Synthetic Tabular Data Generation for Class Imbalance and Fairness: A Comparative Study

Emmanouil Panagiotou[1,2][0000−0001−9134−9387], Arjun Roy[1,2][0000−0002−4279−9442], and Eirini Ntoutsi[2][0000−0001−5729−1003]

[1] Freie Universität Berlin, Department of Mathematics & Computer Science, Berlin, Germany
emmanouil.panagiotou@fu-berlin.de
[2] Universität der Bundeswehr München, Faculty for Informatik, Munich, Germany

Abstract. Due to their data-driven nature, Machine Learning (ML) models are susceptible to bias inherited from data, especially in classification problems where class and group imbalances are prevalent. Class imbalance (in the classification target) and group imbalance (in protected attributes like sex or race) can undermine both ML utility and fairness. Although class and group imbalances commonly coincide in real-world tabular datasets, limited methods address this scenario. While most methods use oversampling techniques, like interpolation, to mitigate imbalances, recent advancements in synthetic tabular data generation offer promise but have not been adequately explored for this purpose. To this end, this paper conducts a comparative analysis to address class and group imbalances using state-of-the-art models for synthetic tabular data generation and various sampling strategies. Experimental results on four datasets, demonstrate the effectiveness of generative models for bias mitigation, creating opportunities for further exploration in this direction.

Keywords: Synthetic Data · Generative Models · Class Imbalance · Group Fairness · Tabular Data.

1 Introduction

Artificial intelligence (AI) has seamlessly integrated into our daily lives, revolutionizing sectors from personalized online experiences to advanced medical diagnostics. Nonetheless, data collected from real-world sources inherently reflects the biases, prejudices, and inequalities prevalent within society [17]. Consequently, ML models trained on such data have the potential to perpetuate and even exacerbate these biases, leading to unfair or discriminatory outcomes [3].

One significant data-related challenge that can cause biased predictions is population imbalance. The most common, and easily detectable, is class imbalance, which can lead to poor predictive performance for instances in an underrepresented class. Group imbalance, on the other hand, might not directly affect the utility of a model in terms of overall accuracy, but it can lead to unfair treatment of minority groups characterized by some protected attribute (e.g. sex or

race). Several methods have been proposed to address both class-imbalanced learning [15], and fairness [19], yet only a few works study their overlap, which is very common for real-world tabular datasets [17]. Most of these methods are model-specific, meaning they change the workings of existing models to increase performance in minority and majority groups.

Despite the prevalent use of generative methods for synthesizing tabular data, there remains a gap in evaluating their influence on group fairness and class imbalance. In this work, we perform a comparative analysis of *model-independent* generative techniques using oversampling to address class and group imbalances in tabular datasets. While most existing methods rely on the Synthetic Minority Oversampling Technique (SMOTE) [5] for generating additional samples, more recent works on generative AI have developed numerous alternatives for synthesizing tabular data [9]. We cover five generative methods: the probabilistic Gaussian Copula SDV-GC [20], two deep learning models CTGAN and TVAE [27] based on GANs and VAEs respectively, generative non-parametric classification and regression trees CART [22], and the conventional SMOTE-NC [5] for oversampling via interpolation. We also define four sampling strategies for these generative methods and evaluate their performance on ML utility and fairness using four real-world tabular datasets. Our results are benchmarked against training on the original (real) data and a state-of-the-art fair data generator, Tabfairgan [21]. We conclude with an experiment on intersectional fairness, examining the scenario where multiple protected attributes coexist. The full code for this study is available under, github.com/Panagiotou/FairAugment.

The rest of this paper is organized as follows. We describe all relevant works related to fairness, class imbalance, and generative methods in Section 2, we present the problem formulation, dataset details, and evaluation metrics in Section 3, we define all sampling strategies in Section 4, and include all experiments and results in Section 5. We conclude the paper with a discussion and opportunities for future work in Section 6.

2 Related Work

Due to the data-driven nature of ML, inherent biases within the data frequently get amplified or perpetuated, resulting in unfair decision-making. Such bias can arise from imbalanced populations regarding the target class labels, or subgroups defined by protected attributes (e.g. sex, race, etc.). This has led to new research directions towards mitigating such bias and developing fairness-aware models [19,1]. In this section, we cover all related work focusing on overcoming class imbalance, group imbalance (fairness-aware ML), and their simultaneous occurrence. Additionally, we cover synthetic tabular data generation methods, that are relevant to our comparative study.

2.1 Fairness-aware ML

In our study, we focus on *group fairness*, which considers parity over different groups of individuals, distinguished by one or more protected attributes, such

as sex, race, age, etc. Several metrics have been defined to measure the group fairness of a classification model (see Section 3.2). While there are various methods for enhancing group fairness, the main focus is on i) creating methods that specifically optimize for fairness [14], by incorporating constraints to existing models, for example via adding fairness objectives to the loss function [24], and ii) model-agnostic, pre-processing methods [8,16,26] that overcome bias by modifying the training data.

We focus on the second approach, specifically generative pre-processing methods, which rely solely on the training data and mitigate bias by augmenting the existing training data with new samples or sampling an entirely new dataset. For example, the GAN-based Tabfairgan [21] optimizes for accuracy and fairness with consecutive training phases. Once fitted, an entire synthetic dataset is sampled. However, while all of these methods address group fairness, they do not take class imbalance into account.

2.2 Class imbalance in ML

Class imbalance is a common problem in classification problems [17], where a large percentage of the data belongs to a specific class. This scenario is encountered in various domains, such as clinical studies, where the minority class (indicating illness) is under-represented, compared to the majority class (representing healthy individuals). To tackle this issue, similar to fairness methods, many approaches resort to pre-processing techniques like over/under-sampling to mitigate the bias towards the majority class [18,13].

In general, under-sampling methods are not typically favored due to the potential loss of crucial information, which can degrade performance. Similarly, naive over-sampling techniques, such as simply duplicating individuals in the minority class, may lead to overfitting. To overcome this problem, the renowned Synthetic Minority Oversampling Technique (SMOTE) was proposed [5]. SMOTE operates by interpolating between random instances in the minority class and their K-nearest neighbors. This concept has led to various extensions [10] which for example sample specific regions, such as those close to the decision boundary [13], or more sparse areas of the feature space [7]. While such methods improve ML utility by reducing bias towards a certain class, they do not account for group fairness.

2.3 Fairness and class imbalance in ML

Bias in the data related to fairness and class imbalance are not mutually exclusive. More often than not, they occur simultaneously [17], leading to extreme population imbalance for individuals from a minority group who are assigned underrepresented class labels. For example, in the popular Adult dataset, females with a high-income class label are the most under-represented subgroup, accounting for only 11% of the total data (see Figure 1). These populations can become even smaller under "intersectional-fairness" when more than one protected attribute exists [23] or for multi-class classification.

To address this issue, various fair class-balancing methods like FSMOTE [4], FAWOS [25], and other extensions [26,29], have been proposed. The goal is to overcome both fairness and class imbalance via model-independent oversampling. Yet, most of these methods either employ the SMOTE interpolation technique for oversampling or assume a common (discrete) feature type [28], rendering them unsuitable for handling numerical or mixed feature spaces, which are very common in tabular datasets [17].

Nonetheless, recently several generative models have been proposed for generating synthetic mixed tabular data, for example, based on neural networks [27], classification trees [22], probabilistic approaches [11], or even large language models [2]. In this work, we evaluate such synthetic tabular data generation methods (defined in Section 2.4) for class imbalance and fairness, while considering different sampling strategies.

2.4 Synthetic Tabular Data Generation Methods

Various methods have been proposed to learn to generate tabular data [9]. Compared to other modalities such as images or text, tabular datasets consist of a mixture of discrete and continuous feature types, which are difficult to model. Our analysis covers recent approaches for efficient and effective tabular data generation, encompassing state-of-the-art parametric and non-parametric methods.

- **SDV-GC:** Various continuous distributions (e.g. uniform, exponential, etc.) are considered to model all features (discrete features are not explicitly handled, but transformed into continuous). Subsequently, a multivariate Gaussian Copula is used to estimate the covariance between all features. The covariance matrix and the feature distributions are used to sample new synthetic data [20].
- **CTGAN:** The typical generator/critic neural network architecture for Generative Adversarial Networks (GANs) is adapted to learn to generate tabular data. Mode-specific normalization is used during training to overcome imbalances and avoid mode collapse [27].
- **TVAE:** The Tabular Variational Autoencoder [27] trains an encoder/decoder neural network to learn a low-dimensional Gaussian latent space, which is used for sampling new instances through the trained decoder.
- **CART:** A Classification and Regression Tree [22] method for consecutive column-wise data generation via sampling in the leaves, especially suitable for learning inter-dependencies between mixed data due to its non-parametric nature.
- **SMOTE-NC:** SMOTE (Synthetic Minority Over-sampling Technique) [5] is a non-parametric method that generates new samples by interpolating between line segments connecting real instances. The same paper introduces the SMOTE-NC variant, which can support mixed (but not solely discrete) feature spaces.

3 Background

We assume a tabular dataset T containing N_c continuous columns $\{c_1, c_2, \ldots, c_{N_c}\}$ and N_d discrete columns $\{d_1, d_2, \ldots, d_{N_d-1}, d_{prot}\}$ (including categorical, binary, and ordinal features). Additionally, we assume one binary *protected attribute* $d_{prot} \in \{0, 1\}$ (e.g. the sex of an individual), and a binary class label $Y \in \{0, 1\}$. Given such a dataset, any given ML classifier $f()$ can be trained in a supervised manner, on input-target pairs $x_j = \{c_1, c_2, \ldots, c_{N_c}, d_1, d_2, \ldots, d_{N_d-1}\}$ and $y_j \in \{0, 1\}$, $j = \{1, 2, \ldots, n\}$ (the protected attribute is not used during training). Since the class label and the protected attribute are binary features, they partition the tabular dataset T into $|d_{prot} \times Y| = 4$ subgroups $[T_{00}, T_{01}, T_{10}, T_{11}]$.

A generative model G fitted on some subset $\tilde{T}$ of the dataset T, can sample $\tilde{n}$ synthetic rows that comprise a synthetic dataset $\tilde{T}_{syn} = G(\tilde{T}, \tilde{n})$. Further, we refer to a *sampling strategy* $S(n_{00}, n_{01}, n_{10}, n_{11})$ as the method that dictates the number of synthetic samples to be generated from each subgroup in T, to generate a synthetic dataset $T_{syn} = [G(T_{00}, n_{00}), G(T_{01}, n_{01}), G(T_{10}, n_{10}), G(T_{11}, n_{11})]$. The objective of a sampling strategy in our case, is to create an *augmented* final training dataset, denoted as $T_{aug} = T \cup T_{syn}$ which aims to enhance the classifier's performance regarding class imbalance and fairness. We refer to the proportion of the synthetic samples in the augmented dataset as the *augmentation ratio* $r_{aug} = |T_{syn}|/|T_{aug}|$.

In our study, we define and compare various over-sampling methods (Section 4) dictated by the generative models and sampling strategies G, S, aiming to correct both class and group imbalance.

3.1 Datasets

We use four real-world tabular datasets, frequently used in fairness-aware learning [17]. These datasets comprise demographic attributes of individuals, aimed at predicting their financial status, such as occupation, income, credit score, etc. In Table 1 we list the basic characteristics of all datasets, namely, the *Adult*, *German credit*, *Dutch census*, and *Credit card clients*. The protected attribute chosen for all datasets is the binary feature "sex" (Male/Female). We observe class imbalance for the *Adult*, *German credit*, and *Credit card clients* datasets, as well as, a mixed feature space. The *Dutch census* dataset exhibits a less pronounced class imbalance and includes solely discrete features. Both class and group imbalances for all datasets are visualized in the first column of Fig 1.

3.2 Evaluation Metrics

To evaluate the quality of the synthetic data regarding ML utility and fairness, we measure the performance of ML models on the downstream binary classification task for each dataset. In terms of utility, we measure the *Accuracy* and *ROC AUC* score. The last is more suitable for evaluation under class imbalance, as it takes true/false-positive/negative rates into account. With respect to group fairness, we employ widely used fairness metrics, namely equalized odds [12] (Eq.

Dataset	#Instances	#Attributes N_d/N_c	Class Ratio (+)	Protected (Attribute)	Target Class
Adult	45k	9/6	1:3.03	sex	Income
German credit	1k	14/7	2.33:1	sex	Credit score
Dutch census	60k	12/0	1:1.10	sex	Occupation
Credit card clients	30k	10/14	1:3.52	sex	Default payment

Table 1. Overview of all real datasets used in our comparative study

Odds), statistical parity [8] (SP), and equal opportunity [12] (Eq. Opp.). We define all fairness metrics below:

- Equalized Odds (Eq. Odds):
  Assesses the difference between true positive rates (for positive class) and false positive rates (for negative class) for different groups.
  Eq. Odds $= |P[f(X) = 1|Y = 1, d_{prot} = 0] - P[f(X) = 1|Y = 1, d_{prot} = 1]| +$
  $|P[f(X) = 1|Y = 0, d_{prot} = 0] - P[f(X) = 1|Y = 0, d_{prot} = 1]|$
- Statistical Parity (SP):
  Measures whether the probability of a favorable outcome is consistent across different groups defined by the sensitive attribute.
  SP $= P[f(X) = 1|d_{prot} = 0] - P[f(X) = 1|d_{prot} = 1]|$
- Equal Opportunity (Eq. Opp.):
  Measures the difference between the true positive rates (sensitivity) across sensitive attribute groups.
  Eq. Opp. $= |P[f(X) = 1|Y = 1, d_{prot} = 0] - P[f(X) = 1|Y = 1, d_{prot} = 1]|$

All utility metrics should be maximized (the closer to 1 the better), while fairness metrics are minimized (the closer to 0 the better), as they measure differences in performance between subgroups. Since we compare model-independent generative methods, any downstream classification model can be used for evaluation. We choose XGBoost [6], a state-of-the-art gradient boosting model for tabular data classification.

4 Sampling strategies

All generative methods covered in our study (described in Section 2.4) can be trained on a set of tabular data, and then used to generate an arbitrary number of synthetic samples. Given our assumption of a single binary protected attribute and a binary classification task, this results in 4 homogeneous subgroups for each dataset. Additionally, as defined in Section 3, a sampling strategy dictates the number of synthetic samples to draw from each subgroup, to create the final augmented training set. In this work, we propose four such sampling strategies aimed at addressing class imbalance, group imbalance, or both. Namely, *class* and *protected*, sample data to achieve class, and group balance, respectively. Furthermore, *class & protected*, and *class (ratio)*, sample synthetic data to achieve

both class and group parity. We define each sampling strategy in detail hereafter and provide a visual representation of the final distributions of the augmented data for each sampling strategy on all datasets in Figure 1.

Fig. 1. Distributions of class and group imbalance for each real dataset (first column) along with final augmented dataset for each sampling strategy.

- **class:** Separately for each group (Male/Female) we sample instances for the minority class, to match the number of instances in the majority class. Therefore, we achieve a 50/50 class balance for each group.
- **class & protected:** For the largest group (e.g. Male) we sample instances for the minority class, to match the number of instances in the majority class. For all other groups (e.g. Female) we sample for both the majority class and the minority class, to match the number of instances in the majority class in the largest group. Therefore, we achieve the same number of samples for all 4 subgroups. It is worth noting that the *class & protected* strategy is described and used by the FSMOTE method [4] (refer to Section 2.3).
- **protected:** We do not sample for the largest group (e.g. Male), but only for all other groups (e.g. Female), to match the number of instances in the largest group, without considering class labels. Therefore, we achieve the same number of instances for all groups.
- **class (ratio):** We do not sample for the largest group (e.g. Male), but only for all other groups (e.g. Female), to match the class ratio of the largest group. Therefore, we achieve that all groups have the same class ratio as the largest group in the original dataset.

For each sampling strategy in the figure, we report the *augmentation ratio* r_{aug}, as defined in Section 3, i.e. the percentage of synthetic samples in the final augmented dataset. To visualize this, the number of synthetic samples in each bar plot is depicted with a darker color than the real data.

5 Experiments and results

In this section, we present our comparative study, evaluating all generative methods and sampling strategies under utility and fairness. Additionally, we perform an experiment on intersectional fairness, taking multiple protected attributes into account. We conclude with a runtime comparison of all generative methods.

5.1 Experimental setup

We perform experiments for all four datasets, four sampling methods, and five generative models. To ensure robustness, each experiment on the downstream task is 3-fold cross-validated and repeated two times over different random seeds. We report average results over all repetitions, highlighting the best results in bold, and underlining the second-best. For the accumulated results of Section 5.2, we further shade with *blue color* the experiments on synthetic data, which exhibit better performance than training on the original real data (first row).

All experiments are conducted on a single machine equipped with a 12th Gen Intel(R) Core(TM) i9 processor and a Nvidia GeForce RTX 3080 Ti GPU.

5.2 Accumulated results on all datasets

The following Table 3 and Table 4 show the results of our comparative study for the *Adult* and *German credit* datasets, and the *Dutch census* and *Credit card clients* datasets, respectively. As previously mentioned, we present average metrics for all sampling strategies and generative methods. The first two rows in each table (for each dataset) correspond to baselines, i.e. training the classifier using the real data, and synthetic data generated with Tabfairgan [21]. Subsequent rows display results for augmented training data generated through various combinations of the five generative methods and four sampling strategies. We highlight in blue the experiments with superior performance compared to training on the real dataset. Testing (evaluation) is always performed on, previously-unseen, test data from the real dataset.

We interpret the results based on the following criteria:

Accuracy: An initial observation of the results suggests an overall decrease in classifier *accuracy* across datasets when using synthetic data. This is substantiated by relevant literature [27], and can be ascribed to the introduction of out-of-distribution synthetic data by the generative methods.

ROC AUC (class imbalance): Sampling strategies focusing on *class balancing*, such as *class* and *class & protected* improve the ROC AUC score for imbalanced datasets. For the *dutch* dataset, we do not observe any improvement,

due to the lack of inherent class imbalance in the data (see Table 1). Notably, the best ROC AUC score in most cases is achieved with CART-generated data.
Fairness: We observe that the Tabfairgan baseline, although specifically optimized for statistical parity (SP), can also lead to improvements in terms of Eq. Odds and Eq. Opp. However, it is evident that using generative methods, especially with class (ratio) sampling, leads to superior fairness metrics while maintaining higher utility (ROC AUC). This can be attributed to the fact that for most datasets (excluding Dutch census), fewer synthetic samples are needed to achieve equal class ratios between different subgroups, i.e. a lower r_{aug} (see Fig. 1). On the other hand, the class & protected strategy requires the highest number of synthetic samples to maintain class and group balance. This increases the risk of producing out-of-distribution samples, which can degrade performance.
Generative methods: The non-parametric CART model emerges as the top performer in most cases. Notably, despite its simplicity, SMOTE-NC demonstrates performance similar to deep methods, i.e. TVAE, CTGAN. However, it is not applicable for datasets with exclusively discrete feature spaces, such as the Dutch census dataset.

To summarize, sampling strategies like *class* and *class & protected* improve ROC AUC for imbalanced datasets, and CART often achieves the best results. The class (ratio) strategy enhances fairness metrics, generating fewer synthetic samples and maintaining utility.

5.3 Intersectional fairness

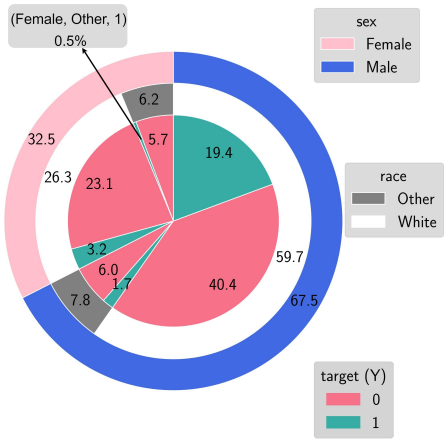

Fig. 2. Sex, race, and class subgroup percentage distributions of the adult dataset.

In all previous experiments, we assume a single binary protected attribute and class label, leading to 4 subgroups (see Section 3). Nonetheless, in some cases

multiple protected attributes can exist, partitioning the data into further groups. To study this effect, we conduct an experiment on the Adult dataset, with race = {White, Other} as the additional protected attribute, splitting the data into 8 subgroups. In Table 5, we present the results when using *race* and *sex & race* (intersection) as protected attributes. In line with our previous results, we observe that the class (ratio) strategy and CART generative method significantly enhance fairness without compromising utility. The non-parametric nature of CART enables consistent generation even from the most under-represented subgroups under extreme data scarcity. For example, in the Adult dataset, the subgroup (sex, race, class) = (Female, Other, 1) accounts for only 0.5% of the total data, i.e. under 200 instances, as seen in Fig. 2.

5.4 Runtime comparison

We conclude our experiments by performing a runtime comparison of all generative methods. We report runtime for, i) training (fitting) on the Adult dataset, ii) sampling 10.000 synthetic instances, and iii) training and sampling, since this overall runtime is the most significant metric in our comparison. From Table 2 it becomes evident that the CART method outperforms all others significantly in terms of overall runtime.

Table 2. Training and sampling runtime comparison for all generative methods on the Adult dataset.

Model	Training time ↓	Sampling time ↓	Overall time ↓
Tabfairgan	187.521 ± 7.64	**0.011** ± 0.005	187.533 ± 7.64
SDV-GC	2.259 ± 0.137	0.163 ± 0.013	<u>2.422</u> ± 0.138
CTGAN	215.49 ± 49.118	<u>0.155</u> ± 0.027	215.65 ± 49.132
TVAE	73.527 ± 7.855	0.159 ± 0.271	73.687 ± 7.957
CART	<u>1.019</u> ± 0.017	0.326 ± 0.002	**1.346** ± 0.018
SMOTE-NC	**0.016** ± 0.004	18.955 ± 1.607	18.971 ± 1.609

6 Conclusion and discussion

Training ML models that take fairness and class imbalance into account is an open problem, with many applications in the real world, especially for tabular datasets. Most model-independent methods perform oversampling by building upon existing methods that generate synthetic samples via interpolation in the minority classes (SMOTE). This comparative study, considers several state-of-the-art generative approaches to synthesize tabular data in each minority class, to overcome bias. Results on four real-world tabular datasets indicate that the non-parametric CART is the better-performing generative method while being the most computationally efficient. In future work, we would like to delve deeper into exploring the capabilities of CART for generating synthetic data, particularly when optimized in the context of fairness.

Table 3. Results for Adult and German credit datasets.

Adult dataset						
Sampling strategy	train-set	Metrics on test-set (real data)				
				sex		
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	**0.868**	0.798	0.122	0.178	0.059
	Tabfairgan	0.539	0.626	0.119	0.123	**0.029**
class	SDV-GC	0.855	0.767	0.116	0.149	0.066
	CTGAN	0.842	0.791	0.147	0.189	0.068
	TVAE	0.846	0.768	0.131	0.168	0.063
	CART	0.836	0.804	0.155	0.174	0.076
	SMOTE-NC	0.839	0.788	0.133	0.181	0.052
class & protected	SDV-GC	0.854	0.766	0.122	0.161	0.065
	CTGAN	0.853	0.792	0.161	0.191	0.083
	TVAE	0.844	0.772	0.169	0.188	0.082
	CART	0.836	**0.814**	0.138	0.194	0.048
	SMOTE-NC	0.840	0.790	0.127	0.188	0.043
protected	SDV-GC	0.853	0.754	0.140	0.155	0.089
	CTGAN	0.854	0.755	0.122	0.148	0.072
	TVAE	0.859	0.768	0.116	0.157	0.064
	CART	0.856	0.763	0.116	0.154	0.065
	SMOTE-NC	0.857	0.758	0.160	0.162	0.105
class (ratio)	SDV-GC	0.857	0.766	0.102	0.151	0.051
	CTGAN	0.860	0.776	0.103	0.152	0.055
	TVAE	0.857	0.767	**0.079**	0.142	0.033
	CART	0.857	0.780	0.137	**0.117**	0.109
	SMOTE-NC	0.854	0.768	0.095	0.124	0.062
German credit dataset						
Sampling strategy	train-set	Metrics on test-set (real data)				
				sex		
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	0.753	0.677	0.088	0.058	0.051
	Tabfairgan	0.639	0.501	0.090	0.042	0.056
class	SDV-GC	0.731	0.686	0.097	**0.019**	0.045
	CTGAN	0.732	**0.704**	0.102	0.037	0.041
	TVAE	0.736	0.675	0.098	0.067	0.044
	CART	0.730	0.695	0.127	0.057	0.047
	SMOTE-NC	0.736	0.687	**0.073**	0.038	0.041
class & protected	SDV-GC	0.726	0.683	0.103	0.024	0.040
	CTGAN	0.730	0.687	0.099	0.052	0.045
	TVAE	0.749	0.690	0.113	0.065	0.052
	CART	0.722	0.680	0.104	0.045	0.045
	SMOTE-NC	0.738	0.681	0.121	0.033	0.063
protected	SDV-GC	0.733	0.668	0.107	0.054	0.063
	CTGAN	0.733	0.652	0.120	0.062	0.053
	TVAE	0.743	0.666	0.113	0.060	0.050
	CART	0.737	0.664	0.132	0.064	**0.033**
	SMOTE-NC	0.739	0.669	0.108	0.046	0.053
class (ratio)	SDV-GC	0.760	0.676	0.084	0.061	0.048
	CTGAN	**0.763**	0.683	0.097	0.044	0.045
	TVAE	0.748	0.675	0.118	0.037	0.053
	CART	0.753	0.675	0.104	0.060	0.056
	SMOTE-NC	0.749	0.667	0.112	0.048	0.055

Table 4. Results for Dutch census and Credit card clients datasets.

Dutch census dataset						
Sampling strategy	train-set	Metrics on test-set (real data)				
				sex		
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	**0.819**	**0.817**	0.092	0.189	0.049
	Tabfairgan	0.804	0.801	**0.078**	0.184	**0.037**
class	SDV-GC	**0.819**	**0.817**	0.093	0.177	0.062
	CTGAN	0.816	0.813	0.097	0.165	0.075
	TVAE	0.815	0.813	0.086	0.171	0.062
	CART	0.816	0.813	0.097	0.157	0.083
	SMOTE-NC					
class & protected	SDV-GC	**0.819**	0.816	0.091	0.178	0.060
	CTGAN	0.817	0.814	0.094	0.167	0.072
	TVAE	0.815	0.813	0.086	0.175	0.059
	CART	0.809	0.805	0.095	0.154	0.080
	SMOTE-NC					
protected	SDV-GC	**0.819**	0.816	0.090	0.188	0.049
	CTGAN	**0.819**	0.816	0.091	0.188	0.049
	TVAE	**0.819**	**0.817**	0.089	0.189	0.047
	CART	0.702	0.689	0.081	**0.090**	0.071
	SMOTE-NC					
class (ratio)	SDV-GC	0.816	0.813	0.091	0.175	0.061
	CTGAN	0.809	0.806	0.090	0.165	0.067
	TVAE	0.809	0.806	0.091	0.168	0.064
	CART	0.807	0.804	0.099	0.155	0.081
	SMOTE-NC					
Credit card clients dataset						
Sampling strategy	train-set	Metrics on test-set (real data)				
				sex		
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	0.812	0.651	0.034	0.023	0.021
	Tabfairgan	0.784	0.597	0.042	0.020	0.032
class	SDV-GC	0.807	0.662	0.045	0.026	0.028
	CTGAN	0.807	0.666	0.044	0.026	0.028
	TVAE	0.804	0.653	0.036	0.026	0.019
	CART	0.757	0.692	0.052	0.033	0.031
	SMOTE-NC	0.766	0.666	0.037	0.023	0.024
class & protected	SDV-GC	0.806	0.662	0.034	0.025	0.019
	CTGAN	0.809	0.644	0.033	**0.019**	0.021
	TVAE	0.806	0.656	0.041	0.021	0.029
	CART	0.760	**0.693**	0.054	0.032	0.034
	SMOTE-NC	0.769	0.668	0.038	0.021	0.026
protected	SDV-GC	0.814	0.649	0.035	0.023	0.021
	CTGAN	0.814	0.652	0.038	0.024	0.025
	TVAE	**0.815**	0.655	0.042	0.028	0.026
	CART	0.813	0.654	0.039	0.025	0.024
	SMOTE-NC	0.813	0.647	0.040	**0.019**	0.027
class (ratio)	SDV-GC	0.813	0.649	**0.032**	0.023	**0.018**
	CTGAN	0.810	0.632	0.034	0.022	0.020
	TVAE	0.814	0.650	0.033	0.021	0.021
	CART	0.813	0.645	0.036	0.022	0.022
	SMOTE-NC	0.813	0.649	0.039	0.023	0.025

Table 5. Results for Adult dataset with *race* and *sex & race* (intersectional) protected attributes.

Adult dataset - race						
Sampling strategy	train-set	Metrics on test-set (real data)				
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	**0.868**	0.798	0.092	0.096	0.064
class	SDV-GC	0.853	0.763	0.073	0.073	0.049
	CTGAN	0.850	0.786	0.063	0.084	0.036
	TVAE	0.846	0.771	0.104	0.088	0.065
	CART	0.834	**0.815**	0.081	0.113	0.028
	SMOTE-NC	0.839	0.792	0.105	0.106	0.052
class & protected	SDV-GC	0.849	0.758	0.069	0.072	0.045
	CTGAN	0.846	0.796	0.057	0.092	**0.022**
	TVAE	0.847	0.766	0.097	0.088	0.060
	CART	0.821	0.790	0.100	0.107	0.046
	SMOTE-NC	0.837	0.793	0.112	0.109	0.057
protected	SDV-GC	0.849	0.761	0.060	0.075	0.042
	CTGAN	0.857	0.767	0.068	0.081	0.045
	TVAE	0.853	0.757	0.066	0.070	0.045
	CART	0.859	0.775	0.067	0.085	0.043
	SMOTE-NC	0.856	0.757	0.081	0.080	0.059
class (ratio)	SDV-GC	0.859	0.770	0.056	0.076	0.034
	CTGAN	0.859	0.769	0.056	0.072	0.036
	TVAE	0.858	0.767	0.070	0.074	0.048
	CART	0.856	0.767	**0.050**	**0.069**	0.030
	SMOTE-NC	0.857	0.770	0.058	0.074	0.034
Adult dataset - sex & race (intersectional)						
Sampling strategy	train-set	Metrics on test-set (real data)				
		Accuracy ↑	ROC AUC ↑	Eq. Odds ↓	SP ↓	Eq. Opp. ↓
	Real	**0.867**	0.798	0.205	0.221	0.131
class	SDV-GC	0.855	0.764	0.209	0.186	0.146
	CTGAN	0.849	0.774	0.192	0.192	0.122
	TVAE	0.848	0.773	0.242	0.213	0.153
	CART	0.835	**0.817**	0.222	0.243	0.114
	SMOTE-NC	0.840	0.790	0.247	0.230	0.143
class & protected	SDV-GC	0.850	0.757	0.188	0.182	0.122
	CTGAN	0.834	0.798	0.211	0.235	**0.110**
	TVAE	0.843	0.756	0.237	0.189	0.163
	CART	0.825	0.797	0.252	0.243	0.133
	SMOTE-NC	0.839	0.797	0.253	0.252	0.135
protected	SDV-GC	0.841	0.752	0.219	0.169	0.168
	CTGAN	0.853	0.763	0.267	0.213	0.188
	TVAE	0.850	0.750	0.210	0.180	0.147
	CART	0.857	0.776	0.233	0.210	0.162
	SMOTE-NC	0.852	0.742	0.263	0.179	0.207
class (ratio)	SDV-GC	0.854	0.757	0.199	0.180	0.139
	CTGAN	0.854	0.764	0.180	0.178	0.120
	TVAE	0.855	0.764	0.173	0.182	0.115
	CART	0.854	0.781	**0.162**	**0.163**	0.118
	SMOTE-NC	0.853	0.768	0.181	0.167	0.128

7 Acknowledgements

Our work is Funded by the Deutsche Forschungsgemeinsschaft (DFG, German Research Foundation) - SFB1463 - 434502799. I further acknowledge the support by the European Union, Horizon Europe project MAMMOth under contract number 101070285.

References

1. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)
2. Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., Kasneci, G.: Language models are realistic tabular data generators. arXiv preprint arXiv:2210.06280 (2022)
3. Brackey, A.: Analysis of racial bias in Northpointe's COMPAS algorithm. Ph.D. thesis, Tulane University School of Science and Engineering (2019)
4. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: Why? how? what to do? In: Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. pp. 429–440 (2021)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
7. Douzas, G., Bacao, F., Last, F.: Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Information sciences **465**, 1–20 (2018)
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
9. Endres, M., Mannarapotta Venugopal, A., Tran, T.S.: Synthetic data generation: a comparative study. In: Proceedings of the 26th International Database Engineered Applications Symposium. pp. 94–102 (2022)
10. Fernández, A., Garcia, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research **61**, 863–905 (2018)
11. Fuchs, R., Pommeret, D., Viroli, C.: Mixed deep gaussian mixture model: a clustering model for mixed datasets. Advances in Data Analysis and Classification **16**(1), 31–53 (2022)
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
13. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. Ieee (2008)

14. Iosifidis, V., Ntoutsi, E.: Adafair: Cumulative fairness adaptive boosting. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp. 781–790 (2019)
15. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
16. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
17. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022)
18. LemaÃŽtre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research **18**(17), 1–5 (2017)
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
20. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). pp. 399–410. IEEE (2016)
21. Rajabi, A., Garibay, O.O.: Tabfairgan: Fair tabular data generation with generative adversarial networks. Machine Learning and Knowledge Extraction **4**(2), 488–501 (2022)
22. Reiter, J.P.: Using cart to generate partially synthetic public use microdata. Journal of official statistics **21**(3), 441 (2005)
23. Roy, A., Iosifidis, V., Ntoutsi, E.: Multi-fairness under class-imbalance. In: International Conference on Discovery Science. pp. 286–301. Springer (2022)
24. Roy, A., Ntoutsi, E.: Learning to teach fairness-aware deep multi-task learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 710–726. Springer (2022)
25. Salazar, T., Santos, M.S., Araújo, H., Abreu, P.H.: Fawos: fairness-aware oversampling algorithm based on distributions of sensitive attributes. IEEE Access **9**, 81370–81379 (2021)
26. Sonoda, R.: Fair oversampling technique using heterogeneous clusters. Information Sciences **640**, 119059 (2023)
27. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)
28. Xu, W., Zhao, J., Iannacci, F., Wang, B.: Ffpdg: Fast, fair and private data generation. arXiv preprint arXiv:2307.00161 (2023)
29. Yan, S., Kao, H.t., Ferrara, E.: Fair class balancing: Enhancing model fairness without observing sensitive attributes. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 1715–1724 (2020)