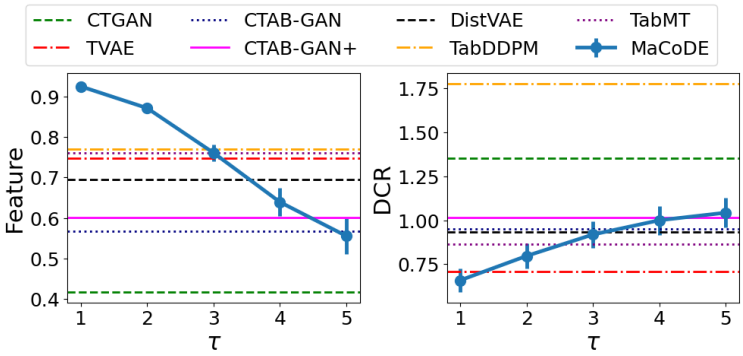


Figure 1: **Trade-off between privacy and quality.** Left: feature selection performance (synthetic data quality). Right: DCR (privacy preservability). The means and standard errors of the mean across 10 datasets and 10 repeated experiments are reported. Error bars represent the standard errors of the mean.

[¶¶] https://github.com/vanderschaarlab/synthcity

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.016_{\pm.002}$	$.029_{\pm.002}$	$.002_{\pm.000}$	$1.019_{\pm.156}$	$.107_{\pm.008}$	$.686_{\pm.023}$	$.887_{\pm.018}$	$.956_{\pm.005}$
CTGAN	$.221_{\pm.014}$	$.561_{\pm.046}$	$.094_{\pm.007}$	$6.435_{\pm1.011}$	$.256_{\pm.016}$	$.411_{\pm.027}$	$.208_{\pm.048}$	$.417_{\pm.043}$
TVAE	$.066_{\pm.003}$	$.119_{\pm.005}$	$.016_{\pm.001}$	$1.631_{\pm.173}$	$.192_{\pm.011}$	$.608_{\pm.021}$	$.486_{\pm.041}$	$.747_{\pm.027}$
CTAB-GAN	$.116_{\pm.008}$	$.196_{\pm.025}$	$.044_{\pm.004}$	$3.327_{\pm.460}$	$.218_{\pm.012}$	$.524_{\pm.026}$	$.263_{\pm.042}$	$.568_{\pm.041}$
CTAB-GAN+	$.136_{\pm.018}$	$.144_{\pm.010}$	$.054_{\pm.007}$	$3.971_{\pm.772}$	$.226_{\pm.017}$	$.530_{\pm.020}$	$.227_{\pm.048}$	$.601_{\pm.041}$
DistVAE	$.059_{\pm.007}$	$.070_{\pm.004}$	$.016_{\pm.001}$	$2.272_{\pm.282}$	$.226_{\pm.017}$	$.588_{\pm.021}$	$.194_{\pm.048}$	$.695_{\pm.030}$
TabDDPM	$.696_{\pm.117}$	$.374_{\pm.087}$	$.057_{\pm.011}$	$42.916_{\pm8.127}$	$.161_{\pm.011}$	$.576_{\pm.022}$	$.507_{\pm.039}$	$.770_{\pm.027}$
TabMT	$.011_{\pm.001}$	$.035_{\pm.003}$	$.012_{\pm.001}$	$2.299_{\pm.346}$	$.188_{\pm.013}$	$.622_{\pm.024}$	$.528_{\pm.039}$	$.761_{\pm.028}$
MaCoDE($\tau = 1$)	$.034_{\pm.004}$	$.072_{\pm.004}$	$.007_{\pm.001}$	$1.630_{\pm.245}$	$.158_{\pm.010}$	$.635_{\pm.023}$	$.599_{\pm.035}$	$.925_{\pm.007}$
MaCoDE($\tau = 2$)	$.047_{\pm.004}$	$.117_{\pm.009}$	$.021_{\pm.002}$	$2.450_{\pm.302}$	$.177_{\pm.012}$	$.601_{\pm.024}$	$.432_{\pm.038}$	$.871_{\pm.011}$
MaCoDE($\tau = 3$)	$.071_{\pm.006}$	$.290_{\pm.046}$	$.035_{\pm.003}$	$3.191_{\pm.354}$	$.196_{\pm.013}$	$.562_{\pm.024}$	$.305_{\pm.041}$	$.761_{\pm.020}$
MaCoDE($\tau = 4$)	$.088_{\pm.008}$	$.555_{\pm.109}$	$.043_{\pm.003}$	$3.601_{\pm.383}$	$.210_{\pm.014}$	$.528_{\pm.024}$	$.128_{\pm.041}$	$.639_{\pm.035}$
MaCoDE($\tau = 5$)	$.099_{\pm.010}$	$.847_{\pm.181}$	$.045_{\pm.003}$	$3.822_{\pm.402}$	$.219_{\pm.014}$	$.501_{\pm.025}$	$.065_{\pm.047}$	$.555_{\pm.044}$

Table 4: **Trade-off between privacy and quality** (statistical fidelity and machine learning utility). The means and standard errors of the mean across 10 datasets and 10 repeated experiments are reported. 'Baseline' refers to the result obtained using half of the real training dataset. ↑ (↓) denotes higher (lower) is better.

Model	k-anonimity(%) ↑	DCR ↑	AD ↓
Baseline	$1.151_{\pm0.089}$	$0.501_{\pm0.051}$	$0.631_{\pm0.023}$
CTGAN	$2.413_{\pm0.139}$	$1.355_{\pm0.152}$	$0.398_{\pm0.024}$
TVAE	$1.648_{\pm0.116}$	$0.709_{\pm0.056}$	$0.583_{\pm0.023}$
CTAB-GAN	$2.048_{\pm0.126}$	$0.951_{\pm0.076}$	$0.514_{\pm0.025}$
CTAB-GAN+	$2.215_{\pm0.121}$	$1.016_{\pm0.113}$	$0.471_{\pm0.018}$
DistVAE	$2.345_{\pm0.136}$	$0.932_{\pm0.071}$	$0.540_{\pm0.019}$
TabDDPM	$1.249_{\pm0.108}$	$1.779_{\pm0.391}$	$0.567_{\pm0.024}$
TabMT	$1.656_{\pm0.135}$	$0.862_{\pm0.072}$	$0.566_{\pm0.021}$
MaCoDE($\tau = 1$)	$1.405_{\pm0.104}$	$0.658_{\pm0.067}$	$0.589_{\pm0.022}$
MaCoDE($\tau = 2$)	$1.656_{\pm0.109}$	$0.797_{\pm0.070}$	$0.547_{\pm0.022}$
MaCoDE($\tau = 3$)	$1.959_{\pm0.147}$	$0.919_{\pm0.077}$	$0.508_{\pm0.022}$
MaCoDE($\tau = 4$)	$2.075_{\pm0.159}$	$1.000_{\pm0.082}$	$0.476_{\pm0.022}$
MaCoDE($\tau = 5$)	$2.262_{\pm0.166}$	$1.043_{\pm0.085}$	$0.457_{\pm0.022}$

Table 5: **Privacy preservability**. The means and standard errors of the mean across 10 datasets and 10 repeated experiments are reported. 'Baseline' refers to the result obtained using half of the real training dataset. ↑ (↓) denotes higher (lower) is better.

Dataset	abalone		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$3.634_{\pm 0.318}$	$0.423_{\pm 0.012}$	$0.346_{\pm 0.009}$
TVAE	$2.029_{\pm 0.234}$	$0.333_{\pm 0.010}$	$0.385_{\pm 0.018}$
CTAB-GAN	$2.769_{\pm 0.315}$	$0.449_{\pm 0.027}$	$0.312_{\pm 0.023}$
CTAB-GAN+	$2.718_{\pm 0.285}$	$0.311_{\pm 0.010}$	$0.323_{\pm 0.010}$
DistVAE	$2.074_{\pm 0.149}$	$0.316_{\pm 0.006}$	$0.349_{\pm 0.009}$
TabDDPM	$0.233_{\pm 0.047}$	$0.145_{\pm 0.003}$	$0.371_{\pm 0.016}$
TabMT	$1.595_{\pm 0.617}$	$0.423_{\pm 0.007}$	$0.337_{\pm 0.010}$
MaCoDE	$1.182_{\pm 0.292}$	$0.148_{\pm 0.002}$	$0.368_{\pm 0.014}$

Dataset	banknote		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$2.862_{\pm 0.146}$	$0.341_{\pm 0.020}$	$0.763_{\pm 0.033}$
TVAE	$2.607_{\pm 0.305}$	$0.220_{\pm 0.007}$	$0.920_{\pm 0.021}$
CTAB-GAN	$3.118_{\pm 0.202}$	$0.275_{\pm 0.008}$	$0.847_{\pm 0.025}$
CTAB-GAN+	$2.799_{\pm 0.151}$	$0.279_{\pm 0.012}$	$0.560_{\pm 0.038}$
DistVAE	$3.081_{\pm 0.267}$	$0.290_{\pm 0.010}$	$0.757_{\pm 0.027}$
TabDDPM	$0.985_{\pm 0.180}$	$0.168_{\pm 0.007}$	$0.997_{\pm 0.003}$
TabMT	$2.662_{\pm 0.262}$	$0.257_{\pm 0.009}$	$0.910_{\pm 0.015}$
MaCoDE	$2.552_{\pm 0.255}$	$0.076_{\pm 0.005}$	$0.930_{\pm 0.016}$

Dataset	breast		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$1.582_{\pm 0.383}$	$5.536_{\pm 0.167}$	$0.446_{\pm 0.050}$
TVAE	$0.813_{\pm 0.239}$	$2.148_{\pm 0.026}$	$0.900_{\pm 0.042}$
CTAB-GAN	$0.857_{\pm 0.198}$	$2.807_{\pm 0.078}$	$0.850_{\pm 0.037}$
CTAB-GAN+	$1.473_{\pm 0.289}$	$4.025_{\pm 0.315}$	$0.596_{\pm 0.078}$
DistVAE	$2.264_{\pm 0.390}$	$2.666_{\pm 0.060}$	$0.810_{\pm 0.036}$
TabDDPM	$2.835_{\pm 0.301}$	$12.733_{\pm 0.896}$	$0.740_{\pm 0.089}$
TabMT	$0.967_{\pm 0.164}$	$2.718_{\pm 0.054}$	$0.825_{\pm 0.041}$
MaCoDE	$0.527_{\pm 0.140}$	$2.264_{\pm 0.058}$	$0.906_{\pm 0.032}$

Dataset	concrete		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$3.750_{\pm 0.305}$	$1.236_{\pm 0.076}$	$0.156_{\pm 0.040}$
TVAE	$2.464_{\pm 0.386}$	$0.681_{\pm 0.018}$	$0.340_{\pm 0.054}$
CTAB-GAN	$3.519_{\pm 0.216}$	$0.973_{\pm 0.044}$	$0.310_{\pm 0.054}$
CTAB-GAN+	$4.066_{\pm 0.201}$	$0.972_{\pm 0.034}$	$0.310_{\pm 0.033}$
DistVAE	$4.235_{\pm 0.231}$	$1.012_{\pm 0.037}$	$0.358_{\pm 0.051}$
TabDDPM	$1.663_{\pm 0.253}$	$0.425_{\pm 0.011}$	$0.388_{\pm 0.044}$
TabMT	$4.029_{\pm 0.228}$	$0.949_{\pm 0.038}$	$0.308_{\pm 0.053}$
MaCoDE	$2.864_{\pm 0.293}$	$0.239_{\pm 0.019}$	$0.348_{\pm 0.043}$

Dataset	covtype		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$0.215_{\pm 0.000}$	$0.451_{\pm 0.003}$	$0.490_{\pm 0.023}$
TVAE	$0.215_{\pm 0.000}$	$0.415_{\pm 0.002}$	$0.639_{\pm 0.009}$
CTAB-GAN	$0.215_{\pm 0.000}$	$0.418_{\pm 0.004}$	$0.579_{\pm 0.004}$
CTAB-GAN+	$0.215_{\pm 0.000}$	$0.353_{\pm 0.001}$	$0.645_{\pm 0.001}$
DistVAE	$0.215_{\pm 0.000}$	$0.566_{\pm 0.002}$	$0.612_{\pm 0.002}$
TabDDPM	$0.211_{\pm 0.002}$	$0.313_{\pm 0.001}$	$0.653_{\pm 0.002}$
TabMT	$0.215_{\pm 0.000}$	$0.409_{\pm 0.001}$	$0.671_{\pm 0.003}$
MaCoDE	$0.215_{\pm 0.000}$	$0.266_{\pm 0.000}$	$0.696_{\pm 0.001}$

Dataset	kings		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$1.243_{\pm 0.168}$	$0.502_{\pm 0.006}$	$0.601_{\pm 0.004}$
TVAE	$0.848_{\pm 0.066}$	$0.361_{\pm 0.003}$	$0.685_{\pm 0.003}$
CTAB-GAN	$1.158_{\pm 0.103}$	$0.461_{\pm 0.004}$	$0.631_{\pm 0.003}$
CTAB-GAN+	$1.028_{\pm 0.126}$	$0.411_{\pm 0.006}$	$0.655_{\pm 0.003}$
DistVAE	$0.699_{\pm 0.113}$	$0.488_{\pm 0.002}$	$0.658_{\pm 0.004}$
TabDDPM	$0.236_{\pm 0.032}$	$0.325_{\pm 0.003}$	$0.668_{\pm 0.005}$
TabMT	$0.396_{\pm 0.103}$	$0.427_{\pm 0.003}$	$0.652_{\pm 0.003}$
MaCoDE	$0.445_{\pm 0.151}$	$0.301_{\pm 0.005}$	$0.677_{\pm 0.004}$

Dataset	letter		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$3.520_{\pm 0.180}$	$1.583_{\pm 0.042}$	$0.089_{\pm 0.009}$
TVAE	$2.657_{\pm 0.232}$	$0.779_{\pm 0.009}$	$0.274_{\pm 0.008}$
CTAB-GAN	$3.343_{\pm 0.267}$	$1.438_{\pm 0.016}$	$0.100_{\pm 0.006}$
CTAB-GAN+	$3.112_{\pm 0.253}$	$1.218_{\pm 0.008}$	$0.207_{\pm 0.004}$
DistVAE	$1.959_{\pm 0.265}$	$1.378_{\pm 0.006}$	$0.273_{\pm 0.005}$
TabDDPM	$1.435_{\pm 0.173}$	$0.744_{\pm 0.005}$	$0.260_{\pm 0.008}$
TabMT	$2.112_{\pm 0.236}$	$0.822_{\pm 0.008}$	$0.402_{\pm 0.005}$
MaCoDE	$1.819_{\pm 0.089}$	$1.174_{\pm 0.011}$	$0.380_{\pm 0.007}$

Dataset	loan		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$2.450_{\pm 0.244}$	$0.241_{\pm 0.007}$	$0.660_{\pm 0.012}$
TVAE	$2.442_{\pm 0.338}$	$0.160_{\pm 0.004}$	$0.716_{\pm 0.008}$
CTAB-GAN	$2.407_{\pm 0.156}$	$0.208_{\pm 0.004}$	$0.687_{\pm 0.010}$
CTAB-GAN+	$2.570_{\pm 0.159}$	$0.216_{\pm 0.015}$	$0.577_{\pm 0.005}$
DistVAE	$2.348_{\pm 0.092}$	$0.238_{\pm 0.004}$	$0.683_{\pm 0.002}$
TabDDPM	$0.992_{\pm 0.107}$	$0.123_{\pm 0.002}$	$0.699_{\pm 0.006}$
TabMT	$1.895_{\pm 0.126}$	$0.190_{\pm 0.003}$	$0.687_{\pm 0.003}$
MaCoDE	$1.942_{\pm 0.212}$	$0.127_{\pm 0.003}$	$0.697_{\pm 0.007}$

Dataset	redwine		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$2.291_{\pm 0.411}$	$1.794_{\pm 0.032}$	$0.192_{\pm 0.034}$
TVAE	$1.009_{\pm 0.325}$	$0.976_{\pm 0.016}$	$0.500_{\pm 0.041}$
CTAB-GAN	$1.345_{\pm 0.140}$	$1.191_{\pm 0.027}$	$0.454_{\pm 0.027}$
CTAB-GAN+	$2.056_{\pm 0.208}$	$1.227_{\pm 0.027}$	$0.448_{\pm 0.024}$
DistVAE	$3.815_{\pm 0.220}$	$1.171_{\pm 0.016}$	$0.436_{\pm 0.026}$
TabDDPM	$2.697_{\pm 0.289}$	$1.330_{\pm 0.046}$	$0.480_{\pm 0.025}$
TabMT	$1.071_{\pm 0.205}$	$1.236_{\pm 0.024}$	$0.420_{\pm 0.021}$
MaCoDE	$1.016_{\pm 0.086}$	$0.940_{\pm 0.018}$	$0.454_{\pm 0.026}$

Dataset	whitewine		
Model	k-anonimity(%) ↑	DCR ↑	AD ↓
CTGAN	$2.580_{\pm 0.425}$	$1.445_{\pm 0.018}$	$0.239_{\pm 0.020}$
TVAE	$1.399_{\pm 0.222}$	$1.016_{\pm 0.005}$	$0.475_{\pm 0.032}$
CTAB-GAN	$1.746_{\pm 0.287}$	$1.292_{\pm 0.027}$	$0.373_{\pm 0.022}$
CTAB-GAN+	$2.116_{\pm 0.171}$	$1.145_{\pm 0.010}$	$0.385_{\pm 0.018}$
DistVAE	$2.762_{\pm 0.184}$	$1.189_{\pm 0.007}$	$0.461_{\pm 0.016}$
TabDDPM	$0.896_{\pm 0.064}$	$1.043_{\pm 0.009}$	$0.441_{\pm 0.009}$
TabMT	$1.618_{\pm 0.303}$	$1.192_{\pm 0.013}$	$0.446_{\pm 0.014}$
MaCoDE	$1.539_{\pm 0.183}$	$0.988_{\pm 0.013}$	$0.446_{\pm 0.016}$

Table 6: **Privacy preservability** for each dataset. The means and standard errors of the mean across 10 repeated experiments are reported. ↑ (↓) denotes higher (lower) is better.

A.8.2 Sensitivity Analysis

- **Evaluation procedure:** We generated missing values into the `kings` dataset at rates of 0.1, 0.3, 0.5, and 0.7 for each missing mechanism and proceeded to train the model. Subsequently, we evaluated the machine learning utility using metrics such as SMAPE and F1-score, along with assessing the multiple imputation performance of the model using the same methodology described earlier.
- **Result:** Figure 2 and 3 show the sensitivity analysis conducted by varying the missingness rate of `kings` dataset across four missing data mechanisms (MCAR, MAR, MNARL, MNARQ). In Figure 2, it's evident that MaCoDE maintains competitive performance in terms of SMAPE, even with increasing missingness rates across all missing data mechanisms. Regarding the F_1 score, MaCoDE outperforms other models at missingness rates of 0.1 and 0.3, but its performance diminishes beyond a missingness rate of 0.5 (except for the MNARQ missing data mechanism). Additionally, as shown in Figure 3, MaCoDE consistently exhibits competitive performance in the multiple imputation, regardless of the increasing missingness rate, without significant performance degradation compared to other imputation models.

  Hence, Figures 2 and 3 demonstrate that MaCoDE maintains comparable performance to other models even when trained on a dataset with missing values (i.e., incomplete dataset) without compromising the quality of the synthetic data it generates. This is a notable advantage of our proposed model over other baseline models, which struggle with training on datasets containing missing values.

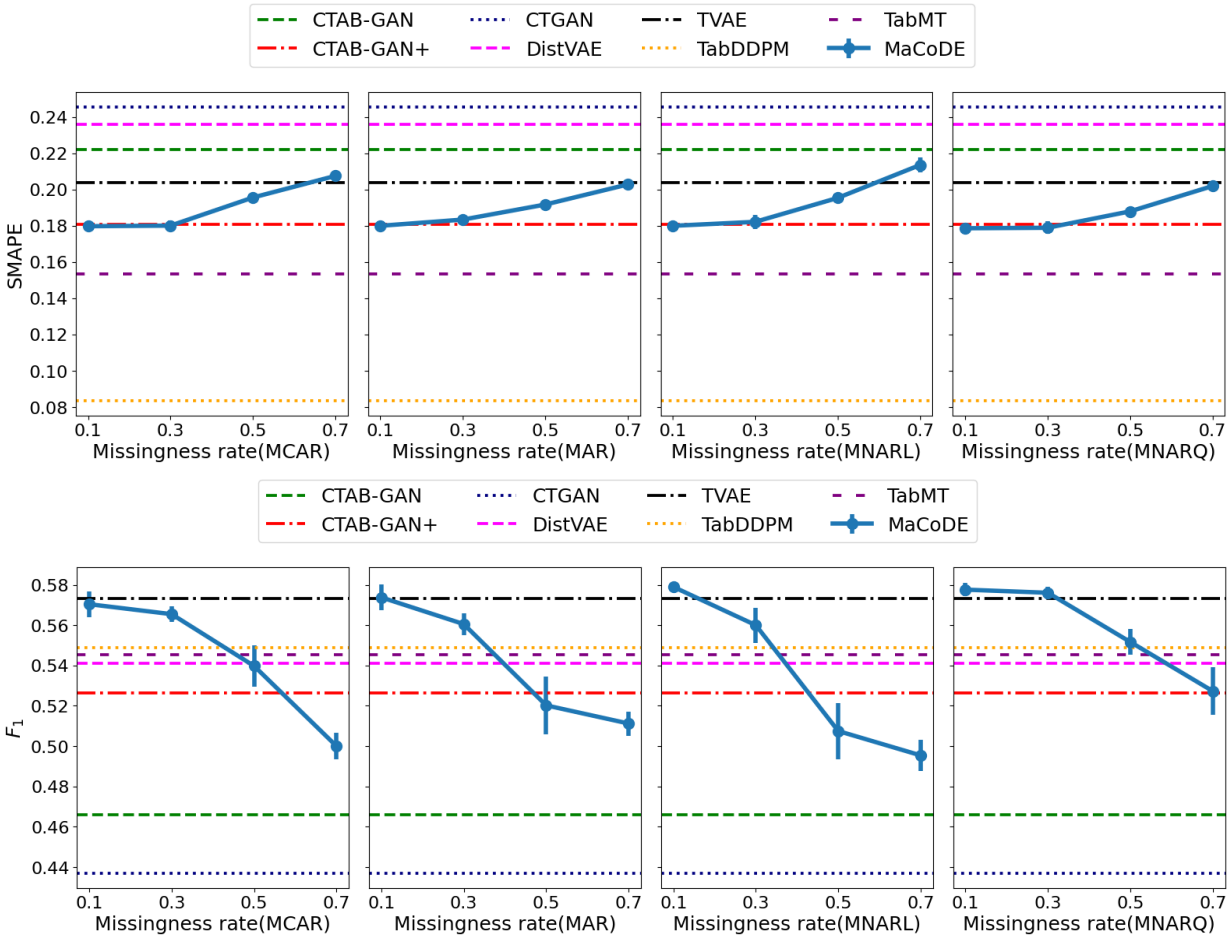


Figure 2: **Q2.** Sensitivity analysis of machine learning utility according to missingness rate. Machine learning utility is evaluated using `kings` dataset under four missing mechanisms. The means and standard errors of the mean across 10 repeated experiments are reported. Error bars represent standard errors.

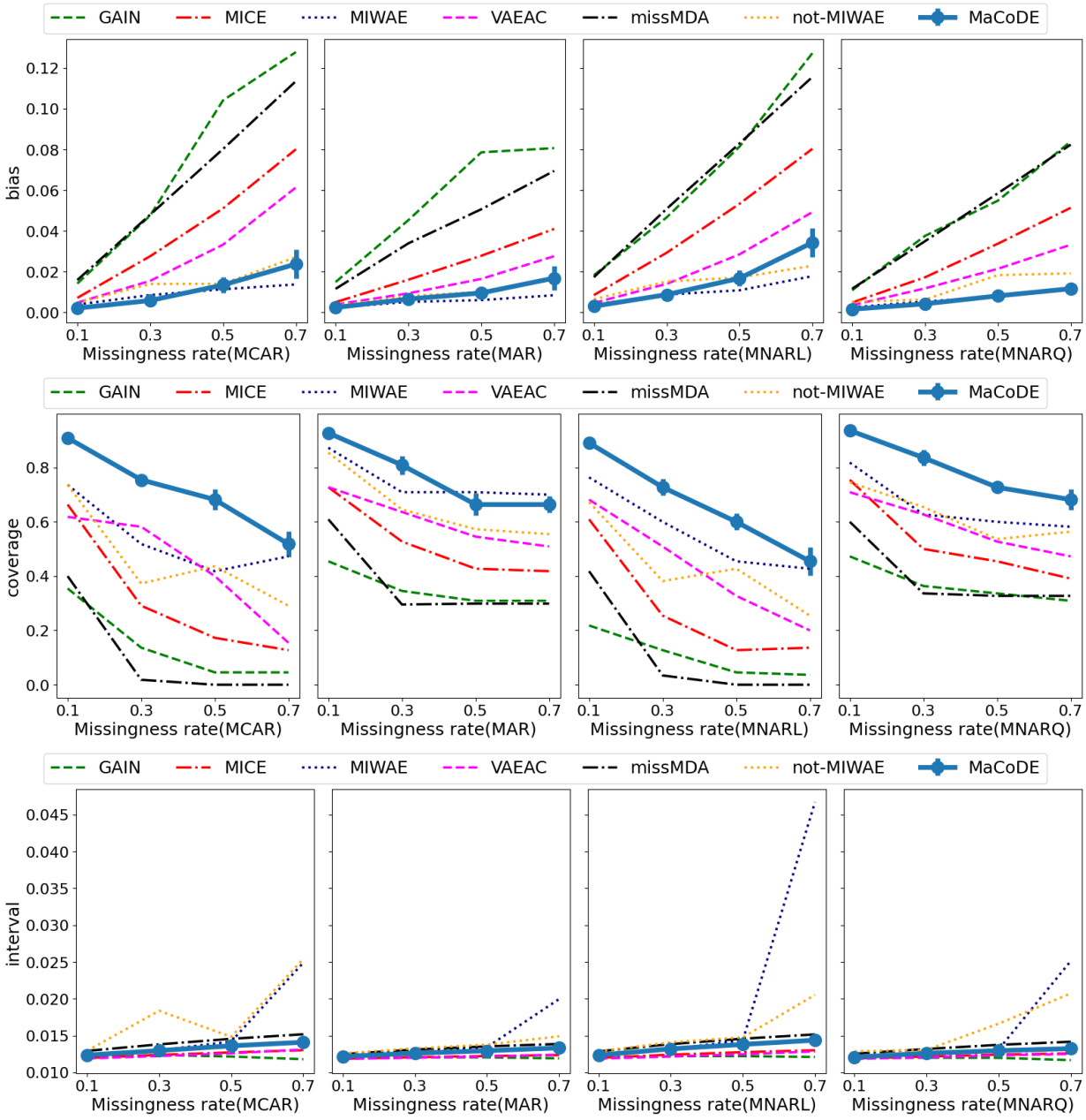


Figure 3: **Q3.** Sensitivity analysis of multiple imputations according to missingness rate. Multiple imputation is evaluated using `kings` dataset under four missing mechanisms. The means and standard errors of the mean across 10 repeated experiments are reported. Error bars represent standard errors.

A.9 Detailed Experimental Results

A.9.1 Q1. Synthetic Data Quality

`abalone`

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.006_{\pm.000}$	$.023_{\pm.001}$	$.001_{\pm.000}$	$.152_{\pm.010}$	$.031_{\pm.000}$	$.229_{\pm.003}$	$.790_{\pm.057}$	$.981_{\pm.006}$
CTGAN	$.114_{\pm.011}$	$.289_{\pm.058}$	$.109_{\pm.010}$	$1.174_{\pm.069}$	$.101_{\pm.003}$	$.166_{\pm.004}$	$.625_{\pm.087}$	$.283_{\pm.095}$
TVAE	$.064_{\pm.003}$	$.121_{\pm.004}$	$.016_{\pm.001}$	$.387_{\pm.007}$	$.120_{\pm.005}$	$.247_{\pm.003}$	$.152_{\pm.146}$	$.417_{\pm.105}$
CTAB-GAN	$.208_{\pm.012}$	$.403_{\pm.054}$	$.096_{\pm.013}$	$1.161_{\pm.120}$	$.109_{\pm.005}$	$.162_{\pm.005}$	$.274_{\pm.161}$	$.395_{\pm.158}$
CTAB-GAN+	$.035_{\pm.008}$	$.091_{\pm.020}$	$.022_{\pm.006}$	$.485_{\pm.094}$	$.068_{\pm.006}$	$.189_{\pm.013}$	$.805_{\pm.147}$	$.771_{\pm.100}$
DistVAE	$.010_{\pm.001}$	$.043_{\pm.003}$	$.005_{\pm.000}$	$.263_{\pm.007}$	$.061_{\pm.001}$	$.216_{\pm.002}$	$.825_{\pm.050}$	$.288_{\pm.034}$
TabDDPM	$.119_{\pm.030}$	$.905_{\pm.577}$	$.005_{\pm.002}$	$4.592_{\pm1.055}$	$.033_{\pm.000}$	$.222_{\pm.002}$	$.747_{\pm.059}$	$.957_{\pm.008}$
TabMT	$.003_{\pm.000}$	$.022_{\pm.001}$	$.033_{\pm.001}$	$.949_{\pm.033}$	$.081_{\pm.001}$	$.198_{\pm.003}$	$.744_{\pm.050}$	$.852_{\pm.022}$
MaCoDE	$.017_{\pm.002}$	$.064_{\pm.006}$	$.005_{\pm.001}$	$.150_{\pm.009}$	$.039_{\pm.000}$	$.217_{\pm.003}$	$.744_{\pm.049}$	$.952_{\pm.010}$

`banknote`

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.016_{\pm.001}$	$.040_{\pm.003}$	$.002_{\pm.000}$	$.059_{\pm.004}$	$.232_{\pm.002}$	$.957_{\pm.002}$	$.903_{\pm.027}$	$1.000_{\pm.000}$
CTGAN	$.146_{\pm.013}$	$.187_{\pm.013}$	$.095_{\pm.009}$	$1.260_{\pm.119}$	$.646_{\pm.009}$	$.745_{\pm.020}$	$-.321_{\pm.143}$	$.720_{\pm.085}$
TVAE	$.047_{\pm.009}$	$.084_{\pm.008}$	$.019_{\pm.004}$	$.260_{\pm.023}$	$.436_{\pm.006}$	$.879_{\pm.006}$	$.616_{\pm.079}$	$.940_{\pm.031}$
CTAB-GAN	$.030_{\pm.003}$	$.066_{\pm.003}$	$.021_{\pm.001}$	$.415_{\pm.018}$	$.519_{\pm.004}$	$.826_{\pm.011}$	$.207_{\pm.166}$	$.880_{\pm.033}$
CTAB-GAN+	$.176_{\pm.076}$	$.137_{\pm.042}$	$.072_{\pm.038}$	$.847_{\pm.227}$	$.670_{\pm.031}$	$.634_{\pm.059}$	$-.388_{\pm.274}$	$.280_{\pm.605}$
DistVAE	$.075_{\pm.002}$	$.068_{\pm.001}$	$.031_{\pm.001}$	$.563_{\pm.011}$	$.651_{\pm.005}$	$.763_{\pm.009}$	$-.513_{\pm.081}$	$1.000_{\pm.000}$
TabDDPM	$.052_{\pm.009}$	$.054_{\pm.004}$	$.006_{\pm.001}$	$.364_{\pm.059}$	$.315_{\pm.004}$	$.942_{\pm.003}$	$.839_{\pm.033}$	$1.000_{\pm.000}$
TabMT	$.012_{\pm.001}$	$.033_{\pm.001}$	$.012_{\pm.001}$	$.301_{\pm.008}$	$.495_{\pm.005}$	$.883_{\pm.007}$	$.468_{\pm.126}$	$.960_{\pm.027}$
MaCoDE	$.031_{\pm.003}$	$.074_{\pm.005}$	$.009_{\pm.001}$	$.148_{\pm.011}$	$.374_{\pm.005}$	$.918_{\pm.005}$	$.652_{\pm.071}$	$1.000_{\pm.000}$

`breast`

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.081_{\pm.005}$	$.080_{\pm.003}$	$.007_{\pm.001}$	$5.511_{\pm.116}$	$.056_{\pm.001}$	$.946_{\pm.002}$	$.662_{\pm.106}$	$.865_{\pm.015}$
CTGAN	$.419_{\pm.029}$	$.345_{\pm.013}$	$.242_{\pm.008}$	$35.684_{\pm1.061}$	$.237_{\pm.006}$	$.645_{\pm.046}$	$.057_{\pm.167}$	$-.163_{\pm.045}$
TVAE	$.049_{\pm.002}$	$.087_{\pm.002}$	$.010_{\pm.001}$	$6.519_{\pm.079}$	$.090_{\pm.001}$	$.938_{\pm.004}$	$.687_{\pm.071}$	$.895_{\pm.012}$
CTAB-GAN	$.218_{\pm.010}$	$.186_{\pm.005}$	$.092_{\pm.005}$	$16.515_{\pm.534}$	$.151_{\pm.005}$	$.877_{\pm.010}$	$.350_{\pm.129}$	$.558_{\pm.034}$
CTAB-GAN+	$.564_{\pm.255}$	$.356_{\pm.092}$	$.229_{\pm.090}$	$25.009_{\pm9.867}$	$.242_{\pm.046}$	$.696_{\pm.138}$	$.265_{\pm.525}$	$.187_{\pm.275}$
DistVAE	$.071_{\pm.002}$	$.090_{\pm.001}$	$.033_{\pm.001}$	$1.407_{\pm.107}$	$.123_{\pm.002}$	$.889_{\pm.006}$	$.274_{\pm.071}$	$.750_{\pm.022}$
TabDDPM	$3.110_{\pm.025}$	$.587_{\pm.006}$	$.306_{\pm.001}$	$262.226_{\pm6.849}$	$.345_{\pm.008}$	$.638_{\pm.016}$	$-.260_{\pm.151}$	$.364_{\pm.041}$
TabMT	$.032_{\pm.002}$	$.060_{\pm.002}$	$.028_{\pm.001}$	$12.405_{\pm.123}$	$.128_{\pm.002}$	$.911_{\pm.003}$	$.503_{\pm.131}$	$.756_{\pm.017}$
MaCoDE	$.036_{\pm.003}$	$.079_{\pm.006}$	$.012_{\pm.001}$	$8.634_{\pm.122}$	$.098_{\pm.002}$	$.921_{\pm.004}$	$.332_{\pm.167}$	$.807_{\pm.010}$

`concrete`

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.026_{\pm.002}$	$.055_{\pm.002}$	$.003_{\pm.000}$	$.433_{\pm.006}$	$.121_{\pm.002}$	$.466_{\pm.009}$	$.912_{\pm.017}$	$.950_{\pm.009}$
CTGAN	$.384_{\pm.027}$	$.830_{\pm.129}$	$.134_{\pm.012}$	$4.443_{\pm.289}$	$.334_{\pm.004}$	$.091_{\pm.020}$	$.070_{\pm.151}$	$.114_{\pm.138}$
TVAE	$.125_{\pm.007}$	$.219_{\pm.007}$	$.017_{\pm.002}$	$1.057_{\pm.031}$	$.276_{\pm.003}$	$.398_{\pm.010}$	$.030_{\pm.161}$	$.290_{\pm.045}$
CTAB-GAN	$.204_{\pm.015}$	$.226_{\pm.010}$	$.051_{\pm.006}$	$2.383_{\pm.098}$	$.290_{\pm.004}$	$.325_{\pm.011}$	$.035_{\pm.061}$	$-.260_{\pm.097}$
CTAB-GAN+	$.263_{\pm.054}$	$.278_{\pm.027}$	$.066_{\pm.023}$	$2.472_{\pm.380}$	$.285_{\pm.012}$	$.331_{\pm.045}$	$-.009_{\pm.175}$	$.017_{\pm.344}$
DistVAE	$.186_{\pm.026}$	$.159_{\pm.002}$	$.024_{\pm.001}$	$1.808_{\pm.011}$	$.289_{\pm.003}$	$.374_{\pm.007}$	$-.160_{\pm.092}$	$.107_{\pm.104}$
TabDDPM	$.036_{\pm.001}$	$.087_{\pm.007}$	$.001_{\pm.000}$	$.532_{\pm.031}$	$.156_{\pm.004}$	$.446_{\pm.007}$	$.537_{\pm.049}$	$.938_{\pm.019}$
TabMT	$.031_{\pm.001}$	$.115_{\pm.002}$	$.019_{\pm.001}$	$1.810_{\pm.034}$	$.284_{\pm.003}$	$.355_{\pm.010}$	$.392_{\pm.060}$	$.121_{\pm.086}$
MaCoDE	$.089_{\pm.019}$	$.118_{\pm.008}$	$.006_{\pm.001}$	$.684_{\pm.025}$	$.229_{\pm.002}$	$.405_{\pm.009}$	$.439_{\pm.081}$	$.919_{\pm.021}$

`covtype`

	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.000_{\pm.000}$	$.002_{\pm.000}$	$.000_{\pm.000}$	$.454_{\pm.002}$	$.067_{\pm.000}$	$.804_{\pm.000}$	$1.000_{\pm.000}$	$1.000_{\pm.000}$
CTGAN	$.105_{\pm.007}$	$.493_{\pm.063}$	$.017_{\pm.002}$	$1.338_{\pm.053}$	$.154_{\pm.001}$	$.495_{\pm.023}$	$.320_{\pm.098}$	$.878_{\pm.011}$
TVAE	$.048_{\pm.004}$	$.078_{\pm.004}$	$.013_{\pm.002}$	$.950_{\pm.021}$	$.151_{\pm.001}$	$.634_{\pm.008}$	$.420_{\pm.055}$	$.930_{\pm.019}$
CTAB-GAN	$.020_{\pm.002}$	$.064_{\pm.005}$	$.005_{\pm.000}$	$.892_{\pm.019}$	$.147_{\pm.001}$	$.574_{\pm.005}$	$.250_{\pm.067}$	$.922_{\pm.010}$
CTAB-GAN+	$.002_{\pm.000}$	$.020_{\pm.003}$	$.001_{\pm.000}$	$.601_{\pm.007}$	$.122_{\pm.001}$	$.652_{\pm.003}$	$.300_{\pm.000}$	$.958_{\pm.006}$
DistVAE	$.007_{\pm.000}$	$.028_{\pm.001}$	$.005_{\pm.000}$	$1.228_{\pm.010}$	$.182_{\pm.001}$	$.626_{\pm.002}$	$-.080_{\pm.033}$	$.895_{\pm.016}$
TabDDPM	$.003_{\pm.000}$	$.011_{\pm.000}$	$.001_{\pm.000}$	$.556_{\pm.006}$	$.115_{\pm.000}$	$.665_{\pm.001}$	$.300_{\pm.000}$	$.964_{\pm.000}$
TabMT	$.001_{\pm.000}$	$.009_{\pm.000}$	$.001_{\pm.000}$	$.621_{\pm.005}$	$.119_{\pm.000}$	$.701_{\pm.002}$	$.700_{\pm.000}$	$.960_{\pm.002}$
MaCoDE	$.001_{\pm.000}$	$.026_{\pm.001}$	$.001_{\pm.000}$	$.509_{\pm.002}$	$.098_{\pm.000}$	$.742_{\pm.001}$	$.900_{\pm.000}$	$.956_{\pm.002}$

Table 7: **Q1**: Statistical fidelity and machine learning utility for each dataset. The means and the standard errors of the mean across 10 repeated experiments are reported. 'Baseline' refers to the result obtained using half of the real training dataset. ↑ (↓) denotes higher (lower) is better.

kings								
	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.002_{\pm.000}$	$.008_{\pm.000}$	$.000_{\pm.000}$	$.414_{\pm.014}$	$.088_{\pm.001}$	$.605_{\pm.003}$	$1.000_{\pm.000}$	$.994_{\pm.001}$
CTGAN	$.141_{\pm.009}$	$.753_{\pm.083}$	$.036_{\pm.004}$	$2.702_{\pm.347}$	$.245_{\pm.003}$	$.437_{\pm.010}$	$.880_{\pm.025}$	$.881_{\pm.008}$
TVAE	$.032_{\pm.001}$	$.097_{\pm.002}$	$.004_{\pm.001}$	$.942_{\pm.038}$	$.204_{\pm.000}$	$.573_{\pm.003}$	$.940_{\pm.016}$	$.894_{\pm.008}$
CTAB-GAN	$.056_{\pm.004}$	$.421_{\pm.223}$	$.017_{\pm.001}$	$1.388_{\pm.114}$	$.222_{\pm.002}$	$.466_{\pm.010}$	$.947_{\pm.016}$	$.898_{\pm.008}$
CTAB-GAN+	$.037_{\pm.030}$	$.100_{\pm.048}$	$.039_{\pm.037}$	$2.107_{\pm1.594}$	$.181_{\pm.007}$	$.526_{\pm.022}$	$.937_{\pm.055}$	$.956_{\pm.022}$
DistVAE	$.159_{\pm.017}$	$.085_{\pm.002}$	$.005_{\pm.000}$	$1.039_{\pm.035}$	$.236_{\pm.001}$	$.541_{\pm.004}$	$.930_{\pm.030}$	$.907_{\pm.003}$
TabDDPM	$.096_{\pm.017}$	$1.621_{\pm.684}$	$.001_{\pm.000}$	$27.567_{\pm6.882}$	$.083_{\pm.001}$	$.549_{\pm.007}$	$.900_{\pm.038}$	$.978_{\pm.005}$
TabMT	$.001_{\pm.000}$	$.010_{\pm.000}$	$.001_{\pm.000}$	$.741_{\pm.011}$	$.154_{\pm.001}$	$.545_{\pm.005}$	$.940_{\pm.016}$	$.964_{\pm.002}$
MaCoDE	$.011_{\pm.001}$	$.046_{\pm.002}$	$.009_{\pm.002}$	$.998_{\pm.089}$	$.179_{\pm.001}$	$.580_{\pm.004}$	$.920_{\pm.025}$	$.975_{\pm.002}$

letter								
	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.001_{\pm.000}$	$.010_{\pm.000}$	$.000_{\pm.000}$	$.921_{\pm.002}$	$.059_{\pm.000}$	$.821_{\pm.002}$	$1.000_{\pm.000}$	$.990_{\pm.001}$
CTGAN	$.171_{\pm.004}$	$.294_{\pm.004}$	$.022_{\pm.002}$	$4.743_{\pm.182}$	$.169_{\pm.003}$	$.169_{\pm.010}$	$.120_{\pm.135}$	$.555_{\pm.055}$
TVAE	$.040_{\pm.002}$	$.064_{\pm.002}$	$.005_{\pm.000}$	$1.604_{\pm.013}$	$.106_{\pm.001}$	$.496_{\pm.011}$	$.550_{\pm.056}$	$.944_{\pm.005}$
CTAB-GAN	$.113_{\pm.005}$	$.121_{\pm.008}$	$.013_{\pm.001}$	$3.921_{\pm.059}$	$.153_{\pm.002}$	$.197_{\pm.010}$	$.130_{\pm.087}$	$.729_{\pm.030}$
CTAB-GAN+	$.051_{\pm.012}$	$.073_{\pm.015}$	$.005_{\pm.001}$	$2.782_{\pm.112}$	$.123_{\pm.001}$	$.471_{\pm.021}$	$.440_{\pm.126}$	$.899_{\pm.023}$
DistVAE	$.021_{\pm.001}$	$.044_{\pm.001}$	$.004_{\pm.000}$	$2.526_{\pm.015}$	$.120_{\pm.000}$	$.577_{\pm.003}$	$.110_{\pm.046}$	$.854_{\pm.009}$
TabDDPM	$.009_{\pm.000}$	$.031_{\pm.001}$	$.002_{\pm.000}$	$1.371_{\pm.023}$	$.075_{\pm.000}$	$.448_{\pm.005}$	$.500_{\pm.001}$	$.531_{\pm.008}$
TabMT	$.002_{\pm.000}$	$.010_{\pm.000}$	$.001_{\pm.000}$	$1.579_{\pm.009}$	$.083_{\pm.000}$	$.745_{\pm.003}$	$.900_{\pm.001}$	$.976_{\pm.003}$
MaCoDE	$.095_{\pm.002}$	$.144_{\pm.003}$	$.017_{\pm.001}$	$2.135_{\pm.035}$	$.106_{\pm.001}$	$.689_{\pm.003}$	$.850_{\pm.017}$	$.965_{\pm.003}$

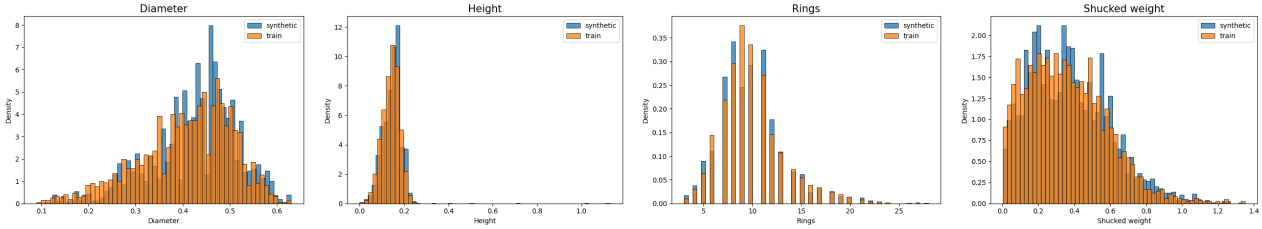
loan								
	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.002_{\pm.000}$	$.011_{\pm.000}$	$.001_{\pm.000}$	$.071_{\pm.001}$	$.267_{\pm.001}$	$.930_{\pm.001}$	$.907_{\pm.037}$	$.965_{\pm.008}$
CTGAN	$.079_{\pm.009}$	$.177_{\pm.014}$	$.056_{\pm.006}$	$.963_{\pm.070}$	$.368_{\pm.003}$	$.886_{\pm.004}$	$.272_{\pm.079}$	$.682_{\pm.045}$
TVAE	$.121_{\pm.004}$	$.170_{\pm.003}$	$.023_{\pm.000}$	$.751_{\pm.028}$	$.339_{\pm.002}$	$.850_{\pm.007}$	$.685_{\pm.102}$	$.916_{\pm.017}$
CTAB-GAN	$.037_{\pm.004}$	$.087_{\pm.004}$	$.023_{\pm.003}$	$.502_{\pm.032}$	$.333_{\pm.004}$	$.904_{\pm.003}$	$.440_{\pm.097}$	$.802_{\pm.014}$
CTAB-GAN+	$.050_{\pm.028}$	$.084_{\pm.009}$	$.039_{\pm.024}$	$.505_{\pm.251}$	$.304_{\pm.020}$	$.892_{\pm.011}$	$.144_{\pm.313}$	$.810_{\pm.043}$
DistVAE	$.011_{\pm.001}$	$.058_{\pm.002}$	$.030_{\pm.000}$	$.673_{\pm.007}$	$.361_{\pm.001}$	$.903_{\pm.003}$	$.517_{\pm.069}$	$.781_{\pm.016}$
TabDDPM	$.004_{\pm.000}$	$.010_{\pm.000}$	$.001_{\pm.000}$	$.075_{\pm.007}$	$.281_{\pm.001}$	$.925_{\pm.001}$	$.677_{\pm.039}$	$.978_{\pm.006}$
TabMT	$.002_{\pm.000}$	$.013_{\pm.001}$	$.003_{\pm.000}$	$.192_{\pm.008}$	$.307_{\pm.001}$	$.913_{\pm.002}$	$.623_{\pm.043}$	$.891_{\pm.012}$
MaCoDE	$.031_{\pm.004}$	$.047_{\pm.003}$	$.007_{\pm.001}$	$.126_{\pm.008}$	$.273_{\pm.001}$	$.919_{\pm.002}$	$.726_{\pm.033}$	$.896_{\pm.021}$

redwine								
	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.025_{\pm.004}$	$.042_{\pm.002}$	$.003_{\pm.000}$	$1.180_{\pm.027}$	$.080_{\pm.001}$	$.576_{\pm.005}$	$.802_{\pm.072}$	$.905_{\pm.014}$
CTGAN	$.463_{\pm.017}$	$.914_{\pm.178}$	$.164_{\pm.007}$	$8.166_{\pm.528}$	$.177_{\pm.005}$	$.228_{\pm.031}$	$.116_{\pm.144}$	$.071_{\pm.119}$
TVAE	$.057_{\pm.002}$	$.115_{\pm.002}$	$.012_{\pm.001}$	$1.842_{\pm.042}$	$.103_{\pm.001}$	$.569_{\pm.005}$	$.238_{\pm.117}$	$.684_{\pm.043}$
CAB-GAN	$.111_{\pm.012}$	$.143_{\pm.008}$	$.042_{\pm.006}$	$3.075_{\pm.150}$	$.139_{\pm.003}$	$.506_{\pm.007}$	$.010_{\pm.109}$	$.445_{\pm.056}$
CTAB-GAN+	$.118_{\pm.025}$	$.157_{\pm.031}$	$.044_{\pm.014}$	$2.969_{\pm.312}$	$.152_{\pm.012}$	$.488_{\pm.049}$	$-.145_{\pm.413}$	$.510_{\pm.343}$
DistVAE	$.035_{\pm.001}$	$.064_{\pm.001}$	$.015_{\pm.000}$	$2.423_{\pm.018}$	$.134_{\pm.001}$	$.533_{\pm.005}$	$-.108_{\pm.091}$	$.776_{\pm.030}$
TabDDPM	$2.678_{\pm.108}$	$.652_{\pm.050}$	$.214_{\pm.009}$	$86.984_{\pm4.091}$	$.108_{\pm.002}$	$.481_{\pm.010}$	$.412_{\pm.122}$	$.372_{\pm.066}$
TabMT	$.023_{\pm.002}$	$.052_{\pm.001}$	$.015_{\pm.000}$	$2.534_{\pm.024}$	$.127_{\pm.002}$	$.507_{\pm.009}$	$.010_{\pm.112}$	$.567_{\pm.057}$
MaCoDE	$.022_{\pm.001}$	$.071_{\pm.002}$	$.005_{\pm.000}$	$1.526_{\pm.016}$	$.105_{\pm.001}$	$.524_{\pm.007}$	$.197_{\pm.092}$	$.870_{\pm.032}$

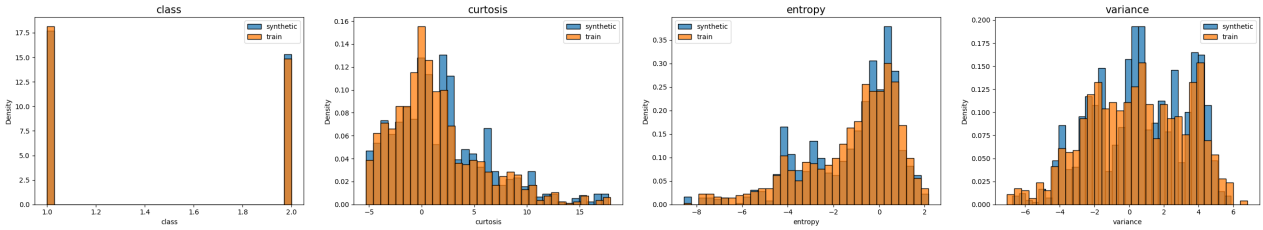
whitewine								
	Statistical fidelity				Machine learning utility			
Model	KL ↓	GoF ↓	MMD ↓	WD ↓	SMAPE ↓	F_1 ↑	Model ↑	Feature ↑
Baseline	$.006_{\pm.001}$	$.022_{\pm.001}$	$.001_{\pm.000}$	$.993_{\pm.008}$	$.065_{\pm.000}$	$.523_{\pm.004}$	$.897_{\pm.027}$	$.913_{\pm.012}$
CTGAN	$.186_{\pm.008}$	$1.330_{\pm.159}$	$.068_{\pm.005}$	$3.881_{\pm.153}$	$.123_{\pm.002}$	$.248_{\pm.019}$	$-.060_{\pm.064}$	$.148_{\pm.083}$
TVAE	$.078_{\pm.005}$	$.152_{\pm.005}$	$.040_{\pm.004}$	$1.996_{\pm.064}$	$.100_{\pm.001}$	$.492_{\pm.003}$	$.545_{\pm.127}$	$.561_{\pm.066}$
CTAB-GAN	$.166_{\pm.016}$	$.241_{\pm.019}$	$.080_{\pm.011}$	$3.016_{\pm.147}$	$.117_{\pm.002}$	$.407_{\pm.010}$	$-.010_{\pm.089}$	$.311_{\pm.090}$
CTAB-GAN+	$.064_{\pm.022}$	$.139_{\pm.022}$	$.029_{\pm.012}$	$1.929_{\pm.265}$	$.109_{\pm.004}$	$.424_{\pm.021}$	$-.080_{\pm.326}$	$.620_{\pm.189}$
DistVAE	$.018_{\pm.001}$	$.059_{\pm.001}$	$.009_{\pm.001}$	$1.787_{\pm.017}$	$.106_{\pm.001}$	$.459_{\pm.003}$	$.150_{\pm.095}$	$.595_{\pm.026}$
TabDDPM	$.675_{\pm.041}$	$.161_{\pm.011}$	$.022_{\pm.002}$	$4.292_{\pm2.951}$	$.079_{\pm.001}$	$.439_{\pm.008}$	$.535_{\pm.070}$	$.685_{\pm.056}$
TabMT	$.005_{\pm.001}$	$.031_{\pm.000}$	$.007_{\pm.000}$	$1.861_{\pm.010}$	$.102_{\pm.001}$	$.462_{\pm.004}$	$-.001_{\pm.068}$	$.562_{\pm.035}$
MaCoDE	$.010_{\pm.001}$	$.049_{\pm.002}$	$.002_{\pm.000}$	$1.238_{\pm.008}$	$.087_{\pm.000}$	$.465_{\pm.003}$	$.240_{\pm.070}$	$.903_{\pm.013}$

Table 8: **Q1**: Statistical fidelity and machine learning utility for each dataset. The means and the standard errors of the mean across 10 repeated experiments are reported. 'Baseline' refers to the result obtained using half of the real training dataset. ↑ (↓) denotes higher (lower) is better.

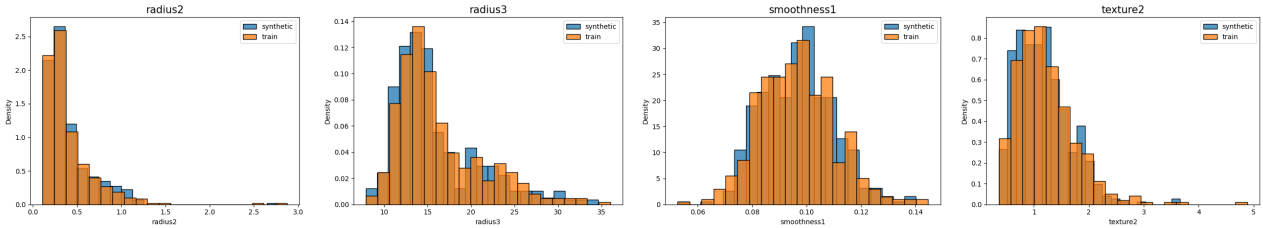
A.9.2 Q1. Visualization of Marginal Histogram



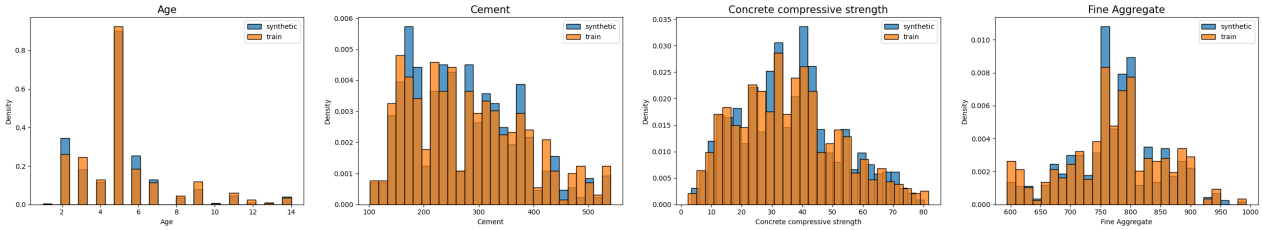
(a) `abalone`



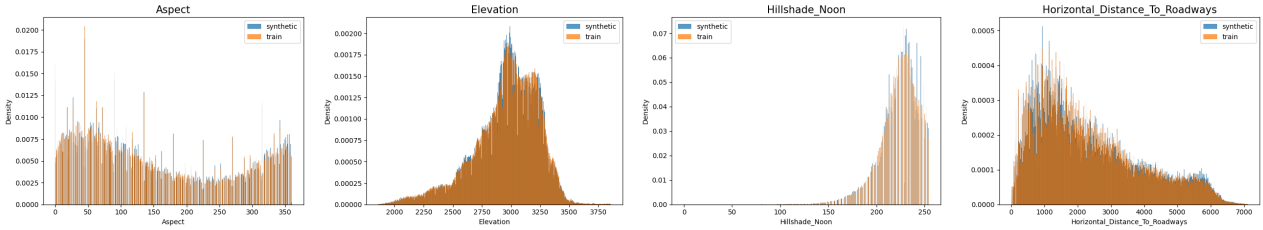
(b) `banknote`



(c) `breast`

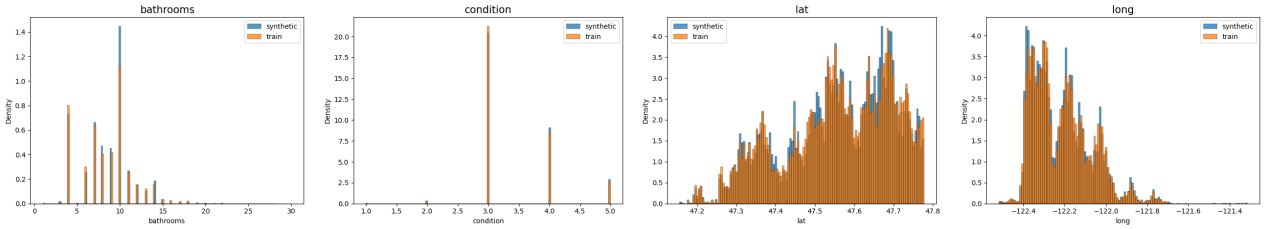


(d) `concreate`

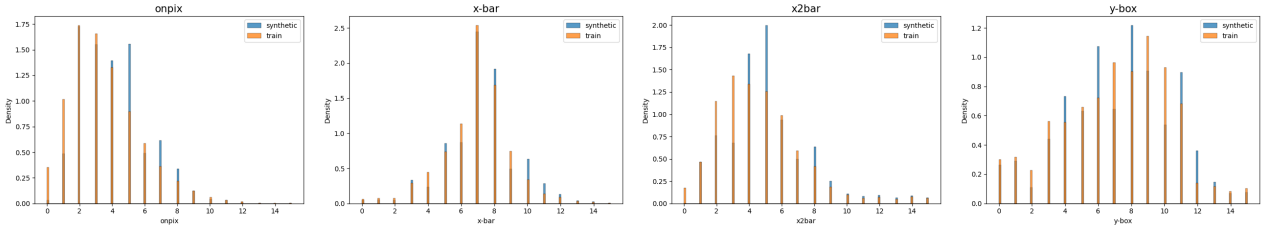


(e) `covtype`

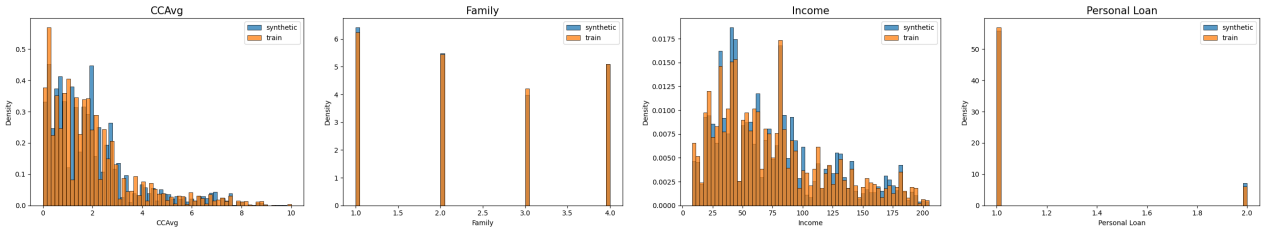
Figure 4: Histograms of observed dataset and synthetic dataset, generated by MaCoDE.



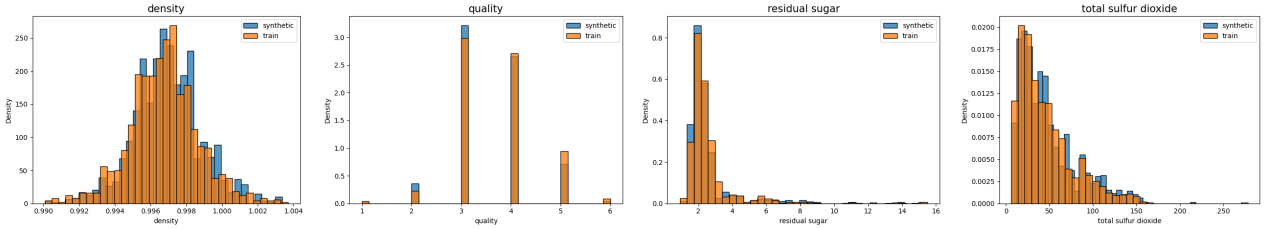
(a) `kings`



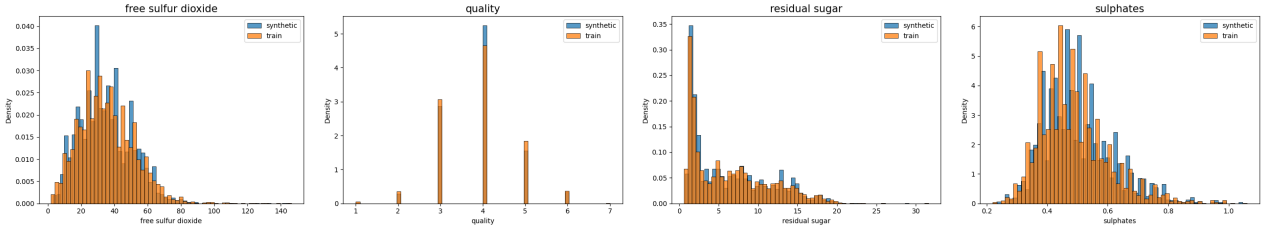
(b) `letter`



(c) `loan`



(d) `redwine`



(e) `whitewine`

Figure 5: Histograms of observed dataset and synthetic dataset, generated by MaCoDE.

A.9.3 Q2: Synthetic Data Quality in Scenarios with Incomplete Training Dataset

	MCAR	
Dataset	SMAPE ↓	F_1 ↑
abalone	$0.042_{\pm 0.000}$	$0.216_{\pm 0.003}$
banknote	$0.417_{\pm 0.004}$	$0.879_{\pm 0.006}$
breast	$0.111_{\pm 0.001}$	$0.909_{\pm 0.006}$
concrete	$0.240_{\pm 0.002}$	$0.373_{\pm 0.010}$
covtype	$0.103_{\pm 0.000}$	$0.721_{\pm 0.001}$
kings	$0.180_{\pm 0.003}$	$0.565_{\pm 0.004}$
letter	$0.107_{\pm 0.000}$	$0.674_{\pm 0.004}$
loan	$0.278_{\pm 0.003}$	$0.915_{\pm 0.002}$
redwine	$0.112_{\pm 0.001}$	$0.515_{\pm 0.009}$
whitewine	$0.091_{\pm 0.000}$	$0.461_{\pm 0.005}$

	MAR	
Dataset	SMAPE ↓	F_1 ↑
abalone	$0.041_{\pm 0.000}$	$0.211_{\pm 0.003}$
banknote	$0.410_{\pm 0.005}$	$0.891_{\pm 0.005}$
breast	$0.107_{\pm 0.002}$	$0.916_{\pm 0.005}$
concrete	$0.244_{\pm 0.003}$	$0.383_{\pm 0.007}$
covtype	$0.102_{\pm 0.000}$	$0.723_{\pm 0.001}$
kings	$0.183_{\pm 0.001}$	$0.561_{\pm 0.005}$
letter	$0.107_{\pm 0.000}$	$0.682_{\pm 0.004}$
loan	$0.278_{\pm 0.001}$	$0.914_{\pm 0.002}$
redwine	$0.110_{\pm 0.001}$	$0.518_{\pm 0.005}$
whitewine	$0.090_{\pm 0.001}$	$0.465_{\pm 0.006}$

	MNARL	
Dataset	SMAPE ↓	F_1 ↑
abalone	$0.042_{\pm 0.000}$	$0.212_{\pm 0.004}$
banknote	$0.418_{\pm 0.007}$	$0.883_{\pm 0.006}$
breast	$0.112_{\pm 0.002}$	$0.908_{\pm 0.006}$
concrete	$0.243_{\pm 0.003}$	$0.374_{\pm 0.011}$
covtype	$0.104_{\pm 0.000}$	$0.718_{\pm 0.001}$
kings	$0.182_{\pm 0.004}$	$0.560_{\pm 0.009}$
letter	$0.108_{\pm 0.001}$	$0.676_{\pm 0.003}$
loan	$0.277_{\pm 0.001}$	$0.915_{\pm 0.002}$
redwine	$0.112_{\pm 0.001}$	$0.511_{\pm 0.009}$
whitewine	$0.091_{\pm 0.001}$	$0.464_{\pm 0.006}$

	MNARQ	
Dataset	SMAPE ↓	F_1 ↑
abalone	$0.041_{\pm 0.000}$	$0.211_{\pm 0.003}$
banknote	$0.399_{\pm 0.003}$	$0.899_{\pm 0.009}$
breast	$0.105_{\pm 0.001}$	$0.918_{\pm 0.003}$
concrete	$0.239_{\pm 0.003}$	$0.384_{\pm 0.009}$
covtype	$0.101_{\pm 0.000}$	$0.730_{\pm 0.001}$
kings	$0.179_{\pm 0.004}$	$0.576_{\pm 0.003}$
letter	$0.107_{\pm 0.001}$	$0.676_{\pm 0.004}$
loan	$0.275_{\pm 0.001}$	$0.916_{\pm 0.001}$
redwine	$0.108_{\pm 0.001}$	$0.527_{\pm 0.006}$
whitewine	$0.089_{\pm 0.001}$	$0.460_{\pm 0.004}$

Table 9: **Q2**: Machine learning utility for each dataset under **MCAR**, **MAR**, **MNARL**, and **MNARQ** at 0.3 missingness. The means and standard errors of the mean 10 repeated experiments are reported. ↑ (↓) denotes higher (lower) is better.

A.9.4 Q3: Multiple Imputation Performance

	MCAR		
Model	Bias ↓	Coverage	Width ↓
MICE	$0.014_{\pm 0.001}$	$0.795_{\pm 0.022}$	$0.040_{\pm 0.002}$
GAIN	$0.024_{\pm 0.002}$	$0.591_{\pm 0.037}$	$0.040_{\pm 0.002}$
missMDA	$0.018_{\pm 0.001}$	$0.644_{\pm 0.015}$	$0.045_{\pm 0.002}$
VAEAC	$0.010_{\pm 0.001}$	$0.874_{\pm 0.022}$	$0.041_{\pm 0.002}$
MIWAE	$0.008_{\pm 0.000}$	$0.936_{\pm 0.014}$	$0.045_{\pm 0.002}$
notMIWAE	$0.008_{\pm 0.000}$	$0.910_{\pm 0.013}$	$0.044_{\pm 0.002}$
EGC	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.060_{\pm 0.002}$
MaCoDE(MCAR)	$0.008_{\pm 0.000}$	$0.950_{\pm 0.011}$	$0.053_{\pm 0.002}$

	MAR		
Model	Bias ↓	Coverage	Width ↓
MICE	$0.010_{\pm 0.001}$	$0.845_{\pm 0.019}$	$0.040_{\pm 0.002}$
GAIN	$0.019_{\pm 0.002}$	$0.633_{\pm 0.033}$	$0.040_{\pm 0.002}$
missMDA	$0.015_{\pm 0.001}$	$0.700_{\pm 0.022}$	$0.043_{\pm 0.002}$
VAEAC	$0.008_{\pm 0.001}$	$0.905_{\pm 0.016}$	$0.040_{\pm 0.002}$
MIWAE	$0.006_{\pm 0.000}$	$0.952_{\pm 0.012}$	$0.043_{\pm 0.002}$
notMIWAE	$0.006_{\pm 0.000}$	$0.949_{\pm 0.012}$	$0.042_{\pm 0.002}$
EGC	$0.006_{\pm 0.000}$	$0.996_{\pm 0.004}$	$0.058_{\pm 0.002}$
MaCoDE(MAR)	$0.006_{\pm 0.000}$	$0.963_{\pm 0.009}$	$0.051_{\pm 0.003}$

	MNARL		
Model	Bias ↓	Coverage	Width ↓
MICE	$0.012_{\pm 0.001}$	$0.790_{\pm 0.026}$	$0.037_{\pm 0.002}$
GAIN	$0.026_{\pm 0.002}$	$0.538_{\pm 0.033}$	$0.040_{\pm 0.002}$
missMDA	$0.020_{\pm 0.001}$	$0.592_{\pm 0.025}$	$0.045_{\pm 0.002}$
VAEAC	$0.011_{\pm 0.001}$	$0.851_{\pm 0.023}$	$0.041_{\pm 0.002}$
MIWAE	$0.009_{\pm 0.001}$	$0.920_{\pm 0.015}$	$0.045_{\pm 0.002}$
notMIWAE	$0.009_{\pm 0.001}$	$0.908_{\pm 0.012}$	$0.044_{\pm 0.002}$
EGC	$0.007_{\pm 0.000}$	$0.994_{\pm 0.004}$	$0.061_{\pm 0.002}$
MaCoDE(MNARL)	$0.008_{\pm 0.000}$	$0.948_{\pm 0.013}$	$0.052_{\pm 0.003}$

	MNARQ		
Model	Bias ↓	Coverage	Width ↓
MICE	$0.007_{\pm 0.001}$	$0.880_{\pm 0.018}$	$0.040_{\pm 0.002}$
GAIN	$0.011_{\pm 0.001}$	$0.819_{\pm 0.020}$	$0.040_{\pm 0.002}$
missMDA	$0.011_{\pm 0.001}$	$0.785_{\pm 0.016}$	$0.043_{\pm 0.002}$
VAEAC	$0.006_{\pm 0.001}$	$0.909_{\pm 0.020}$	$0.040_{\pm 0.002}$
MIWAE	$0.005_{\pm 0.000}$	$0.969_{\pm 0.009}$	$0.043_{\pm 0.002}$
notMIWAE	$0.005_{\pm 0.000}$	$0.958_{\pm 0.010}$	$0.043_{\pm 0.002}$
EGC	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.002}$
MaCoDE(MNARQ)	$0.005_{\pm 0.000}$	$0.944_{\pm 0.012}$	$0.051_{\pm 0.003}$

Table 10: **Q3**: Multiple imputation under **MCAR**, **MAR**, **MNARL**, and **MNARQ** at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes the lower is better.

Dataset		`abalone`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.030_{\pm 0.000}$
GAIN	$0.009_{\pm 0.001}$	$0.829_{\pm 0.042}$	$0.030_{\pm 0.000}$
missMDA	$0.014_{\pm 0.000}$	$0.686_{\pm 0.036}$	$0.034_{\pm 0.001}$
VAEAC	$0.005_{\pm 0.002}$	$0.943_{\pm 0.074}$	$0.030_{\pm 0.000}$
MIWAE	$0.006_{\pm 0.001}$	$0.971_{\pm 0.019}$	$0.034_{\pm 0.001}$
not-MIWAE	$0.009_{\pm 0.001}$	$0.886_{\pm 0.019}$	$0.037_{\pm 0.001}$
EGC	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.040_{\pm 0.000}$
MaCoDE	$0.007_{\pm 0.001}$	$0.914_{\pm 0.023}$	$0.036_{\pm 0.002}$

Dataset		`banknote`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.018_{\pm 0.001}$	$0.725_{\pm 0.025}$	$0.053_{\pm 0.000}$
GAIN	$0.009_{\pm 0.001}$	$0.829_{\pm 0.042}$	$0.030_{\pm 0.000}$
missMDA	$0.018_{\pm 0.000}$	$0.625_{\pm 0.042}$	$0.056_{\pm 0.000}$
VAEAC	$0.009_{\pm 0.002}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$
MIWAE	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.055_{\pm 0.000}$
not-MIWAE	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.055_{\pm 0.000}$
EGC	$0.004_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.066_{\pm 0.001}$
MaCoDE	$0.006_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$

Dataset		`breast`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.007_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.080_{\pm 0.000}$
GAIN	$0.012_{\pm 0.000}$	$0.990_{\pm 0.005}$	$0.080_{\pm 0.000}$
missMDA	$0.025_{\pm 0.000}$	$0.883_{\pm 0.013}$	$0.086_{\pm 0.000}$
VAEAC	$0.008_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
MIWAE	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.084_{\pm 0.000}$
not-MIWAE	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
EGC	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.084_{\pm 0.000}$
MaCoDE	$0.009_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.083_{\pm 0.000}$

Dataset		`redwine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.020_{\pm 0.000}$	$0.764_{\pm 0.020}$	$0.048_{\pm 0.000}$
GAIN	$0.024_{\pm 0.001}$	$0.627_{\pm 0.050}$	$0.048_{\pm 0.000}$
missMDA	$0.026_{\pm 0.000}$	$0.618_{\pm 0.012}$	$0.055_{\pm 0.000}$
VAEAC	$0.014_{\pm 0.001}$	$0.855_{\pm 0.028}$	$0.050_{\pm 0.001}$
MIWAE	$0.009_{\pm 0.001}$	$0.945_{\pm 0.020}$	$0.054_{\pm 0.000}$
not-MIWAE	$0.010_{\pm 0.000}$	$0.909_{\pm 0.000}$	$0.053_{\pm 0.000}$
EGC	$0.007_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.070_{\pm 0.000}$
MaCoDE	$0.008_{\pm 0.001}$	$0.973_{\pm 0.014}$	$0.056_{\pm 0.000}$

Dataset		`whitewine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.011_{\pm 0.000}$	$0.691_{\pm 0.031}$	$0.028_{\pm 0.000}$
GAIN	$0.026_{\pm 0.003}$	$0.482_{\pm 0.036}$	$0.028_{\pm 0.000}$
missMDA	$0.015_{\pm 0.000}$	$0.645_{\pm 0.016}$	$0.035_{\pm 0.000}$
VAEAC	$0.010_{\pm 0.001}$	$0.700_{\pm 0.038}$	$0.028_{\pm 0.000}$
MIWAE	$0.011_{\pm 0.001}$	$0.827_{\pm 0.029}$	$0.036_{\pm 0.001}$
not-MIWAE	$0.009_{\pm 0.000}$	$0.845_{\pm 0.033}$	$0.033_{\pm 0.001}$
EGC	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.062_{\pm 0.000}$
MaCoDE	$0.009_{\pm 0.001}$	$0.864_{\pm 0.034}$	$0.037_{\pm 0.001}$

Table 11: **Q3**: Multiple imputation for each dataset under **MCAR** at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes lower is better.

Dataset		`abalone`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.005_{\pm 0.001}$	$0.971_{\pm 0.019}$	$0.030_{\pm 0.000}$
GAIN	$0.011_{\pm 0.002}$	$0.729_{\pm 0.072}$	$0.030_{\pm 0.000}$
missMDA	$0.015_{\pm 0.001}$	$0.571_{\pm 0.048}$	$0.032_{\pm 0.001}$
VAEAC	$0.003_{\pm 0.001}$	$0.986_{\pm 0.014}$	$0.031_{\pm 0.000}$
MIWAE	$0.005_{\pm 0.001}$	$0.971_{\pm 0.019}$	$0.033_{\pm 0.001}$
not-MIWAE	$0.006_{\pm 0.001}$	$0.943_{\pm 0.023}$	$0.032_{\pm 0.001}$
EGC	$0.004_{\pm 0.001}$	$0.986_{\pm 0.014}$	$0.035_{\pm 0.001}$
MaCoDE	$0.005_{\pm 0.001}$	$0.957_{\pm 0.022}$	$0.033_{\pm 0.001}$

Dataset		`banknote`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.012_{\pm 0.002}$	$0.800_{\pm 0.033}$	$0.052_{\pm 0.000}$
GAIN	$0.021_{\pm 0.004}$	$0.675_{\pm 0.084}$	$0.052_{\pm 0.000}$
missMDA	$0.014_{\pm 0.001}$	$0.800_{\pm 0.033}$	$0.055_{\pm 0.000}$
VAEAC	$0.009_{\pm 0.001}$	$0.925_{\pm 0.038}$	$0.053_{\pm 0.000}$
MIWAE	$0.004_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$
not-MIWAE	$0.004_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$
EGC	$0.005_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.071_{\pm 0.003}$
MaCoDE	$0.004_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$

Dataset		`breast`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.080_{\pm 0.000}$
GAIN	$0.011_{\pm 0.001}$	$0.960_{\pm 0.016}$	$0.080_{\pm 0.000}$
missMDA	$0.019_{\pm 0.001}$	$0.887_{\pm 0.015}$	$0.084_{\pm 0.000}$
VAEAC	$0.006_{\pm 0.000}$	$0.997_{\pm 0.002}$	$0.081_{\pm 0.000}$
MIWAE	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
not-MIWAE	$0.005_{\pm 0.001}$	$0.993_{\pm 0.007}$	$0.081_{\pm 0.000}$
EGC	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.083_{\pm 0.000}$
MaCoDE	$0.007_{\pm 0.000}$	$0.997_{\pm 0.003}$	$0.083_{\pm 0.000}$

Dataset		`redwine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.015_{\pm 0.001}$	$0.836_{\pm 0.035}$	$0.048_{\pm 0.000}$
GAIN	$0.020_{\pm 0.001}$	$0.627_{\pm 0.021}$	$0.048_{\pm 0.000}$
missMDA	$0.020_{\pm 0.002}$	$0.709_{\pm 0.035}$	$0.053_{\pm 0.000}$
VAEAC	$0.011_{\pm 0.001}$	$0.882_{\pm 0.019}$	$0.049_{\pm 0.000}$
MIWAE	$0.008_{\pm 0.001}$	$0.945_{\pm 0.024}$	$0.053_{\pm 0.001}$
not-MIWAE	$0.009_{\pm 0.001}$	$0.909_{\pm 0.030}$	$0.051_{\pm 0.000}$
EGC	$0.007_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.067_{\pm 0.001}$
MaCoDE	$0.008_{\pm 0.000}$	$0.980_{\pm 0.013}$	$0.054_{\pm 0.001}$

Dataset		`whitewine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.008_{\pm 0.001}$	$0.773_{\pm 0.034}$	$0.028_{\pm 0.000}$
GAIN	$0.025_{\pm 0.003}$	$0.500_{\pm 0.053}$	$0.028_{\pm 0.000}$
missMDA	$0.011_{\pm 0.000}$	$0.718_{\pm 0.032}$	$0.033_{\pm 0.001}$
VAEAC	$0.008_{\pm 0.001}$	$0.827_{\pm 0.029}$	$0.029_{\pm 0.000}$
MIWAE	$0.007_{\pm 0.001}$	$0.891_{\pm 0.030}$	$0.033_{\pm 0.001}$
not-MIWAE	$0.006_{\pm 0.001}$	$0.945_{\pm 0.024}$	$0.031_{\pm 0.000}$
EGC	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.058_{\pm 0.001}$
MaCoDE	$0.007_{\pm 0.003}$	$0.882_{\pm 0.075}$	$0.033_{\pm 0.003}$

Table 12: **Q3**: Multiple imputation for each dataset under **MAR** at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes lower is better. ↓ denotes lower is better.

Dataset		`abalone`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.007_{\pm 0.001}$	$0.929_{\pm 0.024}$	$0.030_{\pm 0.000}$
GAIN	$0.015_{\pm 0.002}$	$0.629_{\pm 0.061}$	$0.030_{\pm 0.000}$
missMDA	$0.018_{\pm 0.001}$	$0.471_{\pm 0.057}$	$0.034_{\pm 0.001}$
VAEAC	$0.005_{\pm 0.001}$	$0.957_{\pm 0.022}$	$0.030_{\pm 0.000}$
MIWAE	$0.008_{\pm 0.001}$	$0.900_{\pm 0.043}$	$0.034_{\pm 0.001}$
not-MIWAE	$0.010_{\pm 0.002}$	$0.886_{\pm 0.060}$	$0.035_{\pm 0.003}$
EGC	$0.005_{\pm 0.000}$	$0.986_{\pm 0.014}$	$0.039_{\pm 0.001}$
MaCoDE	$0.007_{\pm 0.001}$	$0.929_{\pm 0.024}$	$0.035_{\pm 0.001}$

Dataset		`banknote`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.017_{\pm 0.001}$	$0.775_{\pm 0.025}$	$0.053_{\pm 0.000}$
GAIN	$0.032_{\pm 0.005}$	$0.625_{\pm 0.077}$	$0.053_{\pm 0.000}$
missMDA	$0.019_{\pm 0.001}$	$0.725_{\pm 0.045}$	$0.056_{\pm 0.000}$
VAEAC	$0.010_{\pm 0.001}$	$0.975_{\pm 0.025}$	$0.053_{\pm 0.000}$
MIWAE	$0.005_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.055_{\pm 0.000}$
not-MIWAE	$0.005_{\pm 0.003}$	$1.000_{\pm 0.000}$	$0.055_{\pm 0.000}$
EGC	$0.007_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.072_{\pm 0.002}$
MaCoDE	$0.006_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$

Dataset		`breast`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.008_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.080_{\pm 0.000}$
GAIN	$0.014_{\pm 0.001}$	$0.960_{\pm 0.011}$	$0.080_{\pm 0.000}$
missMDA	$0.026_{\pm 0.001}$	$0.853_{\pm 0.015}$	$0.086_{\pm 0.000}$
VAEAC	$0.009_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
MIWAE	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.084_{\pm 0.000}$
not-MIWAE	$0.007_{\pm 0.002}$	$0.990_{\pm 0.016}$	$0.082_{\pm 0.000}$
EGC	$0.007_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.084_{\pm 0.000}$
MaCoDE	$0.009_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.084_{\pm 0.000}$

Dataset		`redwine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.021_{\pm 0.001}$	$0.673_{\pm 0.034}$	$0.048_{\pm 0.000}$
GAIN	$0.026_{\pm 0.002}$	$0.473_{\pm 0.057}$	$0.048_{\pm 0.000}$
missMDA	$0.028_{\pm 0.000}$	$0.600_{\pm 0.028}$	$0.055_{\pm 0.000}$
VAEAC	$0.016_{\pm 0.001}$	$0.782_{\pm 0.036}$	$0.050_{\pm 0.000}$
MIWAE	$0.012_{\pm 0.001}$	$0.909_{\pm 0.019}$	$0.055_{\pm 0.001}$
not-MIWAE	$0.013_{\pm 0.001}$	$0.855_{\pm 0.047}$	$0.053_{\pm 0.001}$
EGC	$0.009_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.072_{\pm 0.001}$
MaCoDE	$0.011_{\pm 0.001}$	$0.990_{\pm 0.010}$	$0.055_{\pm 0.001}$

Dataset		`whitewine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.012_{\pm 0.001}$	$0.673_{\pm 0.039}$	$0.028_{\pm 0.000}$
GAIN	$0.029_{\pm 0.004}$	$0.427_{\pm 0.045}$	$0.028_{\pm 0.000}$
missMDA	$0.015_{\pm 0.000}$	$0.582_{\pm 0.034}$	$0.035_{\pm 0.001}$
VAEAC	$0.012_{\pm 0.001}$	$0.691_{\pm 0.024}$	$0.029_{\pm 0.000}$
MIWAE	$0.010_{\pm 0.001}$	$0.873_{\pm 0.024}$	$0.036_{\pm 0.001}$
not-MIWAE	$0.008_{\pm 0.001}$	$0.891_{\pm 0.072}$	$0.033_{\pm 0.002}$
EGC	$0.008_{\pm 0.000}$	$0.991_{\pm 0.009}$	$0.063_{\pm 0.001}$
MaCoDE	$0.009_{\pm 0.001}$	$0.827_{\pm 0.037}$	$0.034_{\pm 0.001}$

Table 13: **Q3**: Multiple imputation for each dataset under **MNARL** at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes lower is better.

Dataset		`abalone`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.003_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.030_{\pm 0.000}$
GAIN	$0.008_{\pm 0.001}$	$0.886_{\pm 0.036}$	$0.031_{\pm 0.001}$
missMDA	$0.009_{\pm 0.000}$	$0.800_{\pm 0.032}$	$0.033_{\pm 0.000}$
VAEAC	$0.002_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.030_{\pm 0.000}$
MIWAE	$0.005_{\pm 0.001}$	$0.957_{\pm 0.022}$	$0.032_{\pm 0.001}$
not-MIWAE	$0.007_{\pm 0.001}$	$0.943_{\pm 0.023}$	$0.035_{\pm 0.001}$
EGC	$0.003_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.038_{\pm 0.001}$
MaCoDE	$0.007_{\pm 0.001}$	$0.886_{\pm 0.019}$	$0.033_{\pm 0.001}$

Dataset		`banknote`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.008_{\pm 0.001}$	$0.875_{\pm 0.042}$	$0.052_{\pm 0.000}$
GAIN	$0.013_{\pm 0.002}$	$0.825_{\pm 0.053}$	$0.052_{\pm 0.000}$
missMDA	$0.009_{\pm 0.001}$	$0.775_{\pm 0.045}$	$0.054_{\pm 0.000}$
VAEAC	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.053_{\pm 0.000}$
MIWAE	$0.003_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$
not-MIWAE	$0.003_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.054_{\pm 0.000}$
EGC	$0.004_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.066_{\pm 0.001}$
MaCoDE	$0.003_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.053_{\pm 0.000}$

Dataset		`breast`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.004_{\pm 0.000}$	$0.997_{\pm 0.003}$	$0.080_{\pm 0.000}$
GAIN	$0.007_{\pm 0.000}$	$0.997_{\pm 0.003}$	$0.080_{\pm 0.000}$
missMDA	$0.017_{\pm 0.001}$	$0.927_{\pm 0.012}$	$0.084_{\pm 0.000}$
VAEAC	$0.006_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
MIWAE	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
not-MIWAE	$0.004_{\pm 0.000}$	$0.997_{\pm 0.003}$	$0.081_{\pm 0.000}$
EGC	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.082_{\pm 0.000}$
MaCoDE	$0.005_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.083_{\pm 0.000}$

Dataset		`redwine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.010_{\pm 0.001}$	$0.873_{\pm 0.028}$	$0.048_{\pm 0.000}$
GAIN	$0.011_{\pm 0.001}$	$0.827_{\pm 0.032}$	$0.048_{\pm 0.000}$
missMDA	$0.016_{\pm 0.002}$	$0.755_{\pm 0.024}$	$0.052_{\pm 0.000}$
VAEAC	$0.009_{\pm 0.001}$	$0.900_{\pm 0.025}$	$0.050_{\pm 0.000}$
MIWAE	$0.006_{\pm 0.001}$	$0.955_{\pm 0.020}$	$0.052_{\pm 0.000}$
not-MIWAE	$0.006_{\pm 0.001}$	$0.973_{\pm 0.014}$	$0.050_{\pm 0.000}$
EGC	$0.005_{\pm 0.001}$	$1.000_{\pm 0.000}$	$0.063_{\pm 0.001}$
MaCoDE	$0.006_{\pm 0.001}$	$0.982_{\pm 0.018}$	$0.053_{\pm 0.000}$

Dataset		`whitewine`	
Model	Bias ↓	Coverage	Width ↓
MICE	$0.008_{\pm 0.000}$	$0.773_{\pm 0.020}$	$0.028_{\pm 0.000}$
GAIN	$0.012_{\pm 0.002}$	$0.736_{\pm 0.029}$	$0.028_{\pm 0.000}$
missMDA	$0.010_{\pm 0.000}$	$0.809_{\pm 0.021}$	$0.033_{\pm 0.001}$
VAEAC	$0.009_{\pm 0.001}$	$0.736_{\pm 0.029}$	$0.028_{\pm 0.000}$
MIWAE	$0.006_{\pm 0.001}$	$0.964_{\pm 0.020}$	$0.033_{\pm 0.001}$
not-MIWAE	$0.006_{\pm 0.000}$	$0.918_{\pm 0.025}$	$0.031_{\pm 0.001}$
EGC	$0.004_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.055_{\pm 0.001}$
MaCoDE	$0.006_{\pm 0.001}$	$0.855_{\pm 0.031}$	$0.032_{\pm 0.001}$

Table 14: **Q3**: Multiple imputation for each dataset under **MNARQ** at 0.3 missingness. The means and standard errors of the mean across 5 datasets and 10 repeated experiments are reported. ↓ denotes lower is better.

References

[1] Seunghwan An and Jong-June Jeon. Distributional learning of variational autoencoder: Application to synthetic data generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[2] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[3] Yuri Burda, Roger Baker Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *ICLR*, abs/1509.00519, 2015.

[4] Shoja'eddin Chenouri, Majid Mojirsheibani, and Zahra Montazeri. Empirical measures for incomplete data with applications. *Electronic Journal of Statistics*, 3:1021–1038, 2009.

[5] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

[6] Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

[7] Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Manbir S Gulati and Paul F Roysdon. Tabmt: Generating tabular data with masked transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-{miwae}: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*, 2021.

[10] Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019.

[11] Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. HyperImpute: Generalized iterative imputation with automatic model selection. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9916–9937. PMLR, 17–23 Jul 2022.

[12] Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[13] Jayoung Kim, Chaejeong Lee, and Noseong Park. STasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[15] Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[16] Jin Hyuk Lee and J. Charles Huber. Evaluation of multiple imputation with large proportions of missing data: How much is too much? *Iranian Journal of Public Health*, 50:1372 – 1380, 2021.

[17] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[18] Volker Lohweg. Banknote Authentication. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C55P57.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[20] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, 2019.

[21] Stan Matwin, Jordi Nin, Morvarid Sehatkar, and Tomasz Szapiro. A review of attribute disclosure control. In *Advanced Research in Data Privacy*, 2015.

[22] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, 2020.

[23] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

[24] Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

[25] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.*, 11:1071–1083, 2018.

[26] Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[27] Donald B. Rubin and Nathaniel Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–374, 1986.

[28] David Slate. Letter Recognition. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5ZP40.

[29] Latanya Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10:557–570, 2002.

[30] Stef van Buuren. Flexible imputation of missing data. 2012.

[31] Stef van Buuren and Karin G. M. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

[32] William Wolberg, Olvi Mangasarian, Nick Street, , and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5DW2B.

[33] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C5PK67.

[35] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR, 10–15 Jul 2018.

[36] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*, 2024.

[37] Nanhua Zhang, Chunyan Liu, Steven J Steiner, Richard B Colletti, Robert N. Baldassano, Shiran Chen, Stanley Cohen, Michael D. Kappelman, Shehzad Ahmed Saeed, Laurie S. Conklin, Richard Strauss, Sheri Volger, Eileen C. King, and Kim Hung Lo. Using multiple imputation of real-world data to estimate clinical remission in pediatric inflammatory bowel disease. *Journal of Comparative Effectiveness Research*, 12, 2023.

[38] He Zhao, Ke Sun, Amir Dezfouli, and Edwin V. Bonilla. Transformed distribution matching for missing value imputation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42159–42186. PMLR, 23–29 Jul 2023.

[39] Yuxuan Zhao, Alex Townsend, and Madeleine Udell. Probabilistic missing value imputation for mixed categorical and ordered data. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[40] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. Ctab-gan: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 17–19 Nov 2021.

[41] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Yiyu Chen. Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in Big Data*, abs/2204.00401, 2023.