COPPE
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia
UFRJ

FAIRNESS ASSESSMENT AND MITIGATION IN SYNTHETIC TABULAR DATA GENERATION

Felipe Bevilaqua Foldes Guimarães

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Setembro de 2025

FAIRNESS ASSESSMENT AND MITIGATION IN SYNTHETIC TABULAR DATA GENERATION

Felipe Bevilaqua Foldes Guimarães

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Geraldo Zimbrão da Silva

Aprovada por: Prof. Geraldo Zimbrão da Silva
Prof. Daniel Serrão Schneider
Prof. Victor Stroele de Andrade Menezes

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2025

Guimarães, Felipe Bevilaqua Foldes

Fairness Assessment and Mitigation in Synthetic Tabular Data Generation/Felipe Bevilaqua Foldes Guimarães. – Rio de Janeiro: UFRJ/COPPE, 2025.

XIII, 70 p.: il.; 29, 7cm.

Orientador: Geraldo Zimbrão da Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2025.

Referências Bibliográficas: p. 66 – 70.

1. Fair machine learning. 2. Synthetic data generation. 3. Synthetic tabular data. I. Silva, Geraldo Zimbrão da. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Ao meu avô Hamilton, à Nina e
à Angel.

Agradecimentos

Primeiramente, agradeço aos meus pais, Sergio e Juliana, por serem fonte constante de inspiração e incentivo, sempre me motivando a dar o melhor de mim em todas as etapas da vida.

À minha namorada Larissa, pelo seu apoio em todos os momentos e pela companhia que torna meus dias mais leves.

Sou grato à minha família e aos meus amigos, partes fundamentais da minha vida, por todo o carinho e pela compreensão durante minhas ausências ao longo da realização deste trabalho.

Agradeço ao Prof. Geraldo Zimbrão pela oportunidade de sua orientação, pelos ensinamentos e pelos conselhos durante a minha trajetória acadêmica.

Aos Profs. Filipe Braida, Leandro Alvim e Ygor Cannali, pelas contribuições fundamentais para o desenvolvimento e aprimoramento deste trabalho, enriquecendo-o de forma significativa.

A todos que, direta ou indiretamente, contribuíram para a realização desta dissertação, meu sincero reconhecimento e gratidão.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AVALIAÇÃO E MITIGAÇÃO DE INJUSTIÇA NA GERAÇÃO DE DADOS SINTÉTICOS TABULARES

Felipe Bevilaqua Foldes Guimarães

Setembro/2025

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

A crescente adoção da geração de dados tabulares sintéticos em aplicações de aprendizado de máquina levanta questões importantes sobre as implicações de injustiça dos dados gerados. Este trabalho examina se os modelos de geração de dados sintéticos preservam, amplificam ou reduzem a injustiça dos conjuntos de dados originais, e avalia a eficácia de algoritmos de mitigação de injustiça em dados gerados sinteticamente. Para isso, seis modelos de geração de dados sintéticos foram avaliados em quatro conjuntos de dados de referência da área de injustiça em aprendizado de máquina. Para avaliar a eficácia dos experimentos de mitigação de injustiça, dois algoritmos foram selecionados. Os resultados mostram que os conjuntos de dados sintéticos aumentam sistematicamente a injustiça do classificador em comparação com os dados originais, com aumentos de injustiça variando de modestos a substanciais, dependendo do modelo utilizado. Os algoritmos de mitigação de injustiça permaneceram eficazes em dados sintéticos, alcançando desempenho comparável à sua aplicação em dados reais. Modelos capazes de gerar dados sintéticos de alta utilidade demonstraram as melhores reduções de injustiça após a mitigação. Esses resultados indicam que os dados sintéticos amplificam a injustiça, mas que isso pode ser abordado através de técnicas padrão de mitigação de injustiça quando aplicadas a conjuntos de dados sintéticos de alta qualidade.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

FAIRNESS ASSESSMENT AND MITIGATION IN SYNTHETIC TABULAR DATA GENERATION

Felipe Bevilaqua Foldes Guimarães

September/2025

Advisor: Geraldo Zimbrão da Silva

Department: Systems Engineering and Computer Science

The increasing adoption of synthetic tabular data generation in machine learning applications raises essential questions about the fairness implications of the generated data. This work examines whether synthetic data generation models preserve, amplify, or reduce unfairness from original datasets, and evaluates the effectiveness of fairness mitigation algorithms on synthetically generated data. For this, six synthetic data generation models were evaluated across four fairness benchmark datasets. To assess the efficacy of fairness mitigation experiments, two algorithms were selected. Results show that synthetic datasets systematically increase classifier unfairness compared to original data, with unfairness increases ranging from modest to substantial depending on the model used. Fairness mitigation algorithms remained effective on synthetic data, achieving comparable performance to their application on real data. High-utility synthetic models demonstrated the best fairness improvements after mitigation. The findings indicate that synthetic data amplifies unfairness but that this can be addressed through standard fairness mitigation techniques when applied to high-quality synthetic datasets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Contextualization

The recent emergence and widespread adoption of large language models (LLMs) and image generation models have popularized the term Generative AI. This terminology distinguishes AI models trained to generate new content, such as text, images, videos, sound, or other data types, from traditional AI and ML models designed for predictive tasks such as classification or regression (FEUERRIEGEL *et al.*, 2024). While image and text generation currently drive the generative AI trend, an important question arises regarding the potential role of generative models in the context of tabular data, one of the most widespread data types in practical applications.

Currently, one of the most prominent applications of generative AI in tabular data is synthetic tabular data generation (STDG). In this context, synthetic tabular data refers to data created using machine learning models trained to learn the underlying distribution and relationships from an original dataset. Once trained, these models are able to generate new, realistic data instances that maintain the statistical properties of the original dataset. An illustrative image of this process is presented in Figure 1.1.

This has been particularly explored for privacy-preserving data sharing in sensitive industries such as healthcare and finance (ASSEFA *et al.*, 2020; HERNANDEZ *et al.*, 2022). Additionally, it can be used for augmenting existing datasets in cases involving small or imbalanced data, imputing missing values, among other analytical purposes (FONSECA e BACAO, 2023).

These applications expands beyond research contexts, demonstrating commercial viability across multiple sectors with some significant investments being made in startups in the area. This is exemplified by Mostly AI's $25 million funding round in 2022 (MOSTLY AI, 2022) and NVIDIA's recent acquisition of Gretel for approximately $320 million (SILICONANGLE, 2025).

This theme also has been discussed in the news and courtrooms. A case involving the startup Frank, acquired by JP Morgan for $175 million, underscores potential ethical and legal concerns of the use of synthetic tabular data generation. Post-acquisition, JP Morgan discovered that instead of the claimed four million customers, Frank had only 300,000 real clients. The remaining entries in the database were synthetic data points, generated with assistance from a data science professor, based on the actual customer data (DUNNE *et al.*, 2024).

Generating high-quality synthetic tabular data, however, presents challenges not encountered in image or text generation, particularly in preserving complex inter-column dependencies and ensuring the practical utility of the generated data. These challenges are compounded by the diverse nature of tabular data types, ranging from categorical and numerical to temporal features, each requiring specific consideration in the generation process (XU *et al.*, 2019; ZHAO *et al.*, 2021).

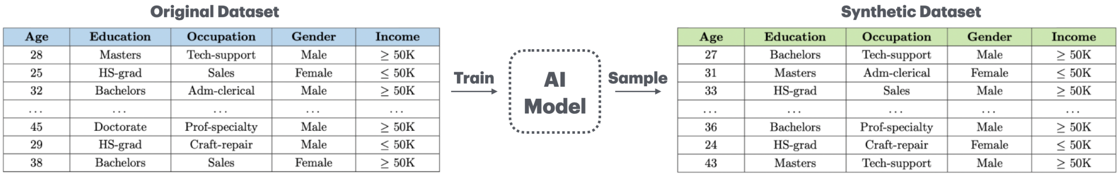
Another common challenge when working with tabular data is the fairness of machine learning models, especially when these models are employed towards automated decision making. Machine learning models can treat in a disparate way individuals from distinct demographical groups, ocasionally incurring in prejudices for individuals that belongs to groups with less privileges in society.

While some research attempts to generate fairer data by introducing modifications in the STDG models training (XU *et al.*, 2018; RAJABI e GARIBAY, 2022; VAN BREUGEL *et al.*, 2021), this approach presents significant challenges. One of the main issues is ensuring that downstream models trained in the synthetic data will behave similarly when evaluated in real data (EITAN *et al.*, 2022).

On the other side, there is evidence that the conventional generation of synthetic tabular data, i.e., without fairness constraints or modifications to generate fairer data, could led to downstream models that show worst fairness behavior than those trained on real data (BHANOT *et al.*, 2021; GANEV *et al.*, 2022; LIU *et al.*, 2025). This is a serious and yet overlooked aspect in the STDG literature that could amplify existing biases in the datasets or even introduce new ones.

This theme has multiple implications across distinct domains where synthetic tabular data is increasingly being adopted. For instance, in healthcare, where synthetic patient data is used for research while protecting privacy, understanding fairness implications is important to prevent the amplification of existing healthcare disparities. In financial services, synthetic data for credit scoring model development must maintain fairness to avoid discriminatory lending practices.

In this context, this research aims to address existing gaps in the STDG literature by conducting an empirical analysis of fairness in state-of-the-art STDG models and exploring potential strategies to mitigate identified biases.



Original Dataset

Age	Education	Occupation	Gender	Income
28	Masters	Tech-support	Male	$\geq$ 50K
25	HS-grad	Sales	Female	$\leq$ 50K
32	Bachelors	Adm-clerical	Male	$\geq$ 50K
...	...	...	...	...
45	Doctorate	Prof-specialty	Male	$\geq$ 50K
29	HS-grad	Craft-repair	Male	$\leq$ 50K
38	Bachelors	Sales	Female	$\geq$ 50K

Synthetic Dataset

Age	Education	Occupation	Gender	Income
27	Bachelors	Tech-support	Male	$\geq$ 50K
31	Masters	Adm-clerical	Female	$\leq$ 50K
33	HS-grad	Sales	Male	$\geq$ 50K
...	...	...	...	...
36	Bachelors	Prof-specialty	Male	$\geq$ 50K
24	HS-grad	Craft-repair	Female	$\leq$ 50K
43	Masters	Tech-support	Male	$\geq$ 50K

Figure 1.1: Illustrative example of AI-driven synthetic tabular data generation.

1.2 Objectives and Research Questions

The primary objective of this work is to investigate the fairness of machine learning classifiers trained on synthetic datasets.

To achieve this, three fundamental research questions were formulated:

RQ1: Do synthetic datasets preserve the unfairness present in the original data?

RQ2: Do synthetic data generation models potentially exacerbate unfairness compared to the original datasets?

RQ3: Can fairness mitigation algorithms effectively reduce unfairness when applied to classifiers trained on synthetic datasets?

These questions address gaps in our understanding of fairness implications in synthetic tabular data generation and will guide the development of this work.

1.3 Contributions

This research makes the following contributions to the field of synthetic tabular data generation:

- A rigorous evaluation framework for assessing fairness in synthetic tabular data generation across six STDG models and four benchmark datasets.
- Empirical evidence that synthetic datasets frequently amplify unfairness present in original data.
- Demonstration that fairness mitigation algorithms effectively reduce unfairness in classifiers trained on high-quality synthetic data.
- Practical guidance for selecting STDG models and applying fairness interventions in fairness-sensitive applications.

1.4 Results Summary

The evaluation of fairness in machine learning classifiers trained on synthetic tabular datasets highlights important implications for fairness-sensitive applications.

Across datasets, most synthetic data generation models led to notable increases in unfairness for at least one fairness measure.

Despite this, applying fairness mitigation algorithms proved effective in reducing these disparities while maintaining competitive performance. Models with stronger predictive performance also tended to show the greatest reductions in unfairness after mitigation.

Overall, these findings suggest that in fairness-sensitive scenarios, a practical strategy is to select synthetic data generation models that produce data closely aligned with the original distribution, and then apply fairness mitigation techniques to address residual bias.

1.5 Document Structure

This Dissertation is structured as follows: Chapter 2 provides a literature review covering synthetic tabular data generation, including main model architectures and evaluation procedures; fairness in machine learning systems, including common fairness measures and mitigation algorithms with emphasis on group fairness concepts in binary classification settings; and the intersection of fairness and synthetic data generation, reviewing both works that attempt to generate fairer datasets and studies indicating fairness implications of synthetic tabular data use. Chapter 3 presents the research motivations, describes the most closely related works, provides an overview of the datasets employed, and concludes with a detailed description of the methodology and experimental setup. Chapter 4 presents the experimental results and their analysis. Chapter 5 concludes with a summary of key findings and recommendations for future research directions

Chapter 2

Literature Review

2.1 Synthetic Tabular Data Generation

In the past decade, the emergence of Generative Adversarial Networks (GANs) (GOODFELLOW *et al.*, 2014) has revolutionized the domain of image generation due to their remarkable ability to produce high-quality and faithful images. This advancement has served as a catalyst for the STDG community to explore the development of machine learning and deep learning techniques for STDG applications. Initially, this exploration focused on GANs and Variational Autoencoders (VAEs) and subsequently expanded to include diffusion models, transformer-based methods, and other new approaches.

2.1.1 STDG Models

One of the pioneering techniques in successfully utilizing GANs for generating synthetic tabular data is known as Conditional Tabular GAN (CTGAN) (XU *et al.*, 2019). The authors of this study addressed several challenges associated with generating synthetic tabular data using GANs, including the presence of both numerical and categorical data types within tabular datasets, learning from sparse one-hot-encoded vector representations, and handling non-Gaussian and multimodal data distributions commonly found in such datasets.

To overcome these challenges, the authors make some contributions, such as introducing mode-specific normalization to overcome the non-Gaussian and multimodal distribution, adding a conditional generator and training-by-sampling to deal with the imbalanced discrete columns, and also introducing some changes to the network structure. The architecture adopted consisted of two fully-connected hidden layers in both the generator and the critic.

The generator utilized batch normalization and the ReLU activation function. After the two hidden layers, the synthetic row representation was generated using a combination of activation functions. The scalar values α_i were generated using the tanh function, while the mode indicator β_i and discrete values d_i were generated using the Gumbel softmax. In the critic, the leaky ReLU function and dropout were applied to each hidden layer. To prevent mode collapse, the authors utilized the PacGAN framework with 10 samples in each pac. The model was trained using the WGAN loss with gradient penalty.

In this paper, the authors also introduced TVAE, a modification of VAEs for tabular data generation. TVAE uses a conventional VAE architecture, with an encoder to map inputs to a latent distribution and a decoder to reconstruct data from latent samples. Unlike standard VAEs which typically handle continuous data with uniform loss treatment, TVAE employs a modified ELBO loss with a modification on the reconstruction term, where it differentiates between categorical features and continuous features, adopting a specific loss for each of those types. Additionally, TVAE's architecture includes the same preprocessing used for CTGAN.

In ZHAO *et al.* (2021) while proposing a new GAN based method named CTAB-GAN, the authors explore further the complexities associated with various data types in the synthetic tabular data generation and proposed the introduction of a new category in variable modeling: the mixed data type. This novel category can be interpreted as a fusion of categorical and continuous variables. For illustrative purposes, the authors presented an instance of a 'mortgage loan holder' variable. This variable could either possess a categorical value of '0', indicating a lack of a loan, or a continuous value representing the presence of a loan.

Additionally, they examined the previous methods capacity for handling skewed continuous variables and imbalanced categorical variables, and provided an enhanced methodology for dealing with these issues. For skewed continuous variables, they utilized a log transformation, asserting that despite its simplicity, it yielded positive outcomes. To manage imbalanced categorical variables and their co-occurrence with skewed continuous variables, they introduced a novel encoding technique for the conditional vector.

The same authors latter expanded their work in ZHAO *et al.* (2022) with CTAB-GAN+, in which they introduce a couple novelties to their previous work such as adding downstream losses to conditional GANs for higher utility synthetic data, using Wasserstein loss with gradient penalty for better training convergence, introducing novel encoders targeting mixed continuous-categorical variables and variables with unbalanced or skewed data and adding the possibility of training with DP stochastic gradient descent to impose strict privacy guarantees.

In recent years, the machine learning community has witnessed the emergence and success of diffusion models within the domain of image generation, an area previously dominated by GANs. This inspired the authors of TABDDPM (KOTEL-

NIKOV *et al.*, 2023) to investigate the application of diffusion models to the generation of synthetic tabular data. As in other diffusion models, TABDDPM operates by gradually adding noise according to a fixed schedule in the forward process and then learning to reverse this process by predicting the noise components. To addresses challenges of heterogeneous data types in tabular data, TABDDPM employs a hybrid architecture that combines Gaussian diffusion for continuous features and multinomial diffusion for categorical variables.

Meanwhile, since 2017 the deep learning community has witnessed the success of transformer based architectures in many fields, particularly, Large Language Models (LLMs) became the state-of-art for natural language processing. In this context, BORISOV *et al.* (2023) developed GReaT (Generation of Realistic Tabular data), a model that exploits an auto-regressive generative LLM to generate synthetic tabular data. To adapt this model to be successful for tabular data generation, their methodology, comprise of two main parts, the finetuning of a pretrained auto-regressive LLM on a textual representation of the tabular dataset and the sampling from this finetuned model.

The textual encoding of the tabular dataset is done by transforming the information stored in the rows and columns to a textual representation of each row. For example, an instance of a tabular dataset with a feature named Age with value of 39, a feature named Education with value 'Bachelors', and a feature named Gender with value 'Male', would be represented in this work with the phrase "Age is 39, Education is Bachelors, Gender is Male". Also, random permutations of this textual representation are used to make the model invariant to the order of the features. For the fine-tuning phase, two different pretrained transformer decoder LLM models of various sizes were used, GPT-2 and it smaller distilled version DistilGPT2.

Building on transformer-based architectures for tabular data synthesis, RealTabFormer (Realistic Relational and Tabular Transformer) (SOLATORIO e DUPRIEZ, 2023) represents another significant advancement in the field as it was designed for the generation of singular tabular synthetic datasets and also for relational data. For generating synthetic tabular data, the data is encoded in the same way as in GReaT and the fine-tuning is done in a similar way. One notable difference to GReat is that RealTabFormer uses target masking in the training phase for prevent data copying and uses a early stopping criteria based on quantiles values distance betwen synthetic and real features. For generating relational data, first, a parent table is first generated and then a sequence-to-sequence (Seq2Seq) model is used to generate the relational tables.

Despite the recent focus in deep learning models for synthetic tabular data generation, other approaches also have been explored. One example, is ARF (Adversarial Random Forests) (WATSON *et al.*, 2023), a synthetic tabular data generation model based on the classical machine learning ensemble model random forests. ARF employs an adversarial framework inspired by GANs where the joint density distribution of a dataset is learned by recursively fitting a random forest model to classify between synthetic and real samples untill a convergence criteria is met. For generating synthetic data, the trees structures of the model are explored with another algorithm named FORests for GEnerative modeling (Forge).

2.1.2 STDG Evaluation

Evaluation of synthetic generated data is a critical aspect in the use and development of STDG methods, as it's quality is essential for any practical downstream use of this data. This evaluation can be done with relation to multiple aspects of the data with utility, resemblance and privacy being the most common aspects evaluated (HERNADEZ *et al.*, 2023).

Resemblance, also referred to as fidelity or quality evaluations, measures the degree to which the synthetic data resembles the original data. Evaluation methods in this area range from basic checks, such as comparing numerical ranges and categorical frequencies, to more sophisticated statistical tests and distances calculations to compare the similarity between synthetic and real data. Visualization techniques, like distribution plots and dimensionality reduction, also support qualitative comparison between real and synthetic datasets.

Notably, both CTABGAN+ (ZHAO *et al.*, 2022) and Tabddpm (KOTELNIKOV *et al.*, 2023) adopts Jensen–Shannon Divergence and Wasserstein Distance to measure distribution similarity for categorical and continuous features, respectively. Also, in these works, a measure of distance between correlation matrices of synthetic and real data is also used. RealTabFormer and GReaT (BORISOV *et al.*, 2023) uses visualizations of joint distributions for qualitative evaluation of the generated data.

Machine Learning Utility (or efficacy) is one of the most prevalent measures of evaluation in the recent STDG literature. This measure is based on the performance of machine learning models trained on the synthetic data and evaluated in a test set of the original data. Once calculated, the performance of this machine learning model is compared to the performance of the same model trained on the original data. The closer the performance of the synthetic data-trained model is to its original data-trained counterpart, the higher is the utility of the synthetic data and consequently, of the STDG model itself.

A common approach for calculating this measure is to calculate the mean performance gap across multiple distinct downstream machine learning models. CTGAN, CTABGAN+, GrEAT, and others follow this methodology. In Tabdppm, however, the authors compare this approach to an alternative that uses only Catboost

(PROKHORENKOVA *et al.*, 2018) to evaluate the performance gap. They adopt this single-model approach because they believe it better reflects real-world scenarios faced by researchers and practitioners, as gradient boosting models are standard in most tabular data applications. After comparing these two approaches, they argue in favor of the single-model method, noting that discrepancies in model complexity within the multi-model approach can lead to misleading conclusions about performance gaps. Furthermore, the model used in this single-model approach had its hyperparameters tuned on a real data validation set, which contrasts with previous approaches that used models with default hyperparameters. In Realtabformer, the authors build upon this approach but instead of relying on a real data validation set, they generate a synthetic validation set for tuning the hyperparameters, as it is a scenario closer to what practitioners would face in real scenarios using the synthetic data.

Privacy evaluation is the third critical aspect in assessing synthetic data quality, ensuring that the generated data does not inadvertently disclose sensitive information from the original dataset. This aspect is fundamental for cases where the main goal of synthetic data generation is to facilitate data publishing or to address privacy concerns. However, even when this is not the primary purpose, and synthetic data is generated for augmentations, imputation, and other objectives, privacy-related evaluations can serve valuable purposes such as detecting overfitting (SOLATORIO e DUPRIEZ, 2023).

Distance to Closest Record (DCR) is one of the most common privacy-related measures in the STDG literature. It calculates the distance between synthetic and real records using metrics such as L1 or L2 distances to assess how close the synthetic samples are to the real ones. DCR values close to zero indicate that models might be memorizing records from the real data, while higher DCR values suggest better generalization. However, high DCR could also indicate excessive noise, which is why it should be evaluated alongside other measures such as Machine Learning Utility (KOTELNIKOV *et al.*, 2023).

2.2 Machine Learning Fairness

With the rapid adoption of machine learning systems governing diverse applications that affect people's daily lives, such as automated decision-making systems in credit approval processes, e-commerce recommendations, and social media content curation, concerns regarding ethical implications and potential biases in these systems have attracted significant attention from media outlets, industry stakeholders, and academic researchers.

The study of biases in machine learning systems that can prejudice some unprivileged group represented in the data is known in the academic literature as "Fairness", "Algorithmic Fairness" or "Machine Learning Fairness", among others common terms. Multiple concepts exist regarding what defines a fair machine learning system, which can vary according to the specific context, cultural aspects, among others. Considering this, a variety of measures have been proposed to align with different fairness concepts.

Overall, this phenomenon is typically studied by defining one or more sensitive classes. These classes include unprotected (also known as unprivileged) groups, which usually represent socio-demographical groups with less power or representation in society, and a non-sensitive (also known as privileged) class, which represents a group with more privileges in society. After defining these groups, the fairness of a machine learning model is evaluated by comparing the model's results on these groups on an evaluation set.

There can be many different sources of unfairness in machine learning models, such as biases present in the data due to historical societal biases, inequalities in the data acquisition process, or even from the optimization process of the model (MEHRABI *et al.*, 2021).

While the majority of existing research in this field focuses on binary classification settings with a single sensitive variable, significant work has also emerged addressing regression problems, multi-class classification, and scenarios involving multiple sensitive variables.

2.2.1 Fairness Measures

Existing definitions and respective measures can be categorized into group, individual or counterfactual fairness measures. Group fairness concepts focus on treating different groups equally based on each group statistics, individual fairness definitions, also know as similarity-based criteria, aims to give similar predictions to similar individuals, causal definitions, for instance, are related to individual definitions but are based on the use of counterfactual scenarios for dealing with causal relationships.

Among the group fairness definitions, Demographic Parity, Equalized Odds, Predictive Equality and Equality of Opportunity, are the most used. These measures aims for different objectives and may be non-compatible between themselves. Demographic Parity 2.1 is one of the most commonly adopted in the fairness literature, the concept around it is about giving the same rate of positive predictions among distinct groups. As can be noted, Demographic Parity measures depends only on the predictions $\hat{y}$ and the sensitive variable S and does not depends on the true labels y.

Predictive Equality, Equality of Opportunity and Equalized Odds, for instance, are based on the disparities in error rates between distinct groups. This distinction is crucial because it highlights a fundamental trade-off in fairness metrics. While Demographic Parity ensures equal representation in positive predictions across groups, it may inadvertently penalize a model that accurately reflects legitimate differences in base rates between groups. In contrast, error-rate-based metrics like Predictive Equality, Equality of Opportunity, and Equalized Odds incorporate the ground truth labels, allowing them to assess whether a model's mistakes are distributed fairly while still maintaining predictive accuracy. Predictive Equality 2.2 specifically focuses on equal false positive rates across groups, Equality of Opportunity 2.3 ensures equal true positive rates, and Equalized Odds 2.4 requires both false positive and true positive rates to be equal across groups.

$$\text{Demographic Parity} = \Pr(\hat{y} = 1 \mid S = 1) - \Pr(\hat{y} = 1 \mid S = 0) \tag{2.1}$$

$$\begin{aligned}\text{Predictive Equality} = \Pr(\hat{y} = 1 \mid S = 1, y = 0) - \Pr(\hat{y} = 1 \mid S = 0, y = 0) = \\ \text{FPR}_{\text{Sensitive}} - \text{FPR}_{\text{Non-Sensitive}}\end{aligned} \tag{2.2}$$

$$\begin{aligned}\text{Equality of Opportunity} = \Pr(\hat{y} = 1 \mid S = 1, y = 1) - \Pr(\hat{y} = 1 \mid S = 0, y = 1) = \\ \text{TPR}_{\text{Sensitive}} - \text{TPR}_{\text{Non-Sensitive}}\end{aligned} \tag{2.3}$$

$$\text{Equalized Odds} = \max_{y \in \{0,1\}} \left|\Pr(\hat{y} = 1 \mid S = 1, y) - \Pr(\hat{y} = 1 \mid S = 0, y)\right| \tag{2.4}$$

2.2.2 Fairness Mitigation Techniques

Over the past two decades, researchers have developed various methods to mitigate fairness-related issues as research on algorithmic fairness has evolved and gained attention from both media and academia. These methods are typically categorized into pre-processing, in-processing, or post-processing based on the stage of the prediction pipeline where they are applied. Pre-processing algorithms perform transformations on the data before it reaches a machine learning model to ensure fair outcomes. In-processing algorithms modify the training of machine learning models, while post-processing algorithms modify the outputs of a model that has already been trained (MEHRABI *et al.*, 2021).

One of the earliest fairness pre-processing methods is the Massaging algorithm, developed by KAMIRAN e CALDERS (2009). In this algorithm, a biased ranker is learned to predict the class attribute without considering the sensitive attribute. It identifies two groups of objects in the training data, candidates for promotion and demotion for the target class, and uses the ranker to select the best candidates for both. The training data is modified until discrimination is eliminated, and a new classifier is trained on the modified data. Other pre-processing methods include fair clustering, fair dimensionality reduction, fair feature selection, among others.

In-processing methods are one of the most studied in the fairness literature. Most of the works about in-processing methods are based on the addition of fairness constraints or regularization terms to the optimization process of machine learning models to achieve fairer outcomes. In many cases, these constraints or regularization terms can be the fairness metrics themselves or a proxy variable for them.

An early work in this field is Fairness Through Awareness (DWORK *et al.*, 2012), which formulates a fair classifier as a constrained optimization problem. This problem minimizes classification loss while ensuring that similar individuals have similar outcomes, promoting individual fairness. Another early method is the work of KAMISHIMA *et al.* (2011), which quantified the degree of prejudice based on mutual information and added this as a regularizer in a logistic regression model. Similarly, ZEMEL *et al.* (2013) introduced the concept of fair representation. They used demographic parity as a regularizer to learn data representations that adhere to specific fairness criteria, ensuring that these representations would satisfy these criteria when used by a classifier.

Building on this line, distinct methods have also explored the concept of fair representation, particularly using neural network models such as Autoencoders and VAEs LOUIZOS *et al.* (2016) trained with fairness loss terms to learn these representations. Neural networks have also been used with fair loss terms employed directly from the model predictions, as in PADALA e GUJAR (2020), where distinct fairness notions are expressed in terms of fair loss functions that are added to the usual training of a two-layer MLP. Adversarial approaches also have been explored, as in MADRAS *et al.* (2018) for example, where a discriminator tries to learn from a latent representation if an instance belongs to the protected group or not, and uses the results of this discriminator a loss term in an adversarial framework similar to the ones used in GANs (GOODFELLOW *et al.*, 2014).

Post-processing methods by the other side are less explored, with significantly fewer research studies conducted on this topic compared to pre-processing or in-processing methods. A remarkable example of post-processing methods is the Threshold Optimizer method (HARDT *et al.*, 2016) where the algorithm adjusts the decision threshold after the model to achieve a desired fairness condition. This method allows for flexibility in modifying the performance of the classifier without altering the underlying model. In particular, it can be used to meet specific fairness criteria such as demographic parity or equalized odds. Another noteworthy post-processing technique is the Reject Option Classification (KAMIRAN *et al.*, 2012), which focuses on instances where the classifier is uncertain. By favoring decisions that promote fairness in these ambiguous cases, this method helps improve overall equity without significantly compromising accuracy.

While there is a great variety of fairness mitigation methods published in the literature, most of these methods are not agnostic to the model choice and to the fairness definition. This is a serious limitation for practical applications as it restricts the flexibility and generalization of fairness interventions. In real-world scenarios, organizations often work with diverse datasets, models, and fairness requirements that evolve over time. Furthermore, most methods in the fairness literature require neural networks, while the most adopted and most performative models for tabular problems are still Gradient Boosting variants (BORISOV *et al.*, 2022).

Two examples of methods that can be applied to distinct classifiers and fairness measures are the Threshold Optimizer method (HARDT *et al.*, 2016) and the reductions method (AGARWAL *et al.*, 2018), both available on the Python fairness library Fairlearn (WEERTS *et al.*, 2023).

The reduction algorithms (AGARWAL *et al.*, 2018) treat any standard classification or regression algorithm as a black box, and iteratively re-weight the data points and retrain the model after each re-weighting. The algorithm transforms the complex fair classification problem into a sequence of standard cost-sensitive classification problems and formulates fair classification as a constrained optimization problem where fairness constraints are expressed as linear inequalities on conditional moments. The algorithm iteratively alternates between finding optimal classifiers given current fairness penalty weights and updating these weights based on constraint violations, ultimately converging to a randomized classifier that achieves the optimal accuracy-fairness trade-off.

2.3 Fairness in Synthetic Tabular Data

The intersection between STDG methods for tabular data and fairness has been explored in two main directions, one that attempts to mitigate fairness issues by using STDG models to modify the original dataset and another that focus on studying the fairness issues of synthetic generated data. In the first case, in studies that attempt to mitigate fairness through the use of synthetic data, it usually can be done in two ways, by adopting fairness constraints in the data generation process or by using synthetic generated data to augment original datasets to make then fairer.

One of the first approaches for mitigating fairness by using a entirely synthetic dataset generated by a model trained on real data was FairGAN (XU *et al.*, 2018).

In this work, the authors employ a GAN framework with one generator conditional on the sensitive variable and two discriminators, one for distinguishing between synthetic and real samples, as usual in the GAN framework, and another to distinguish if the synthetic generate data is from the protected group or not. The result is a generator whose synthetic generated data cannot have it's sensitive variable distinguished from the others variables in the dataset.

Another relevant work in this area is DECAf (VAN BREUGEL *et al.*, 2021), a causal GAN-based framework that generates synthetic data using multiple generators (one for each variable) to learn causal conditionals from observed data, with variables synthesized topologically from root to leaf nodes in the causal graph. The method enables bias removal through targeted edge removal at inference time, allowing for the creation of datasets that satisfy various fairness definitions.

Other related work is TabFairGAN (RAJABI e GARIBAY, 2022) where the authors employ a fair loss penalty on the generator based on the disparities between the relative frequencies of the target variable for the sensitive and non-sensitive groups. By doing this, the models are able to generate synthetic data that is balanced in respect to the sensitive variable.

In this context of fair data generation, a very important and challenging aspect is that training a downstream model on fair synthetic data does not necessarily lead to fair models, as stated in EITAN *et al.* (2022). In this work, the authors use TabFairGAN and show empirically that even thought the data generated by the model is mostly fair when judged by the demographic parity of the data, the models trained on this data do not necessarily exhibit a fair demographic parity when evaluated on the real test data.

Addressing this challenge would require the development of methods to generate fair datasets that, when used to train classifiers without explicit fairness constraints, would produce fair predictions when applied to original test data, despite the inherent distribution shift between synthetic and real data. This represents a complex, ill-posed problem that remains inadequately understood in current literature.

Although synthetic data has been used to tackle fairness challenges, it has also been shown that synthetic generated datasets can maintain or even exacerbate existing biases in the real data when these datasets are generated in the usual generation setting, i.e. without considering fairness in the generation process (GANEV *et al.*, 2022; CHENG *et al.*, 2021). This has been shown specially for differentially private synthetic data (GANEV *et al.*, 2022; CHENG *et al.*, 2021) but also for regular synthetic data.

These works that conduct fairness evaluations on synthetic tabular datasets are more directly related to this study. For this reason, they will be discussed in detail in Section 3.2, along with an analysis of how they differ from this work.

Chapter 3

Methodology

This section details the methodology of this work. First, it shares the motivation behind this study. Then, it discusses the most closely related work, highlighting its similarities and differences with this study. Subsequently, it describes the datasets used in this work. Lastly, it presents the proposal for this study and provides a detailed explanation of the experimental setup.

3.1 Motivation

In recent years, there has been a significant increase in research and industry applications of synthetic tabular datasets. These datasets are generated by machine learning models that are trained to learn the complex distribution of tabular datasets in a way that allows them to be used to sample new instances from these distributions.

These datasets can then be utilized in privacy-preserving applications, to augment existing datasets, and for other analytical purposes. Despite the recent growth in the area, several challenges still exist, such as ensuring that the synthetic data closely follows the real distribution of the original data without copying or leaking private information.

Another critical challenge when working with tabular data is ensuring the fairness of machine learning models trained on a given dataset. This theme has garnered attention from the media and research community over the last decade. Still, it has been overlooked in synthetic tabular data generation, despite evidence that synthetic tabular datasets may either preserve or exacerbate fairness issues in the original data (GANEV *et al.*, 2022; LIU *et al.*, 2025).

3.2 Related Work

As discussed in the subsection 2.3, fairness in STDG models is typically examined in two distinct ways. One approach involves utilizing modified STDG models to generate fair data. The other approach, which is more directly related to this work, focuses on assessing the fairness of standard STDG models.

Following this second approach, BHANOT *et al.* (2021) investigated the presence of unfairness in synthetically generated datasets created using the HealthGAN (YALE *et al.*, 2020) model. They evaluated the fairness of these synthetic datasets using two novel fairness metrics, which were designed to assess the similarity between the fairness properties of the synthetic and original datasets, such as the rate of protected groups and the preservation of the relationship between the sensitive variable and time series covariates.

Similarly, GANEV *et al.* (2022) demonstrated that STDG models trained with differential privacy constraints exhibit disparate effects on class and subgroup sizes, which can impact classifier accuracy, particularly for underrepresented classes. These effects intensify with stronger privacy guarantees and higher levels of data imbalance. Also, examining STDG models with differential privacy, PEREIRA *et al.* (2024) compares marginal-based and GAN-based synthetic data generators, revealing mixed results regarding fairness for GANs, with cases where training classifiers on GANs' synthetic data reduced fairness metrics but at a drastic performance cost, probably related to poor fidelity to the original data distribution.

While previous works focused on differential privacy STDG models or had limited scope, testing only a few STDG models and datasets, the work that most closely relates to this study is LIU *et al.* (2025), which conducted a more comprehensive analysis. In their research, the authors assessed the fairness of classifiers trained on synthetic datasets generated by multiple STDG models. Furthermore, they applied fairness preprocessing methods to evaluate their effectiveness on synthetic data. Their findings indicate that different STDG models achieve varying balances between fairness, privacy, and utility, and that preprocessing fairness mitigation algorithms can help reduce unfairness in synthetic tabular datasets.

While their work shares similarities with the present study, there are several differences between them. First, LIU *et al.* (2025) focused on Learning Analytics (LA) datasets, whereas this work examines common fairness benchmark datasets. Second, their exploration included utility-focused, privacy-focused, and fairness-focused STDG models, while this study concentrates specifically on state-of-the-art utility-focused STDG models. Third, the fairness mitigation algorithms adopted differ between the studies. Additionally, the experimental setups vary significantly. In LIU *et al.* (2025), no hyperparameter tuning or cross-validation was performed for STDG model training, although cross-validation and hyperparameter tuning were applied to downstream classifiers. In contrast, this work employs a cross-validation framework for the entire workflow, encompassing both STDG model training and downstream classifier evaluation, with hyperparameter tuning applied to STDG models but not to downstream classifiers.

3.3 Datasets

Datasets featured in fairness-related research span various sectors, including finance, criminal justice, and healthcare. Recent work has reviewed the most common datasets used in fairness machine learning research (LE QUY *et al.*, 2022). The authors mention 15 datasets cited by at least three papers in the fairness domain. Among them, the four most commonly used datasets encountered in the fairness literature were Adult (BECKER e KOHAVI, 1996), COMPAS, German and Bank Marketing, in this exact order. These datasets were chosen for the experiments in this work.

The first of them, Adult, also known as Census Income, is a dataset that contains demographic and socio-economic data from a representative sample of individuals who responded to the March 1994 US Current Population Survey. The amount by which each entry represents the total population is given by the variable 'fnlwgt'. The dataset has 15 variables and 48842 instances and is used in the machine learning literature for the prediction of the variable 'income', which represents if the individual has an annual income greater than 50,000 US dollars. It is used in the fairness literature to study the fairness in relation to attributes such as gender and race. In this work, the variable 'sex', which represents the person's gender, was used with women defined as the protected group.

The second dataset, COMPAS, originates from the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software, a tool designed to predict risk scores for individuals recidivism on criminal charges. This software has been used by several justice courts in the United States and it has received attention from the media and research community after the studies from ProPublica revealed racial disparities in the algorithm's predictions (ANGWIN *et al.*, 2016).

Since then, it has become an important source for public discussions and studies on the topic of fairness in machine learning systems. The dataset is available in two versions, one with violent charges and the other with both violent and non-violent ones. The latest, is adopted in this study, and it comprises 7214 instances and 11 attributes. The attribute 'two_year_recid', a binary variable indicating if the individual has or has not reincidivated in crime offenses, is used as target variable and the feature 'race' is used for the sensitive variable with the class 'Caucasian' being defined as the privileged group and the remaining classes as the protected group.

The German dataset, also known as German Credit Dataset or Statlog, is a dataset of credit approval for 1000 individuals from a German bank. It contains 20 attributes, including both numerical and categorical variables, such as age, employment status, credit history, and purpose of the loan. The target variable for this dataset is the attribute' class label' which represents if the individual has a good or bad credit risk. The sensitive variable adopted for the german dataset is the attribute 'sex' which indicates the person gender.

Lastly, the Bank Marketing dataset is a dataset from a Portuguese banking institution, containing information related to direct marketing campaigns conducted via phone calls. The dataset has 45211 instances and 17 attributes, which include variables such as age, job, marital status, education, among others. The target variable for this dataset is the variable 'y' which indicates if the person has subscribed a term deposit or not. Both the attributes 'marital-status' and 'age' have been used as protected attributes in this dataset LE QUY *et al.* (2022). In this work, the variable 'age' was used as a sensitive variable with persons with ages less than 25 or greater than 60 being the protected group.

These characteristics of the datasets are summarized in Table 3.1. As evident from the table, these datasets have distinct characteristics, including their purpose, size, and the sensitive attribute used. The datasets span a range, from the smallest German dataset with 1,000 rows to the largest Adult dataset containing 48,842 samples. The COMPAS dataset and Bank Marketing dataset fall in between. The number of features also varies, from 10 columns in COMPAS to 22 in the German dataset.

Furthermore, the datasets exhibit distinct imbalance levels. As evident from the target positive rate information in Table 3.1 and the barplots in Figure 3.1, Bank Marketing has the highest imbalance, with only 11.7% of the positive group of the target variable occurring. On the other hand, the COMPAS dataset is the most balanced, with 45.07% of positive values.

The rates of sensitive attributes also differ significantly. The lowest rate is found in the Bank Marketing dataset, with 5.74% of the protected group occurring. The COMPAS dataset has the highest rate of the protected variable with 65.98% of instances from Non-Caucasians.

In all datasets, the rates of positive values in the target variable differ based on the sensitive variable, which can be a source for fairness concepts like Demographic Parity. The Adult and Bank Marketing datasets exhibit the most significant variation, with the difference in target rates between one class being three times greater than in the other. In the Adult dataset, the lower target rate is observed in the protected group, while in the Bank Marketing dataset, the situation is reversed.

The sensitive attributes chosen for each dataset follow conventions in fairness literature. Beyond academic conventions, these attributes reflect critical dimensions of social inequality. For instance, gender in the Adult and German datasets addresses wage gaps and financial discrimination against women. Race in COMPAS highlights systemic bias in criminal justice, particularly given documented racial disparities in recidivism prediction. Age in Bank Marketing addresses discrimination against both younger and older populations in financial services.

Table 3.1: Overview of datasets used in the study with their characteristics and fairness-related attributes.

Dataset	**Adult**	**Compas**	**German**	**Bank Marketing**
Rows	48,842	7,214	1,000	45,211
Columns	15	10	22	17
Target Column	income	two_year_recid	class-label	y
Target Value	>50K	1	2	1
Sensitive Column	sex	race	sex	age
Protected Group	Female	Non-Caucasian	female	<25 or ≥60
Protected Group Rate	33.15%	65.98%	31.00%	5.74%
Target Positive Rate	23.93%	45.07%	30.00%	11.70%
Target Positive Rate (Privilege)	30.38%	48.00%	27.68%	10.52%
Target Positive Rate (Protected)	10.93%	39.36%	35.16%	31.12%

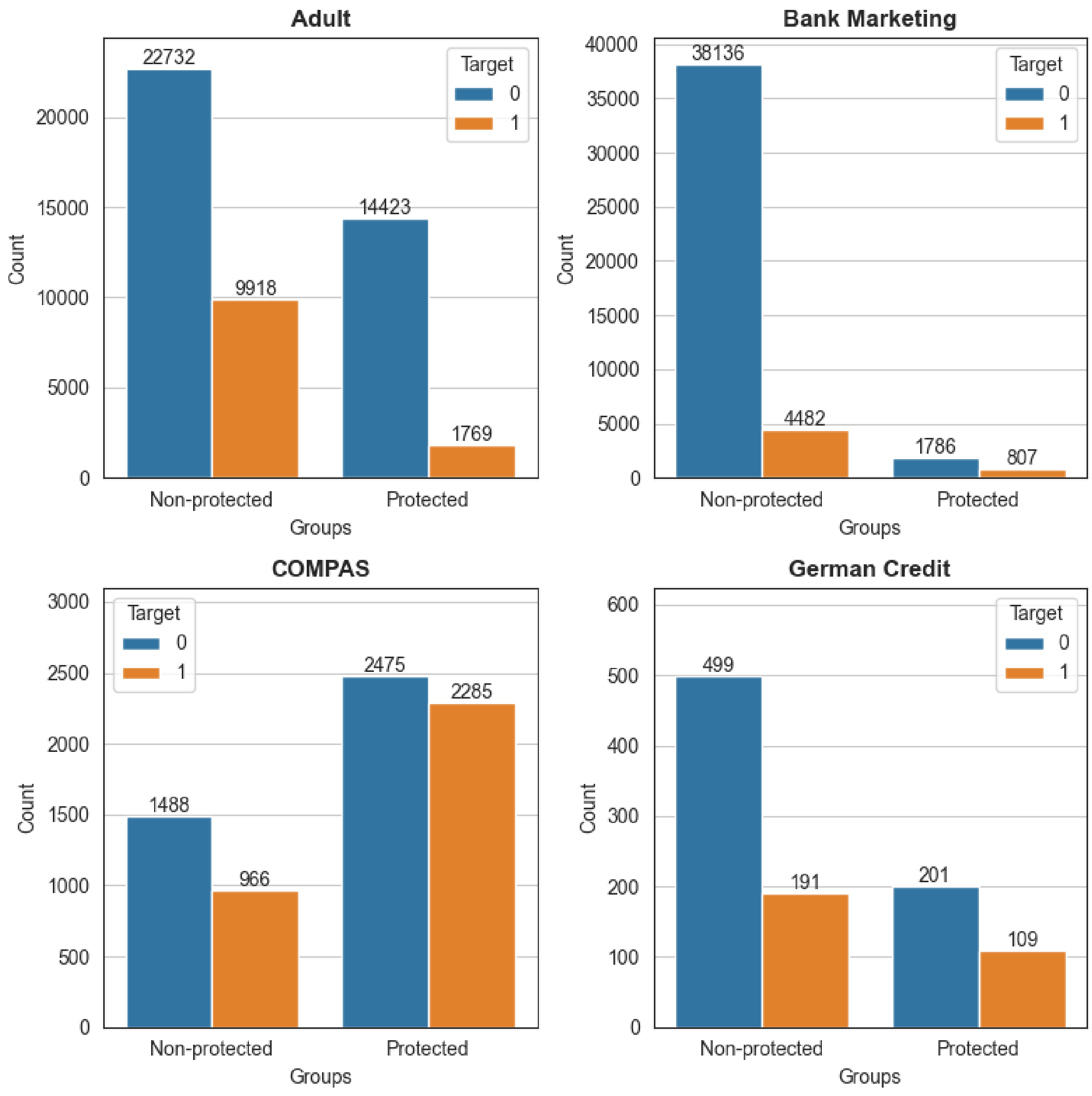


Figure 3.1: Target variable frequencies for protected and non-protected groups.

3.4 Proposal

In this work, six distinct synthetic tabular data generation models will be used for generating synthetic datasets for 4 datasets commonly used in the fairness literature. Then, these synthetic datasets will be evaluated in respect to their fairness by using them for training machine learning classifiers in binary tasks. Lastly, the effectiveness of commonly used fairness mitigation algorithms will be evaluated when they are applied in classifiers trained with synthetic data.

3.5 Experimental Setup

In this section, the experimental setup is fully defined. First, an overview of the methodology is provided. Then, the stages of the study, including hyperparameter tuning of the synthetic data generators, fairness and performance evaluation, and fairness mitigation experiments, are presented in detail.

3.5.1 Overview

An overview of the workflow adopted in this study is illustrated in the Figure 3.2 and is briefly described in this section.

The full workflow used in the experiments in this study, comprising the synthetic data generation models tuning, the synthetic data generation and the downstream fairness evaluating and mitigation experiments was done in a 5 fold K-Fold cross validation procedure stratified by the target variable of each dataset. In each iteration, three folds were used as train set, one as validation and one as test set. The training set is used for training the STDG models, the validation set for the hyperparameter tuning of these models and the test set for evaluating the performance and fairness of classifiers trained on both real and generated data and to evaluate the fairness mitigation experiments.

To generate synthetic data, the models TVAE, CTGAN, CTABGAN+, TabDDPM, ARF, and RealTabFormer were utilized. These models were chosen to represent a diverse set of architectures commonly used in this field, including VAEs, GANs, Diffusion Models, Transformer variants, and more. The hyperparameters for these models were optimized in each iteration of the cross-validation with the Optuna library. The objective function was set to maximize the Matthews Correlation Coefficient (MCC) of a downstream classifier trained on the synthetic data produced by each model and evaluated on the real validation set. The classifier used in all downstream classification experiments in this study, including fairness evaluation and mitigation experiments, was the LightGBM classifier, a well-known and widely adopted gradient boosting model in the industry.

After the tuning procedure, a best hyperparameter configuration for each model in each fold was selected and the corresponding model was used for generating a synthetic dataset with the same size as the original train dataset.

Once generated, these synthetic datasets had their quality evaluated. This evaluation procedure included univariate and multivariate similarity analysis between the synthetic and original data.

Utility and fairness of the synthetic datasets were evaluate by training a classifier on these datasets and evaluating then on real data test set. For evaluating the fairness, Demographic Parity and Equalized Odds were used as measures.

Finally, fairness mitigation experiments were conducted on downstream classifiers trained with the real and synthetic datasets. The fairness mitigation algorithms selected for this task were the Threshold Optimizer and the Reductions method, both available in the Fairlearn Python library (WEERTS *et al.*, 2023).These algorithms are compatible with various classifiers, including the LightGBM classifier used here, and can be applied with multiple fairness measures, such as Demographic Parity and Equalized Odds.

Once the classifiers were trained, they were evaluated using the real data test set to assess both performance and fairness. Following the completion of all experiments, statistics like the mean and standard deviation were calculated across all folds for each experiment.

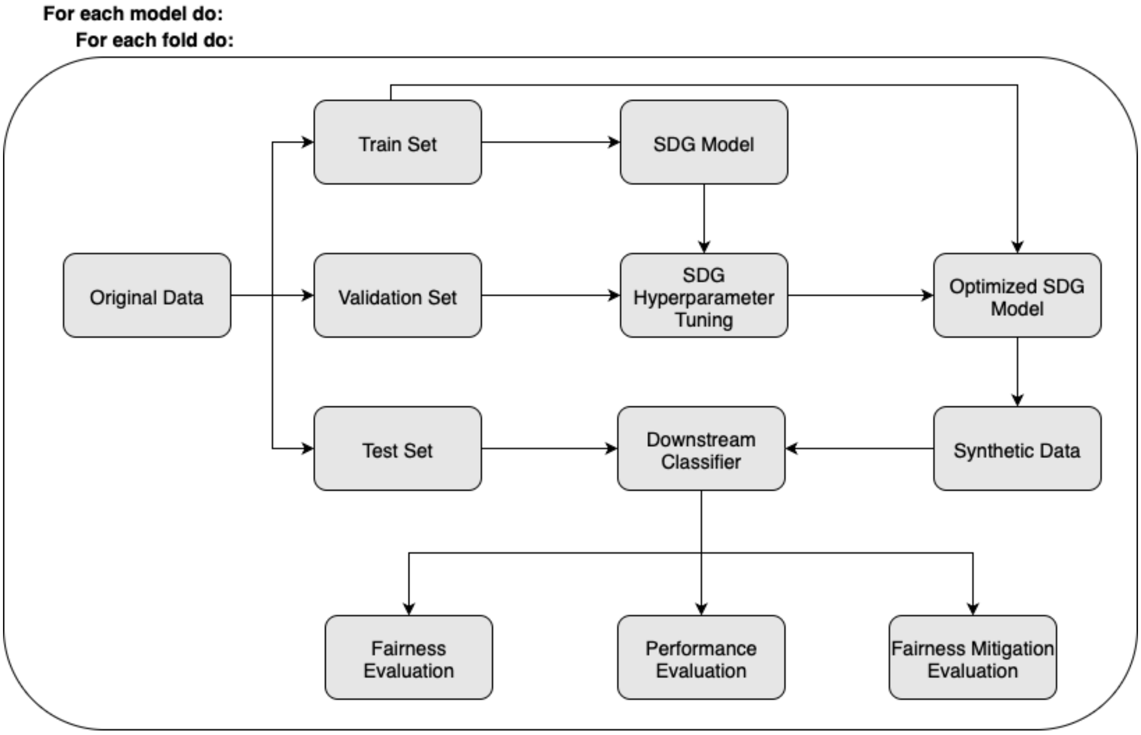


Figure 3.2: Experimental Setup Workflow.

3.5.2 STDG Hyperparameter Tuning

All the STDG models used in this work, except for ARF, are based on neural networks. As is well known, neural network architectures are usually highly sensitive to hyperparameter configurations. Since the primary focus of this work lies in empirical experiments and comparisons, it is crucial to ensure that the hyperparameters of the synthetic data generators are appropriately tuned for each dataset.

To achieve this, a hyperparameter tuning process was conducted for each model within each dataset, spanning an adequate hyperparameter space. For this, the Optuna library (AKIBA *et al.*, 2019) was used. The objective function was set to maximize the MCC of a downstream classifier trained on synthetic data and evaluated on the real validation set. This involved running 20 trials for each combination of models and datasets. The sampler used in the tuning process was the TPESampler. A illustration of this workflow for the hyperparameter tuning stage is shown in Figure 3.3. The exception to this methodology was for the RealTabFormer model, which had expensive computational running times and a limited number of significant hyperparameters to tune in its original implementation. Consequently, only the batch size was tuned with two possible values.

The hyperparameter space for each model was determined based on its specifications. For CTGAN, TVAE, TabDDPM, and ARF, the hyperparameter space was defined following the Synthcity library. In contrast, CTABGAN+ utilized only simple hyperparameters, such as the learning rate and batch size, as its research paper did not provide detailed descriptions of the hyperparameters employed. In most cases, the same hyperparameter space was used for all the datasets, with exception of batch size for TVAE, CTGAN, CTABGAN+ and TABDDPM, that was changed according to the dataset size, and the ARF hyperparameter "min_node_size" in which the minimum value of the range was changed from 3 to 4 for the Bank Marketing dataset as the value of 3 incurred in errors for this dataset.

To train STDG models the specific data preparation process of each STDG model was used. For training classifiers used to evaluate synthetic data utility, categorical features were one-hot-encoded and numerical features were standardized.

This hyperparameter tuning stage was conducted on multiple machines, including personal and cloud-based ones, each equipped with distinct GPUs and specifications. Although each run cannot be directly compared to another due to the varying specifications of the machines used, the total computing time required to perform this hyperparameter tuning for all models across all folds exceeded 40 days of GPU computing time.

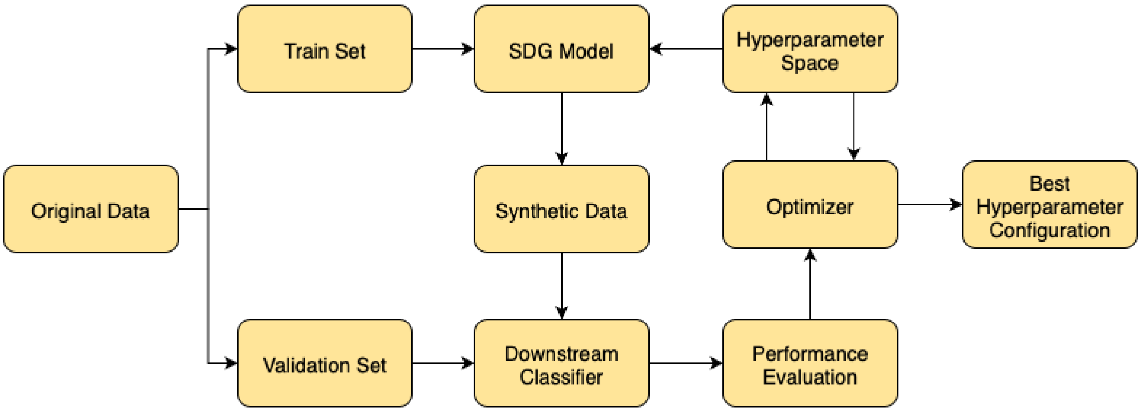


Figure 3.3: Hyperparameter Tuning Workflow.

3.5.3 Synthetic Data Quality Evaluation

Following the generation of synthetic datasets, a comprehensive evaluation framework was implemented to assess the quality of the synthesized data. This evaluation methodology was adapted from the quality assessment procedures established in KOTELNIKOV *et al.* (2023) and encompasses multiple analytical dimension.

The evaluation framework consisted of univariate similarity analysis, which employed statistical measurements and distribution plots to quantify the similarity between real and synthetic feature distributions. For categorical variables, the Jensen-Shannon Divergence (JSD) served as the primary similarity metric:

$$JSD(P,Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3.1}$$

where $M = \frac{1}{2}(P+Q)$ represents the average distribution and D_{KL} denotes the Kullback-Leibler divergence. For continuous numerical variables, the Kolmogorov-Smirnov (KS) test statistic was utilized to measure distributional similarity:

$$D_{KS} = \sup_x |F_n(x) - F_m(x)| \tag{3.2}$$

where $F_n(x)$ and $F_m(x)$ represent the empirical cumulative distribution functions of the original and synthetic datasets, respectively.

Bivariate relationships were also examined through the difference of association matrices between the original and synthetic datasets. These association matrices employed different association measures depending on the variable types: Pearson correlation coefficient for numerical-numerical variable pairs, correlation ratio for numerical-categorical associations, and Cramér's V statistic for categorical-categorical relationships.

The resulting difference matrices were visualized as a heatmap to identify pairs of variables with substantial discrepancies in association. Selected variable pairs exhibiting significant differences were further examined through appropriate visualizations for qualitative assessment.

Additionally, statistical properties related to sensitive and target attributes were computed and compared between the original and synthetic datasets to verify the preservation of critical demographic characteristics.

The Distance to Closest Record (DCR) is one of the most widely adopted privacy measures in the synthetic tabular data literature, especially among utility-focused STDG models, which are not exclusively focused on preserving privacy.

In this work, the Manhattan distance was calculated to determine the closest instance between the synthetic data and the original training data instances. This measure was calculated for the synthetic datasets generated in the first iteration of the cross-validation. Statistics of the DCR, such as the fifth percentile, median, and mean, were used to evaluate and compare the distinct synthetic datasets.

Furthermore, DCR values were also calculated for the validation set in relation to the original training set. This can be used as a baseline to assess whether the STDG models are merely copying the original data or are really learning to represent the underlying data distribution. In the first case, the synthetic DCR distribution would be closer to zero than the DCR of the validation set. In the second case, the DCR of the synthetic data would be similar to the DCR of the validation data. Another possible scenario is that the distribution of DCR values in the synthetic data is larger than that of the validation set. This could indicate that the synthetic samples are deviating from the real data distribution, resulting in poor fidelity and utility of the synthetic data.

3.5.4 Downstream Performance and Fairness Evaluation

Once the hyperparameters for each model in each dataset were optimized, the best configuration was used to train each STDG model on the training set. Subsequently, each model was utilized to generate a synthetic dataset with the same size as the real training set.

These synthetic datasets were then used to train a LightGBM classifier for binary classification tasks on each dataset. To evaluate the classifier, the real data test set was utilized, and performance and fairness metrics were calculated based on the model's results on this real test set. The LightGBM was used with the default configurations in all the datasets.

Accuracy, F1-score, and MCC were calculated as performance metrics, with MCC being the primary focus of analysis. This is because MCC is a measure that can be applied to imbalanced datasets, such as those used in this work, which have varying levels of imbalance. As can be seen in the Equation 3.3, MCC uses true and false positive and negative rates. Demographic parity and equalized odds were calculated as fairness metrics and used for evaluation.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3}$$

3.5.5 Synthetic Data Downstream Fairness Mitigation

To evaluate the effectiveness of fairness mitigation algorithms when applied to classifiers trained on synthetic data, distinct algorithms were introduced into the classifiers' prediction pipeline. These algorithms included the Threshold Optimizer and the Reductions methods. These algorithms were chosen as they can be used with distinct types of classifiers, including LightGBM, and they can be used to mitigate either Demographic Parity or Equalized Odds.

For each fairness metric being evaluated, namely Demographic Parity and Equalized Odds, a LightGBM classifier with these fairness mitigation algorithms applied to it was trained on synthetic datasets with the fairness mitigation algorithm set to mitigate the corresponding fairness metric. The results were then evaluated on the real test set, and fairness and performance metrics were calculated. These metrics were compared with their counterparts prediction pipelines trained with real data. This procedure was applied to each fold of the cross validation and were then aggregated in the end to get final scores.

Chapter 4

Results

This section presents and discusses the main experimental results. First, the outcomes of the hyperparameter tuning for the generation models are presented. Subsequently, the performance and fairness of classifiers trained on datasets generated by the best-performing generator models are compared against classifiers trained on the original data. Finally, the fairness mitigation results of classifiers trained on synthetic data are presented and compared with their counterparts trained on the original dataset.

4.1 Hyperparameter Tuning

Boxplots of the MCC of all the trials of the hyperparameter tuning accumulate along the five folds of the cross-validation of the STDG models are presented in the Figure 4.1. As can be seen, there was some variability in the models' performance across the trials, specifically with CTABGAN+, CTGAN, and TVAE, where, across all datasets, they presented a wide range of MCC depending on the hyperparameter configuration.

In the Adult dataset, RealTabFormer achieved the best performance score, despite the small number of runs, as it was the only model with consistent values over 0.6. TabDDPM, ARF, CTABGAN+, and TVAE were the next with the best maximum scores, with maximum values over 0.6, with a greater variability for CTABGAN+ and TVAE. In CTGAN and CTABGAN+, a greater variability in scores is observed, with values under 0.4 and even cases of total model collapse near 0.

RealTabFormer was also the best performer in the Bank Marketing dataset, with most runs achieving values exceeding 0.5. It's also evident that there is greater variability in the trials for CTGAN, TVAE, and CTABGAN+, with the latter having a significant number of trials that resulted in the model's complete collapse, with values approaching 0.

In the COMPAS dataset, RealTabFormer, which was the best performer in Adult and Bank Marketing, exhibited the worst performance here, with still small variability but with maximum scores over the runs being lower than those of the other models. In this dataset, ARF and TabDDPM achieved the best results, with the best values around 0.35. CTABGGAN+ once again demonstrated a significant number of trials close to 0. Despite CTABGAN+, CTGAN, TVAE, and TABDDPM exhibited the most variability.

In the German dataset, the observed patterns differ from those in the previous dataset. A larger variation in MCC values is observed across all datasets, indicating greater sensitivity to hyperparameter configuration across all models in this dataset.

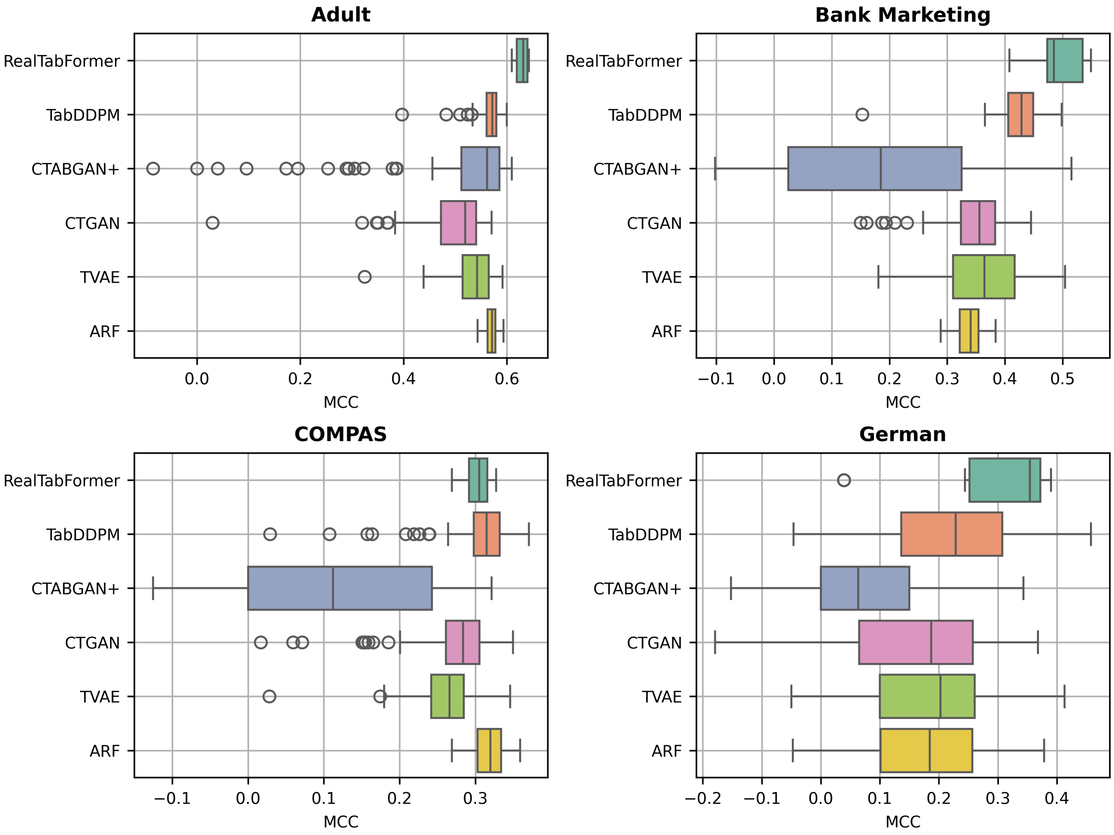


Figure 4.1: Histograms of the MCC of all the trials of the hyperparameter tuning of each model for all datasets.

4.2 Synthetic Data Quality Evaluation

In this section, the results for the quality evaluation of the synthetic datasets are presented and discussed. First, the results for the univariate data analysis are shown, which include plots and statistical measures to evaluate the similarity between the features of the synthetic and real datasets. Then, the joint distributions of the datasets are assessed by comparing their correlation matrix and the difference to the original correlation matrix of the dataset. Also, the multivariate distribution is compared with the MMD metric.

4.2.1 Univariate Quality Analysis

The Tables 4.1 to 4.4 present the results of the univariate analysis of the datasets features similarities. For each dataset, the comparison between the synthetic data and the original data is performed for each feature individually. The JS Divergence is used for categorical variables, while the KS statistic is used for numerical ones.

Observing the tables, it's evident that some models consistently perform better or worse than others. For instance, CTABGAN+ consistently shows poor performance across most features, with particularly high divergence values in the Bank Marketing, COMPAS, and German datasets. On the other hand, TabDDPM, RealTabFormer, and ARF seem to be the models with the most adherence to the original data across the four datasets. TabDDPM specifically excels in numerical features while also maintaining good results for categorical ones. ARF also shows generally excellent scores, mainly with categorical features, except for high-cardinality ones such as native-country (JS=0.260) and occupation (JS=0.227) in the Adult dataset.

RealTabFormer demonstrates overall strong performance without significant deviations in any features across the datasets. An intriguing aspect of RealTabFormer is that, while it excels at high-cardinality categorical variables like those in the Adult dataset, it yields the worst results for simpler categorical features such as income, relationship, and sex within the same dataset. TVAE exhibits mixed performance across datasets, with slightly better results in numerical features than in categorical ones, particularly struggling to capture complex categorical features like native-country in the Adult dataset (JS=0.437). CTGAN also faces challenges with high-cardinality categorical variables, which is particularly evident in the native-country feature (JS=0.543) and capital-gain distributions (KS=0.466).

Table 4.1: Univariate quality analysis with JS Divergence for categorical and KS Statistic for numerical features for the Adult dataset

Feature	**TVAE**	**CTGAN**	**CTABGAN+**	**TabDDPM**	**RealTabFormer**	**ARF**
			Categorical Features (JS Divergence)			
education	0.272	0.357	0.074	0.065	0.050	0.185
income	0.007	0.010	0.024	0.008	0.031	0.003
marital-status	0.051	0.019	0.041	0.028	0.038	0.009
native-country	0.437	0.543	0.073	0.058	0.046	0.260
occupation	0.295	0.213	0.061	0.058	0.055	0.227
race	0.044	0.016	0.036	0.031	0.035	0.005
relationship	0.038	0.015	0.038	0.016	0.039	0.007
sex	0.020	0.004	0.008	0.004	0.020	0.004
workclass	0.048	0.023	0.052	0.027	0.031	0.009
			Numerical Features (KS Statistic)			
age	0.053	0.036	0.113	0.008	0.029	0.014
capital-gain	0.320	0.466	0.033	0.008	0.016	0.187
capital-loss	0.023	0.025	0.025	0.006	0.010	0.140
education-num	0.157	0.153	0.264	0.016	0.039	0.067
fnlwgt	0.049	0.044	0.042	0.009	0.032	0.039
hours-per-week	0.220	0.221	0.046	0.023	0.038	0.176

Note: Blue highlighting indicates best performance (lowest value). Red highlighting indicates worst performance (highest value).

Table 4.2: Univariate quality analysis with JS Divergence for categorical and KS Statistic for numerical features for the Bank Marketing dataset

Feature	TVAE	CTGAN	CTABGAN+	TabDDPM	RealTabFormer	ARF
Categorical Features (JS Divergence)						
contact	0.040	0.010	0.261	0.017	0.021	0.010
default	0.058	0.010	0.039	0.019	0.027	0.003
education	0.031	0.019	0.043	0.019	0.024	0.004
housing	0.004	0.012	0.063	0.003	0.017	0.005
job	0.300	0.279	0.085	0.100	0.034	0.231
loan	0.013	0.015	0.030	0.009	0.020	0.003
marital	0.045	0.019	0.050	0.012	0.030	0.004
month	0.282	0.300	0.240	0.077	0.037	0.227
poutcome	0.029	0.013	0.536	0.008	0.035	0.019
y	0.020	0.014	0.197	0.015	0.027	0.011
Numerical Features (KS Statistic)						
age	0.015	0.010	0.036	0.014	0.011	0.006
balance	0.138	0.099	0.110	0.009	0.030	0.109
campaign	0.160	0.151	0.133	0.020	0.037	0.039
day	0.080	0.044	0.120	0.013	0.043	0.046
duration	0.094	0.106	0.105	0.009	0.039	0.079
pdays	0.451	0.362	0.710	0.011	0.025	0.104
previous	0.026	0.033	0.721	0.009	0.026	0.019

Note: Blue highlighting indicates best performance (lowest value). Red highlighting indicates worst performance (highest value).

Table 4.3: Univariate quality analysis with JS Divergence for categorical and KS Statistic for numerical features for the COMPAS dataset

Feature	TVAE	CTGAN	CTABGAN+	TabDDPM	RealTabFormer	ARF
Categorical Features (JS Divergence)						
age_cat	0.045	0.023	0.061	0.045	0.036	0.039
c_charge_degree	0.023	0.008	0.038	0.009	0.032	0.009
race	0.075	0.036	0.102	0.034	0.046	0.024
sex	0.041	0.032	0.026	0.018	0.026	0.018
Numerical Features (KS Statistic)						
age	0.118	0.066	0.111	0.056	0.053	0.064
juv_fel_count	0.018	0.020	0.099	0.006	0.014	0.011
juv_misd_count	0.015	0.028	0.139	0.011	0.017	0.019
juv_other_count	0.014	0.039	0.179	0.015	0.013	0.017
priors_count	0.089	0.065	0.186	0.078	0.053	0.079
two_year_recid	0.028	0.034	0.099	0.032	0.040	0.027

Note: Blue highlighting indicates best performance (lowest value). Red highlighting indicates worst performance (highest value).

Table 4.4: Univariate quality analysis with JS Divergence for categorical and KS Statistic for numerical features for the German dataset

Feature	TVAE	CTGAN	CTABGAN+	TabDDPM	RealTabFormer	ARF
Categorical Features (JS Divergence)						
checking-account	0.050	0.043	0.054	0.054	0.040	0.028
credit-history	0.084	0.042	0.127	0.057	0.058	0.029
employment-since	0.067	0.048	0.095	0.076	0.068	0.025
foreign-worker	0.030	0.016	0.028	0.029	0.033	0.012
housing	0.037	0.027	0.064	0.060	0.045	0.023
job	0.061	0.029	0.087	0.072	0.047	0.026
other-debtors	0.065	0.027	0.047	0.061	0.032	0.018
other-installment	0.056	0.028	0.052	0.044	0.044	0.014
personal-status	0.061	0.050	0.059	0.062	0.040	0.027
property	0.052	0.036	0.071	0.076	0.058	0.019
purpose	0.109	0.075	0.107	0.096	0.080	0.050
savings-account	0.076	0.054	0.074	0.062	0.052	0.032
sex	0.031	0.024	0.046	0.045	0.028	0.013
telephone	0.009	0.015	0.035	0.032	0.046	0.020
Numerical Features (KS Statistic)						
age	0.118	0.083	0.141	0.113	0.072	0.063
class-label	0.032	0.052	0.076	0.039	0.039	0.010
credit-amount	0.141	0.125	0.095	0.121	0.064	0.113
duration	0.170	0.132	0.137	0.134	0.065	0.134
existing-credits	0.065	0.026	0.184	0.060	0.061	0.018
installment-rate	0.049	0.052	0.190	0.061	0.074	0.016
number-people-provid...	0.051	0.027	0.189	0.050	0.018	0.020
residence-since	0.060	0.042	0.142	0.039	0.062	0.014

Note: Blue highlighting indicates best performance (lowest value). Red highlighting indicates worst performance (highest value).

The results of the analysis with statistical measures can be easily verified and illustrated visually through with graphical representations of the distribution of the features of synthetic and original datasets.

For instance, the attribute ’native-country’ from the Adult dataset, which showcased the worst results for the TVAE and CTGAN models and best results for the RealTabFormer and TabDDPM, can be visualized in Figure 4.2. This attribute, which indicates the country of origin of a person, has a total of 42 possible values in the original dataset. However, their occurrence is highly concentrated in the value ‘United States,’ with approximately 90% of the instances. Consequently, the remaining 41 possible values represent the other 10% of the instances. This imbalance poses a challenge for STDG models, as evident in the figure, where TVAE and CTGAN failed to represent many of the possible values.

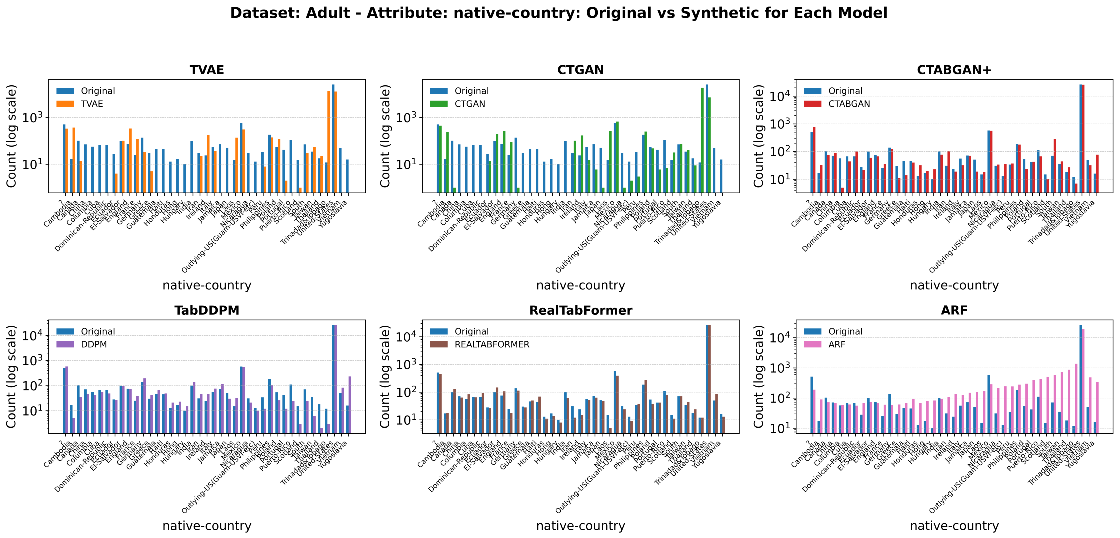


Figure 4.2: Barplots with frequency of values in the variable “native-country” across the synthetic and original versions of the Adult dataset.

A similar pattern can be noted for the ’occupation’ attribute of the same dataset, which indicates the person’s job, represented in the Figure 4.3. This attribute has 15 possible values in the original dataset, with six of the most frequent values accounting for approximately 70% of the instances. Although it is less imbalanced compared to the ‘native-country’ attribute, it exhibits a similar pattern. TVAE and CTGAN failed to accurately represent the original distribution, while TabDDPM, RealTabFormer, and CTABGAN+ demonstrated a much closer alignment with the original dataset.

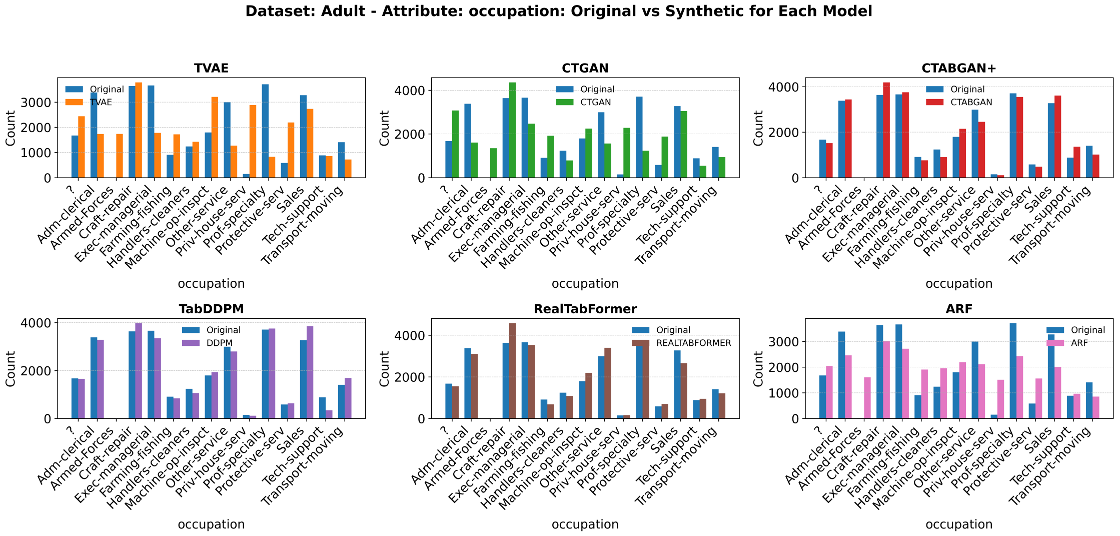


Figure 4.3: Barplots with frequency of values in the variable “occupation” across the synthetic and original versions of the Adult dataset.

When the categorical attributes have fewer possible values and a higher frequency for each value, even those models that performed poorly in the previous cases exhibit a much better performance. This is the case for the variable ‘relationship’ in the Adult dataset, as illustrated in Figure 4.4.

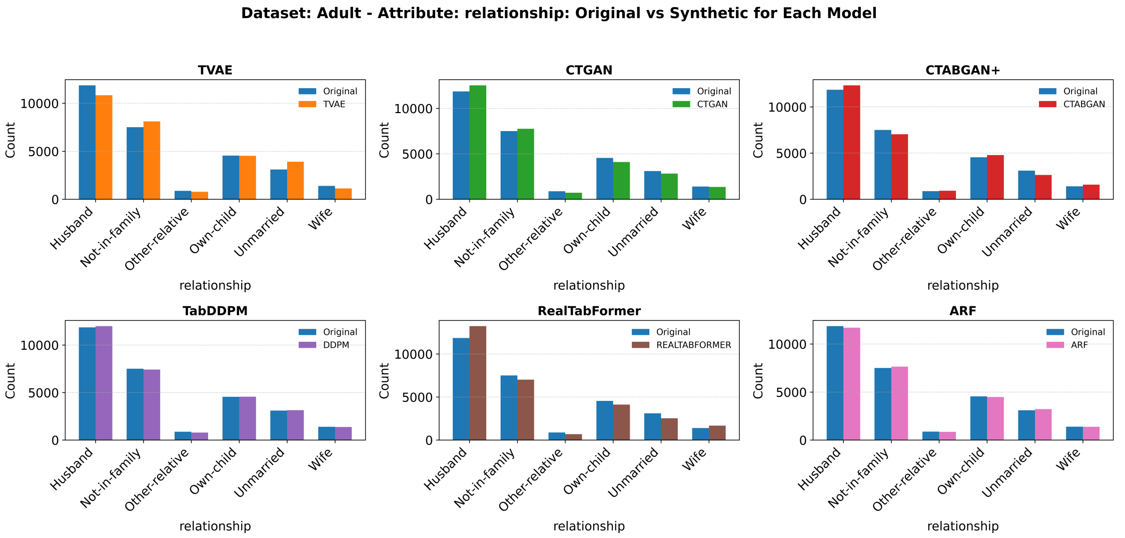


Figure 4.4: Barplots with frequency of values in the variable “relationship” across the synthetic and original versions of the Adult dataset.

A distinct pattern can be noted for numerical attributes. Even in cases where the distribution of a variable is relatively simple, such as the case for the attribute ’age’ in the Adult dataset, illustrated in Figure 4.5, some models, specially TVAE and CTGAN, but also CTABGAN+ and RealTabFormer, failed to represent adequately the distribution of the data.

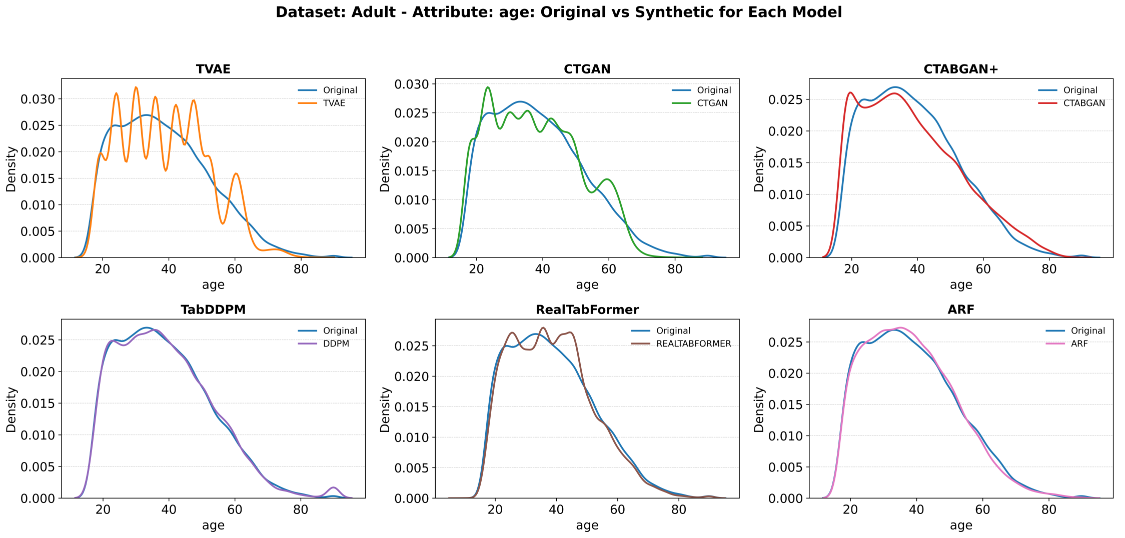


Figure 4.5: Kernel density plots showcasing the distribution of values in the variable “age” across the synthetic and original versions of the Adult dataset.

When the distribution is highly skewed as in the case of the attribute 'capital-gain', represented in Figure 4.6, or multimodal as in the case of the attribute 'education-num', represented in Figure 4.7, these challenges are pronounced. In the case of 'capital-gain', even the model TabDDPM, which had the best performance in numerical features, did not perform well in this case. In 'education-num', observing the Figure 4.7, it is very clear that some models such as TVAE and CTGAN had difficulties related to the multimodal nature of this attribute.

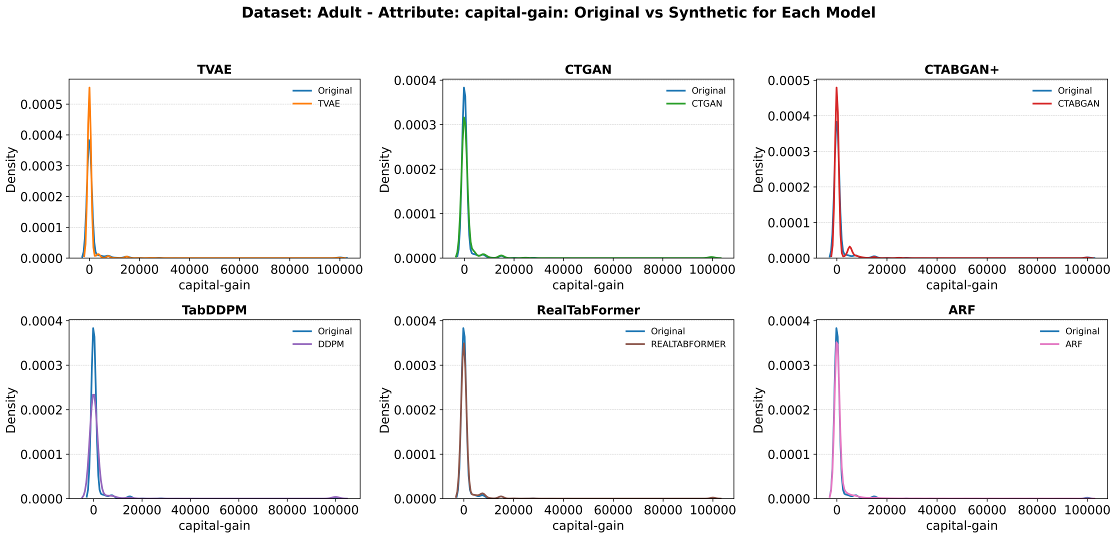


Figure 4.6: Kernel density plots showcasing the distribution of values in the variable "capital-gain" across the synthetic and original versions of the Adult dataset.

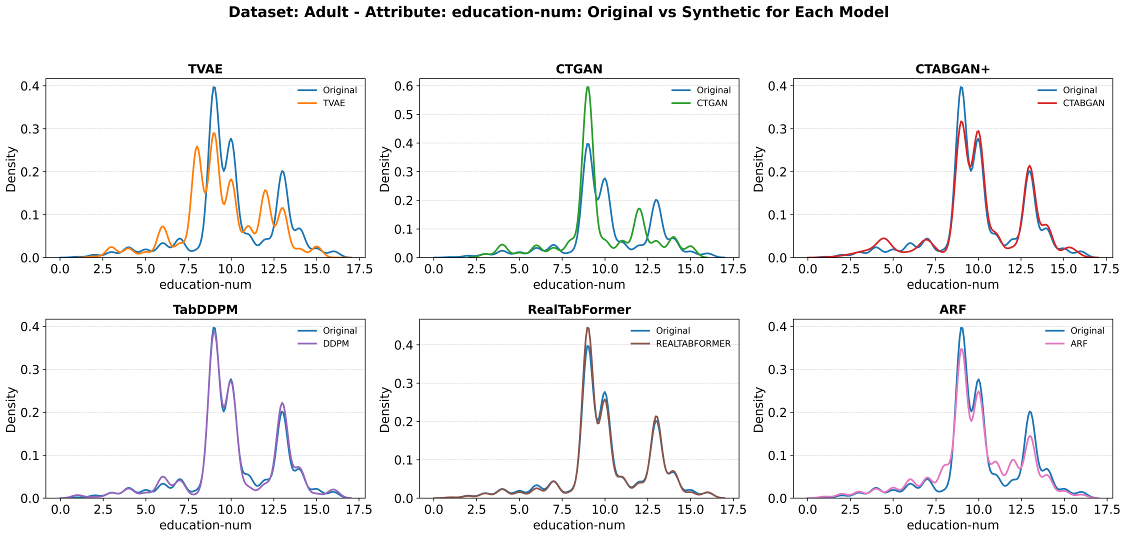


Figure 4.7: Kernel density plots showcasing the distribution of values in the variable "education-num" across the synthetic and original versions of the Adult dataset.

4.2.2 Bivariate Quality Analysis

Beyond univariate distribution, it is important to assess if the synthetic data preserves the original relationship between the variables. For this, heatmaps showing the difference between the association matrices of the original and synthetic datasets are shown in Figures 4.8 to 4.11.

From Figure 4.8, it's evident that RealTabFormer and CTABGAN+ appear to be the most effective models in preserving the relationship between variables in the Adult dataset. In contrast, there are significant divergences in these relationships in TVAE, CTGAN, TabDDPM, and ARF. While TVAE and CTGAN exhibited issues in the univariate distribution, TabDDPM and ARF exhibited remarkable similarity in the univariate analysis. This highlights the importance of evaluating joint distributions for synthetic data.

Additionally, it's clear that different models had varying joint relationships. For instance, TabDDPM appears to have more differences in the relationships with the attributes 'native-country' and 'capital-gain'. TVAE, on the other hand, exhibited differences more distributed across the attributes. CTGAN also particularly struggled with 'native-country', which can be attributed to the difficulties in generating this attribute, as discussed in the previous section. ARF, in contrast, exhibited more differences across the attributes 'occupation' and 'education'.

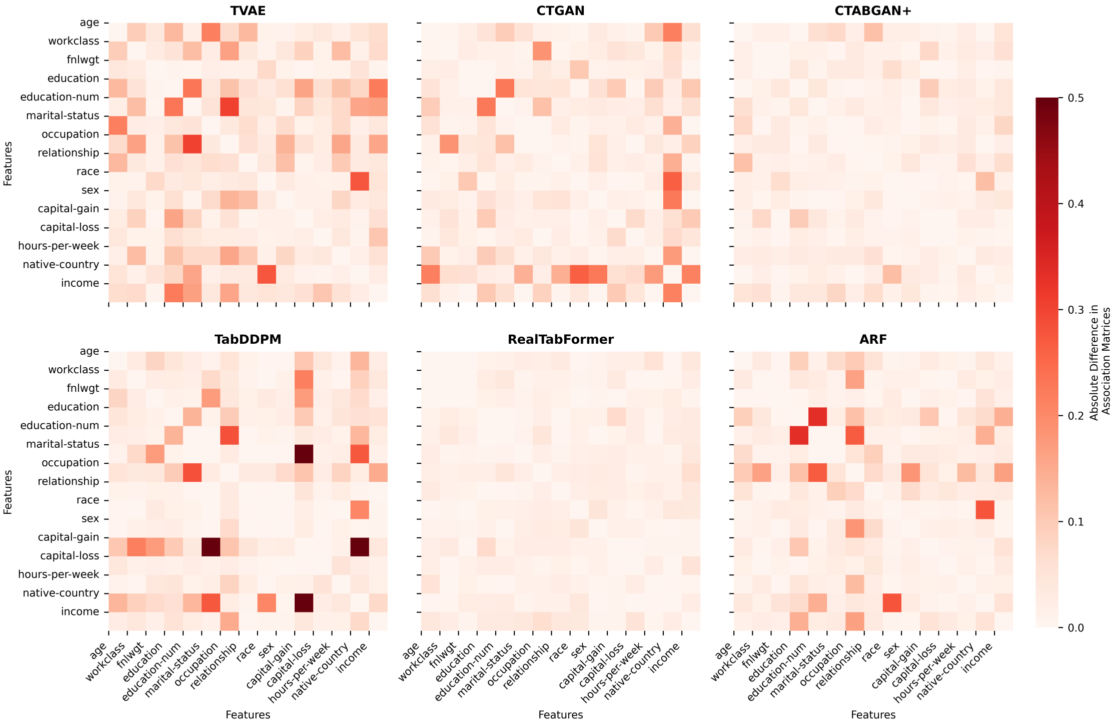


Figure 4.8: Heatmap of the difference between association matrices for the Adult dataset.

In the Bank Marketing dataset, RealTabFormer and TabDDPM appear to preserve the original relationships between variables, as evident from Figure 4.9. In contrast, CTABGAN+ exhibits the opposite behavior and clearly demonstrates the model with the most discrepancies between the original and synthetic relationships. Moreover, it appears that the models encountered challenges in representing the relationship between 'age' and 'job', as well as between the variables 'previous' and 'poutcome', and between 'previous' and 'pdays'.

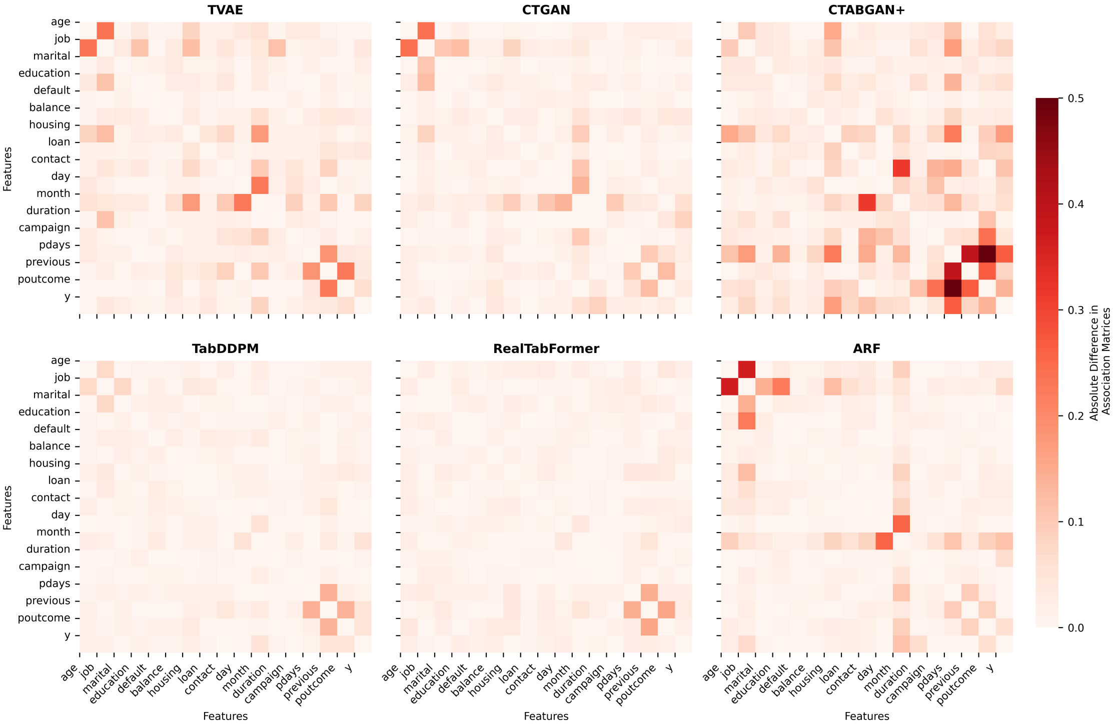


Figure 4.9: Heatmap of the difference between association matrices for the Bank Marketing dataset.

In the COMPAS dataset, most models were able to generate relationships that were quite similar to those observed in the original dataset, as illustrated in Figure 4.10. The exception was CTABGAN+, which exhibited the largest discrepancies in the relationships. ARF appeared to have better captured the joint relationships between the variables, followed by TabDDPM, RealTabForer, and CTGAN.

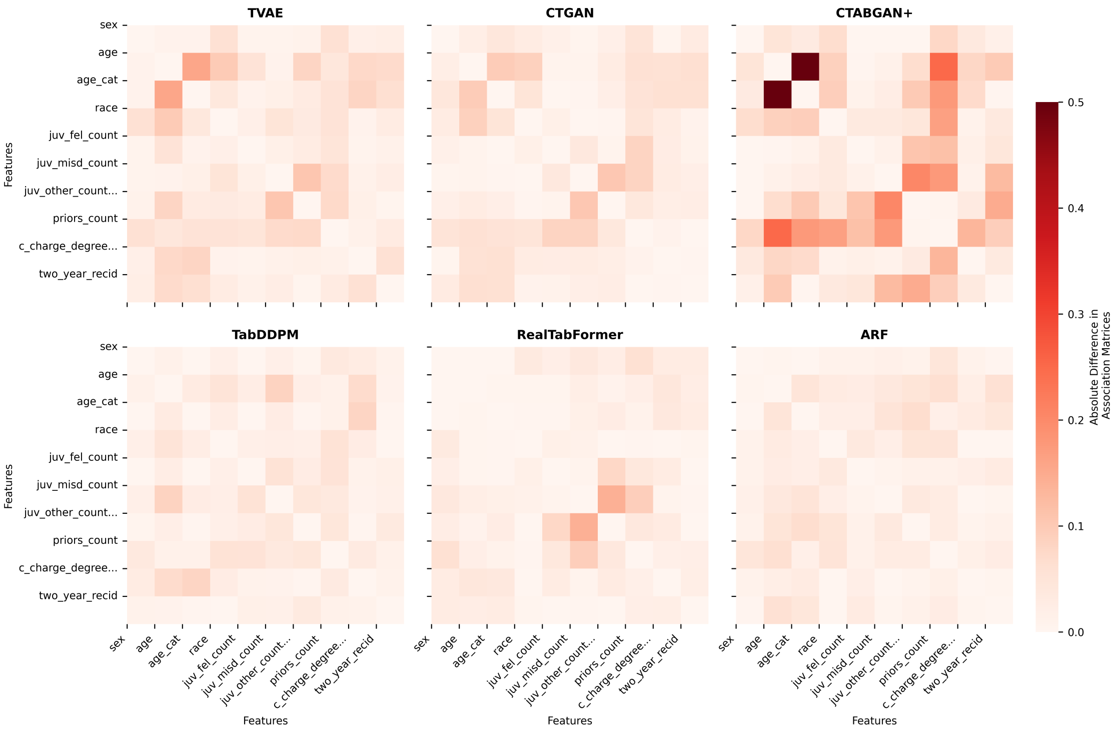


Figure 4.10: Heatmap of the difference between association matrices for the COMPAS dataset.

Lastly, the German dataset exhibited the most significant deviations from the original data relationships across all models, as illustrated in Figure 4.11. While all models exhibited some level of discrepancies across the attributes, CTABGAN+ exhibited the most discrepant results, while RealTabFormer exhibited the least.

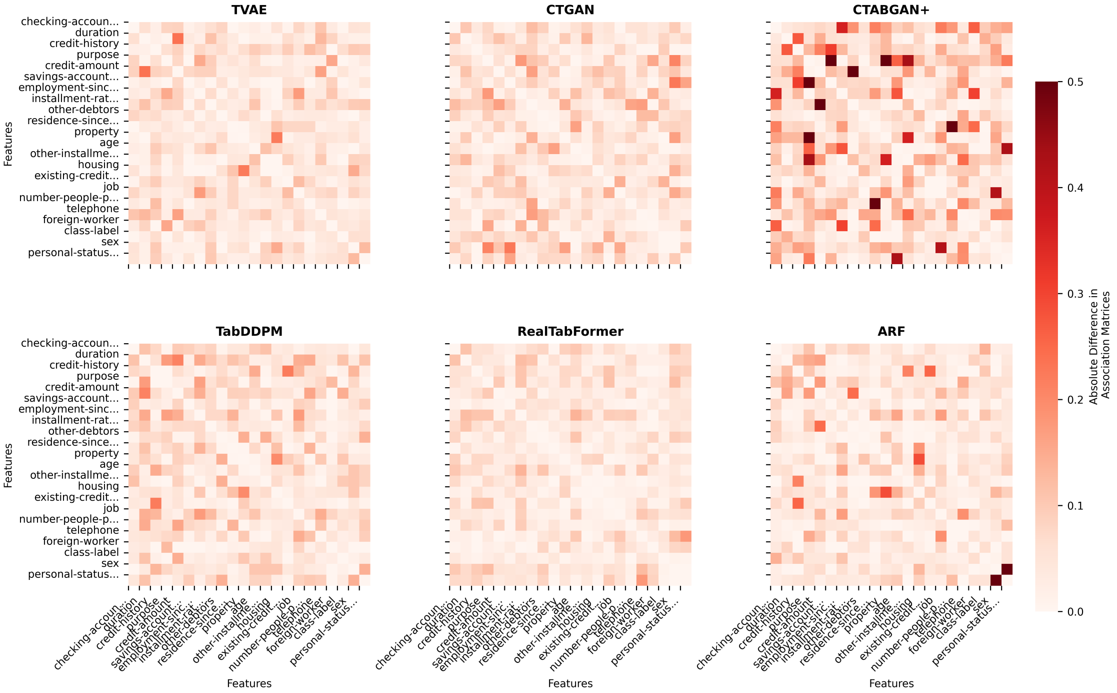


Figure 4.11: Heatmap of the difference between association matrices for the German dataset.

A good example of a numerical to categorical relationship that many models seem to have difficulty representing in the Adult dataset is between the attributes ‘education’ and ‘education-num’. This relationship is quite straightforward, as illustrated in Figure 4.12, which shows the distribution of ‘education-num’ per ‘education’ classes. There’s a direct mapping between the two variables, with each ‘education’ class corresponding to a specific ‘education-num’ numerical value.

Despite the simplicity of this relationship, only RealTabFormer was able to capture it adequately, as shown in Figure 4.13. Except for two outlier values that are misplaced, the model accurately represented the one-to-one mapping. The other models displayed distributed values of ‘education-num’ for each ‘education’ value. CTABGAN+ also showed dispersed values, but it seemed to capture the mean of this distribution closer to the expected value.

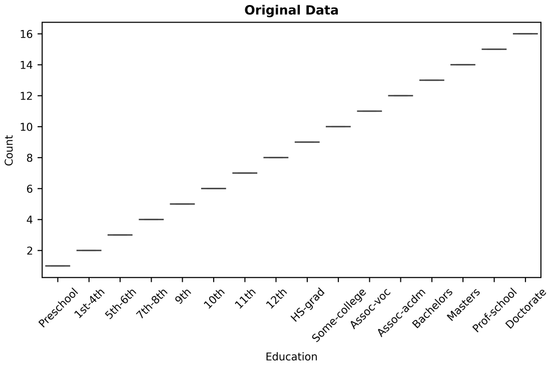


Figure 4.12: Boxplot of the variable 'education-num' per 'education' for the original Adult dataset.

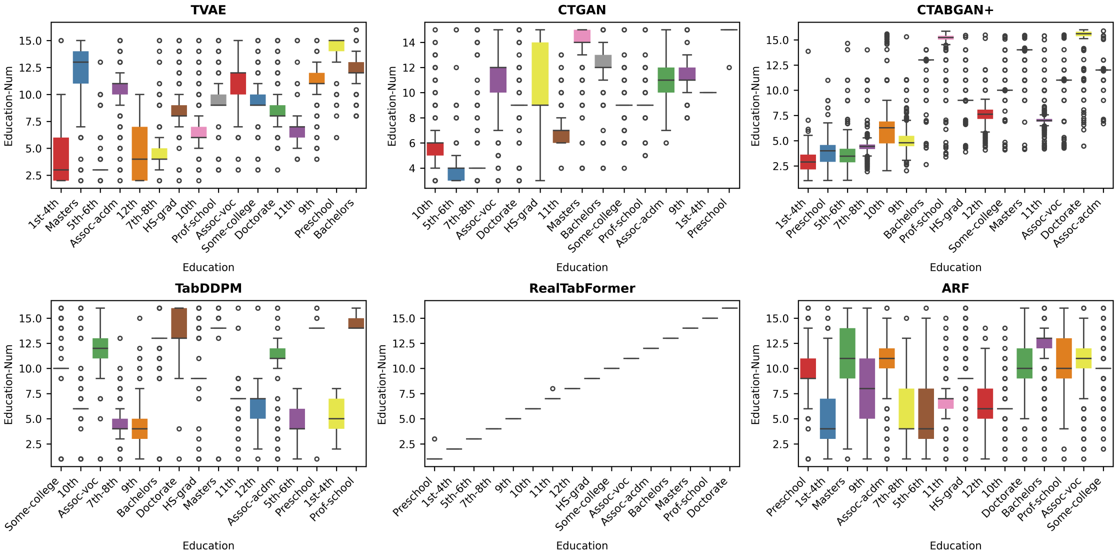


Figure 4.13: Boxplots of the variable 'education-num' per 'education' for the synthetic Adult datasets.

A clear example of discrepancies in a categorical to categorical attributes relationship can be observed when comparing the synthetic relationships depicted in Figure 4.15 with the original in Figure 4.14. In this instance, while most models were able to accurately represent the 'job' attribute for the class '1' of the 'age' attribute, they failed to do so in the same manner for the class '0'. In the original distribution, the two most frequent values in the 'job' attribute are 'retired' and 'student', which makes sense since the class '0' in 'age' represents individuals with ages less than 25 and greater than 60. RealTabFormer and CTABGAN+ were the models that most closely aligned with this pattern, while most other models exhibited distinct distributions of job values for the class '0'. This serves as a clear illustration of a seemingly straightforward semantic relationship between categorical attributes that models failed to represent accurately.

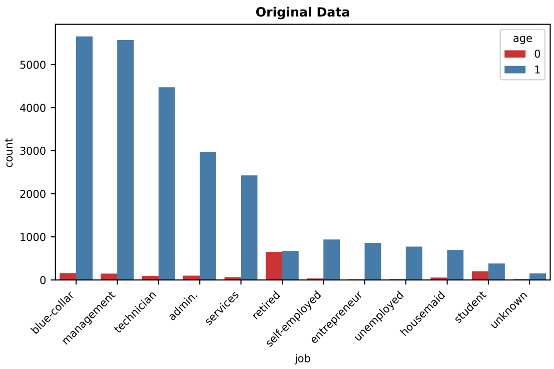


Figure 4.14: Count plot of the variable 'job' per 'age' for the original Bank Marketing dataset.

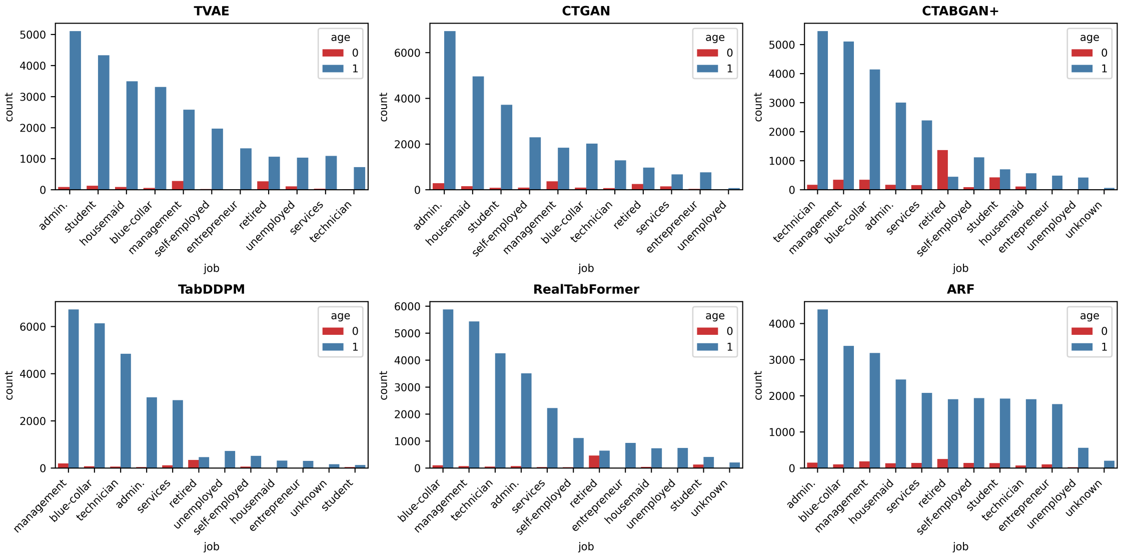


Figure 4.15: Count plots of the variable 'job' per 'age' for the synthetic Bank Marketing datasets.

Discrepancies between numerical attributes relationships were also identified. One example is for the attributes 'credit-amount' and 'duration' of the German dataset, illustrated in Figure 4.16 for the original dataset and in Figure 4.17 for the synthetic datasets. By comparing these figures, it's clear that the models that were able to more closely capture the relationship between the variables were TabDDPM and RealTabFormer.

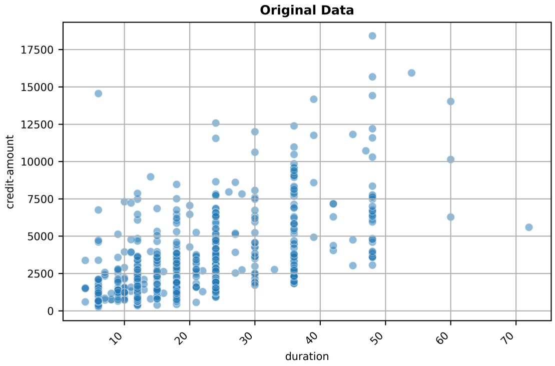


Figure 4.16: Scatterplot of the variable 'duration' per 'credit_amount' for the original German dataset.

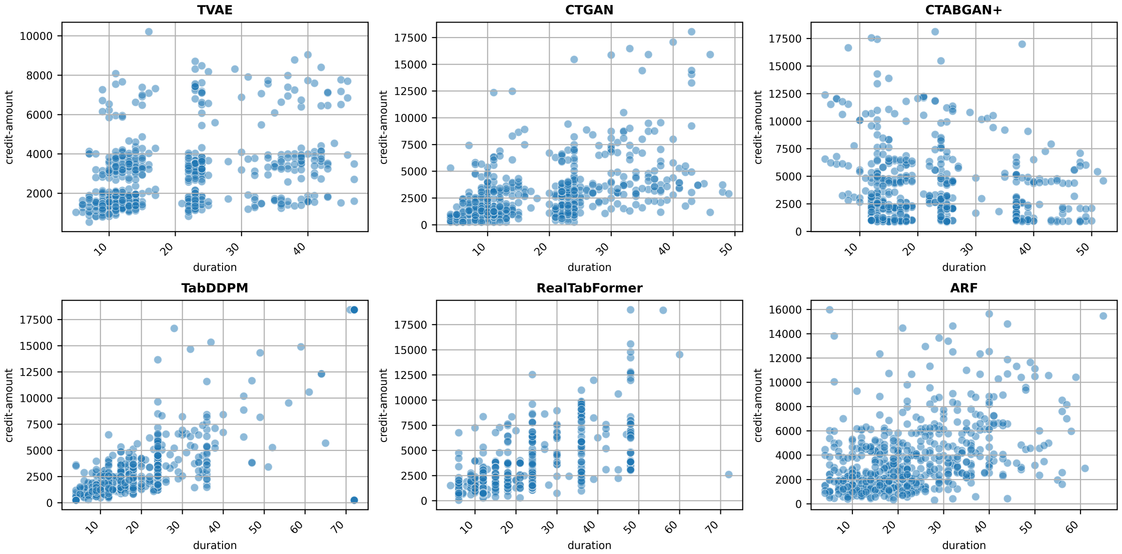


Figure 4.17: Scatterplots of the variable 'duration' per 'credit_amount' for the synthetic German datasets.

4.2.3 Sensitive and Target Attributes Analysis

Table 4.5 presents statistics related to the sensitive and target variables in the synthetic and original datasets for the first fold of the cross-validation.

The results show that in most cases, the synthetic datasets statistics match closely the original ones. Some variation exists, CTABGAN+, for instance, overrepresents protected groups in Bank Marketing, almost doubling from 5.75% to 11.89%. In the same dataset, TVAE, TabDDPM and RealTabFormer underrepresented this same group, with values ranging from 3.51% for TabDDPM to 4.09% for TVAE.

In the Adult dataset, RealTabFormer and TVAE slightly underrepresented the protected group. Additionally, CTABGAN+ underestimated the rates of positives in the target variable, while RealTabformer overestimated them in both the privileged and protected groups. In the CTABGAN+ case, this underestimation occurred due to a lower rate for the privileged group, while in the RealTabFormer case, it resulted in an overestimation of the rate in both groups.

CTABGAN+ and RealTabFormer were also the models with the most divergences in relation to the target positive rate in the COMPAS dataset, although with an inverse behavior. While CTABGAN+ overestimated this rate with a rate of 61.62% versus a reference of 45.08% in the original dataset, RealTabFormer underestimated with a rate of 34.57%. In the German dataset, most models followed closely the original dataset rates, except CTABGAN+, which showed large variations in the target positive rate.

Dataset	**Metric**	**Original**	**TVAE**	**CTGAN**	**CTABGAN+**	**TabDDPM**	**RealTabFormer**	**ARF**
Adult	Protected Group Rate	33.22%	30.56%	33.39%	33.12%	32.82%	29.55%	33.60%
	Target Positive Rate	23.93%	24.80%	24.57%	19.89%	24.17%	30.06%	23.08%
	Target Positive Rate (Privilege)	30.42%	29.76%	31.20%	24.75%	30.82%	36.69%	29.40%
	Target Positive Rate (Protected)	10.87%	13.52%	11.35%	10.09%	10.56%	14.25%	10.59%
Bank Marketing	Protected Group Rate	5.75%	4.09%	5.90%	11.89%	3.51%	3.82%	5.32%
	Target Positive Rate	11.70%	10.41%	12.26%	28.72%	9.96%	9.55%	10.65%
	Target Positive Rate (Privilege)	10.48%	9.46%	10.90%	25.70%	8.98%	8.77%	9.68%
	Target Positive Rate (Protected)	31.71%	32.82%	34.06%	51.12%	36.94%	29.32%	28.09%
Compas	Protected Group Rate	66.98%	69.76%	69.69%	74.31%	67.10%	67.33%	66.87%
	Target Positive Rate	45.08%	49.63%	50.09%	61.62%	48.15%	34.57%	48.66%
	Target Positive Rate (Privilege)	38.70%	48.13%	45.50%	62.86%	43.82%	26.94%	43.38%
	Target Positive Rate (Protected)	48.22%	50.28%	52.09%	61.19%	50.28%	38.26%	51.28%
German	Protected Group Rate	31.00%	30.33%	34.67%	29.00%	38.33%	26.67%	30.67%
	Target Positive Rate	30.00%	32.67%	28.33%	15.83%	26.33%	34.00%	27.83%
	Target Positive Rate (Privilege)	28.50%	29.19%	26.79%	10.80%	25.14%	32.27%	26.68%
	Target Positive Rate (Protected)	33.33%	40.66%	31.25%	28.16%	28.26%	38.75%	30.43%

Table 4.5: Statistics on the sensitive and target attributes for the original and synthetic datasets.

4.2.4 Distance to Closest Record

The DCR was calculated for all synthetic generated datasets. First, the fifth percentile, median, and mean of DCR values for each synthetic dataset are shown in Table 4.6. Then, the distribution of DCR values for a sample of synthetic data with the same size as the validation set are shown together with the DCR of the validation set in Figures 4.18 to 4.21. This comparison with the validation data DCR provides a way to assess whether the synthetic data is merely copying the training data or if it is learning the underlying data distribution.

Looking at Table 4.6, the fifth percentile of DCR values for synthetic datasets is generally under 1. This indicates that many synthetic instances are equal to original values except for one variable. In the COMPAS dataset, the fifth percentile is equal to zero for all the models except CTABGAN+; this likely occurs because this dataset has many duplicated instances and just a small number of features, which makes it easier for the synthetic datasets to reproduce exact copies of the training data.

While generating synthetic datasets with privacy guarantees is not the focus of this work, which focuses on evaluating and mitigating the fairness of SOTA utility-focused STDG models, this DCR evaluation serves as an important consideration regarding the fact that data generated by machine learning models may still present privacy concerns.

Table 4.6: DCR Statistics by Dataset and STDG Model for the first fold of the cross-validation.

Dataset	Model	P05	Median	Mean
Adult	TVAE	1.580	4.699	4.714
	CTGAN	1.886	4.977	5.059
	CTABGAN+	0.333	2.369	2.619
	TabDDPM	0.256	2.175	2.517
	RealTabFormer	0.206	1.547	1.963
	ARF	0.573	3.707	3.965
Bank Marketing	TVAE	0.798	3.196	3.426
	CTGAN	0.838	3.284	3.536
	CTABGAN+	1.304	4.257	4.446
	TabDDPM	0.280	1.350	1.807
	RealTabFormer	0.258	1.467	1.938
	ARF	0.562	2.909	3.228
Compas	TVAE	0.000	0.279	0.541
	CTGAN	0.000	0.247	0.763
	CTABGAN+	0.082	2.164	2.552
	TabDDPM	0.000	0.197	0.622
	RealTabFormer	0.000	0.082	0.327
	ARF	0.000	0.197	0.560
German	TVAE	4.945	9.830	9.598
	CTGAN	5.514	9.676	9.609
	CTABGAN+	7.112	11.173	11.478
	TabDDPM	4.865	9.095	9.231
	RealTabFormer	4.610	8.621	8.484
	ARF	6.046	10.012	10.141

On the other hand, comparing the distribution of synthetic samples with the validation data shows that the models are not simply copying the original data. Otherwise, the synthetic data DCR values would show a distribution much closer to zero than the validation set, which is not the case. As can be seen, in most cases the synthetic data DCR distribution aligns with the validation DCR distribution, with larger deviations for some models, as in the case of TVAE and CTGAN in the Adult dataset.

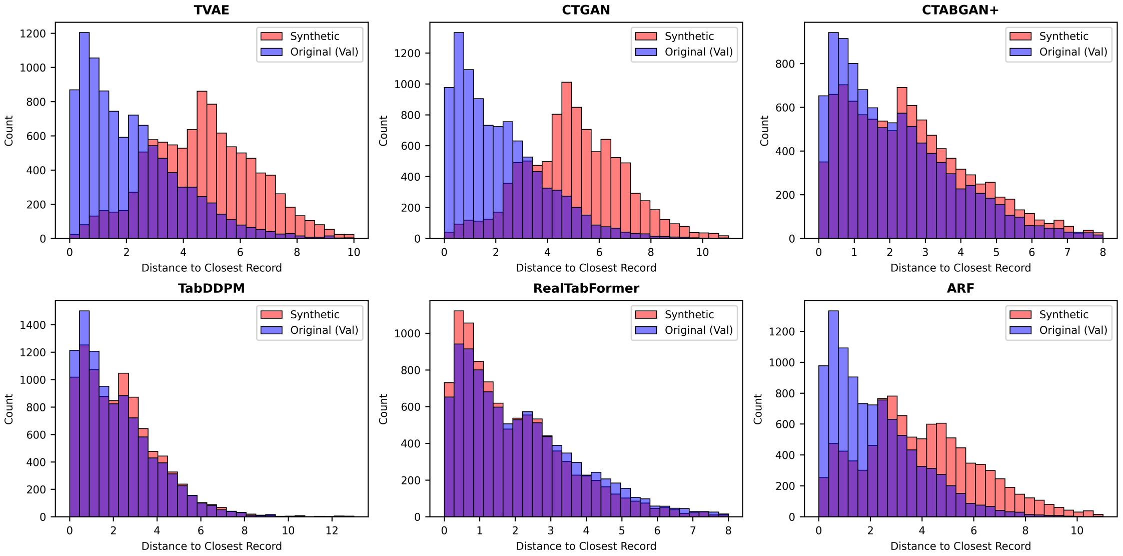


Figure 4.18: Comparison between DCR distribution for synthetic and validation datasets for the Adult dataset in the first fold of the cross-validation.

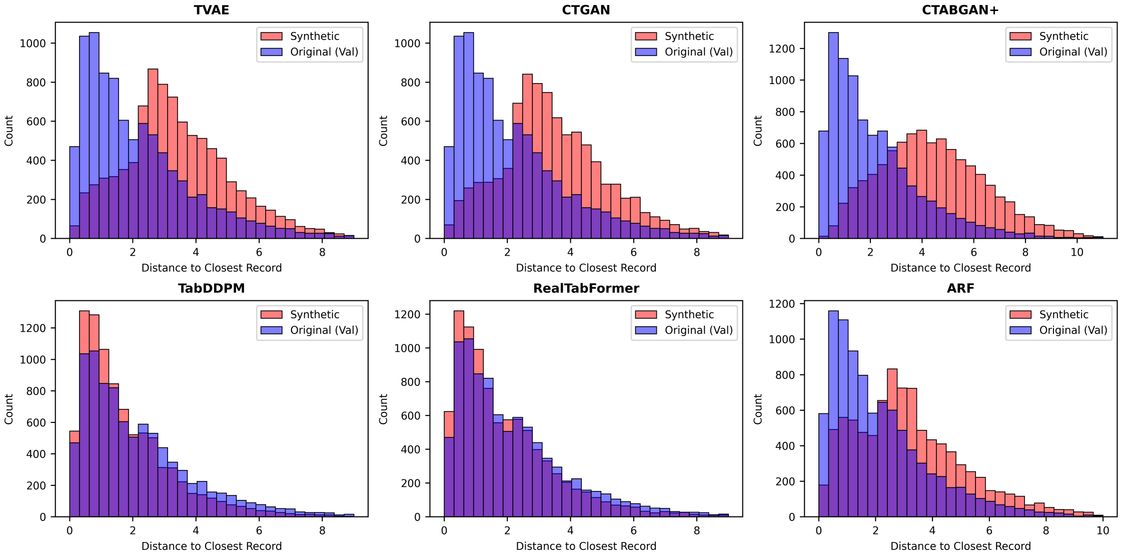


Figure 4.19: Comparison between DCR distribution for synthetic and validation datasets for the Bank Marketing dataset in the first fold of the cross-validation.

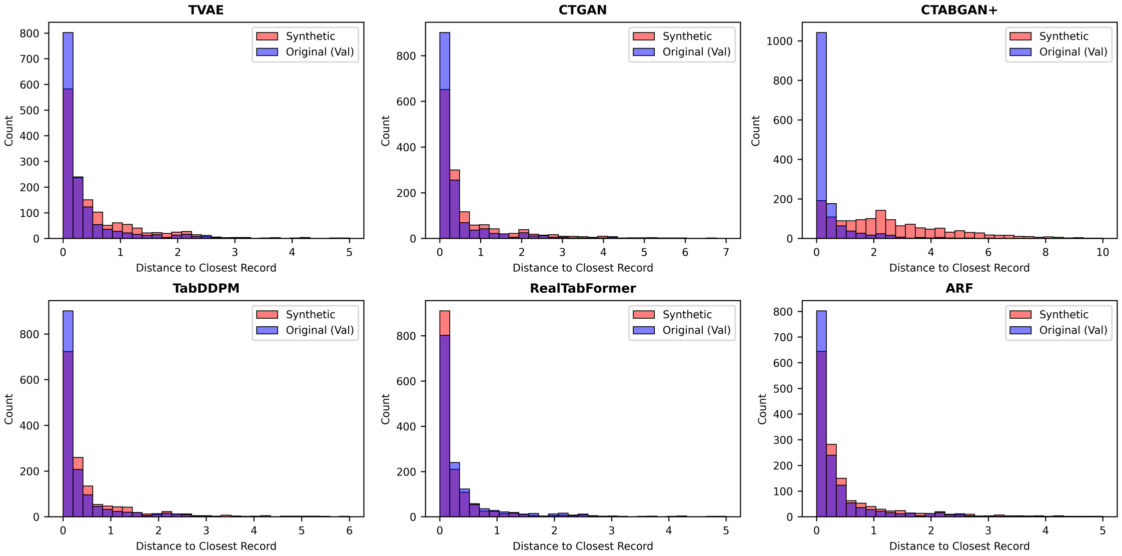


Figure 4.20: Comparison between DCR distribution for synthetic and validation datasets for the COMPAS dataset in the first fold of the cross-validation.

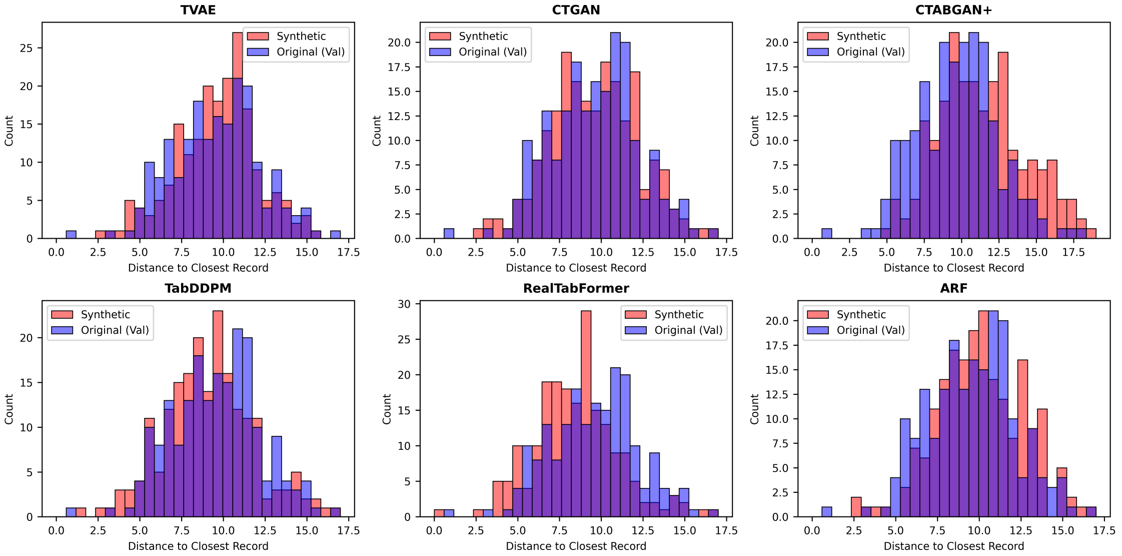


Figure 4.21: Comparison between DCR distribution for synthetic and validation datasets for the German dataset in the first fold of the cross-validation.

4.3 Performance and Fairness of Synthetic Datasets

The generated synthetic datasets were evaluated for their utility and fairness on the real data test set of each iteration of the cross-validation. To assess performance and fairness, LightGBM classifiers were trained on the synthetic datasets and subsequently tested on the corresponding real data test sets. The Figures 4.22 to 4.25 show these results for the Adult, Bank Marketing, Compas and German datasets.

In most cases, the use of synthetic datasets as training data for classifiers results in worse performance metrics compared to those trained on original data. For the Adult, Bank Marketing and Compas datasets, for the majority of the STDG models, the decrease in the perfomance metric was lower than 15% and the best models achieved scores very close to the classifiers trained on original data. For the german dataset however, all the models performed poorly, suggesting that the models evaluated might not perform so well on small datasets.

The worst models for the Adult dataset were CTGAN and CTABGAN, with a decrease in the MCC, respectively of 16.9% and 21.1%, while the best models were RealTabFormer and TabDDPM, that achieved an average decrease of 1.6% and 9.3%, respectivelly. For the bank marketing dataset, the worst models were Arf and CTGAN, with a decrease in the MCC, respectively of 28.5% and 24.2%, while the best models were RealTabFormer and TabDDPM, that achieved an average increase of 2.7% and a decrease of 3.1 %, respectivelly.

For the Compas dataset, the worst models were CTABGAN and TVAE with a decrease in the MCC, respectively of 23.3% and 10.7%. The best models for this dataset were TabDDPM and Arf, with a increase of 0.6% and a decrease of 1.9%. Finally, for the German dataset, as mentioned before, we can observe that all models had a large decrease in the MCC with the best model being RealTabFormer that had a decrease of 28.7% in comparison to the classifier trained on the original data.

Now focusing on the fairness metrics, we can see that for the majority of datasets and models, using synthetic datasets led to unfairer classifiers for both Demographic Parity and Equalized Odds. In the Adult dataset, for instance, the model with the best performance measured by the MCC, RealTabFormer, had an increase of 10.3% and 3.5% in the Demographic Parity and Equalized Odds, as can be seen in Figure 4.22, which means that the classifier trained on the synthetic dataset not only kept the fairness issues present in this dataset but also made them worse.

Furthermore, classifiers trained on the synthetic datasets generated by all the models in this dataset had an increase in Equalized Odds, with the smallest achieving an increase of 3.5% in RealTabFormer and the largest, CTABGAN, with an increase of 188.1%. In relation to Demographic Parity, datasets generated by the models Arf and DDPM showed a slight decrease in this metric, but the remaining four models showed an increase, with the greatest being from CTABGAN that had an increase of 38.6% in this metric.

Table 4.7: Results of performance and fairness metrics for the original and synthetic versions of the Adult dataset with the LightGBM model.

Data Source	**Accuracy**	**F1-Score**	**MCC**	**Demographic Parity**	**Equalized Odds**
Original Data	0.876 ± 0.005	0.714 ± 0.005	0.64 ± 0.01	0.18 ± 0.012	0.086 ± 0.017
RealTabFormer	0.868 ± 0.004	0.71 ± 0.016	0.628 ± 0.011	0.202 ± 0.022	0.094 ± 0.018
TabDDPM	0.856 ± 0.005	0.668 ± 0.008	0.582 ± 0.013	0.176 ± 0.011	0.108 ± 0.037
ARF	0.852 ± 0.004	0.654 ± 0.009	0.568 ± 0.008	0.176 ± 0.011	0.11 ± 0.042
TVAE	0.844 ± 0.009	0.666 ± 0.023	0.566 ± 0.013	0.23 ± 0.039	0.196 ± 0.029
CTGAN	0.83 ± 0.023	0.574 ± 0.165	0.496 ± 0.117	0.192 ± 0.082	0.224 ± 0.083
CTABGAN+	0.824 ± 0.031	0.618 ± 0.063	0.51 ± 0.083	0.234 ± 0.059	0.238 ± 0.138

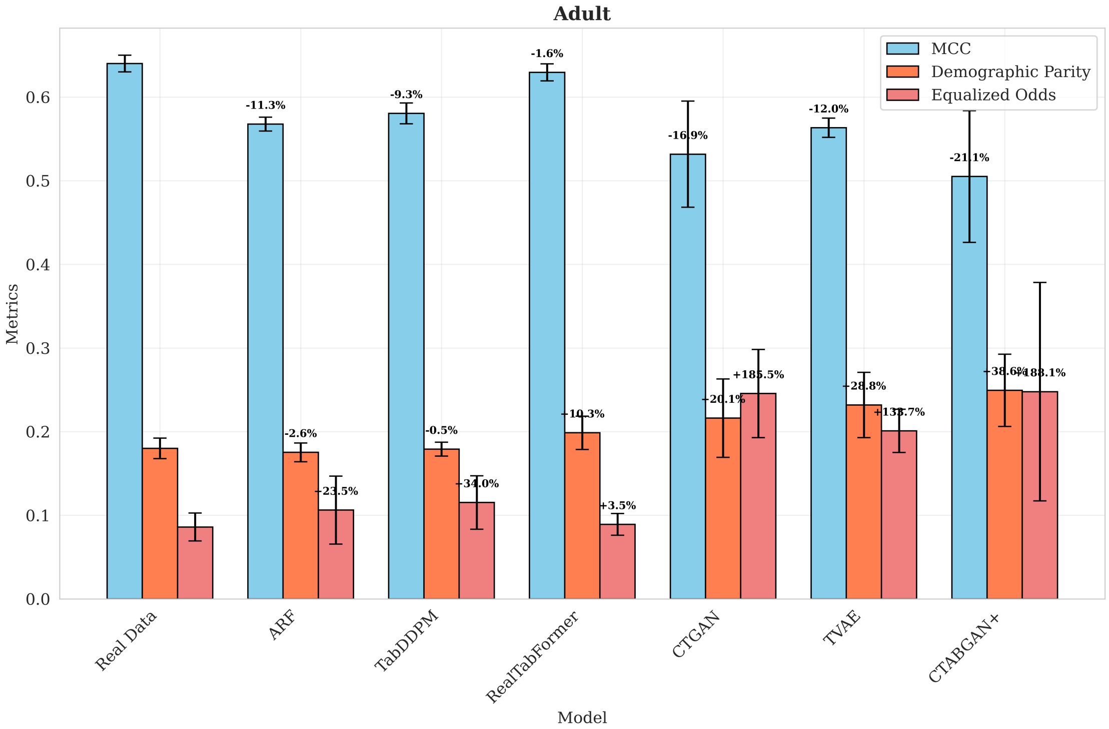


Figure 4.22: Results for MCC, Demographic Parity and Equalized Odds for the Adult dataset.

Examining the fairness metrics for the Bank Marketing dataset, we observe similar trends to those in the Adult dataset. Most synthetic data generation models also led to worse fairness outcomes compared to classifiers trained on real data.

For Demographic Parity, only Arf and TabDDPM had a decrease in this metric. For Equalized Odds, only Arf had a decrease. It is important to notice that this decrease in the fairness metrics for the Arf dataset has come at the cost of the performance metric as this was the model that achieved the lowest MCC on this dataset, as it can be seen in the Figure 4.23. For the model with the best performance, RealTabFormer, there were significant increases in the fairness metrics, with this model obtaining a Demographic Parity increase of 27.1% and a increase in Equalized Odds of 31.9%.

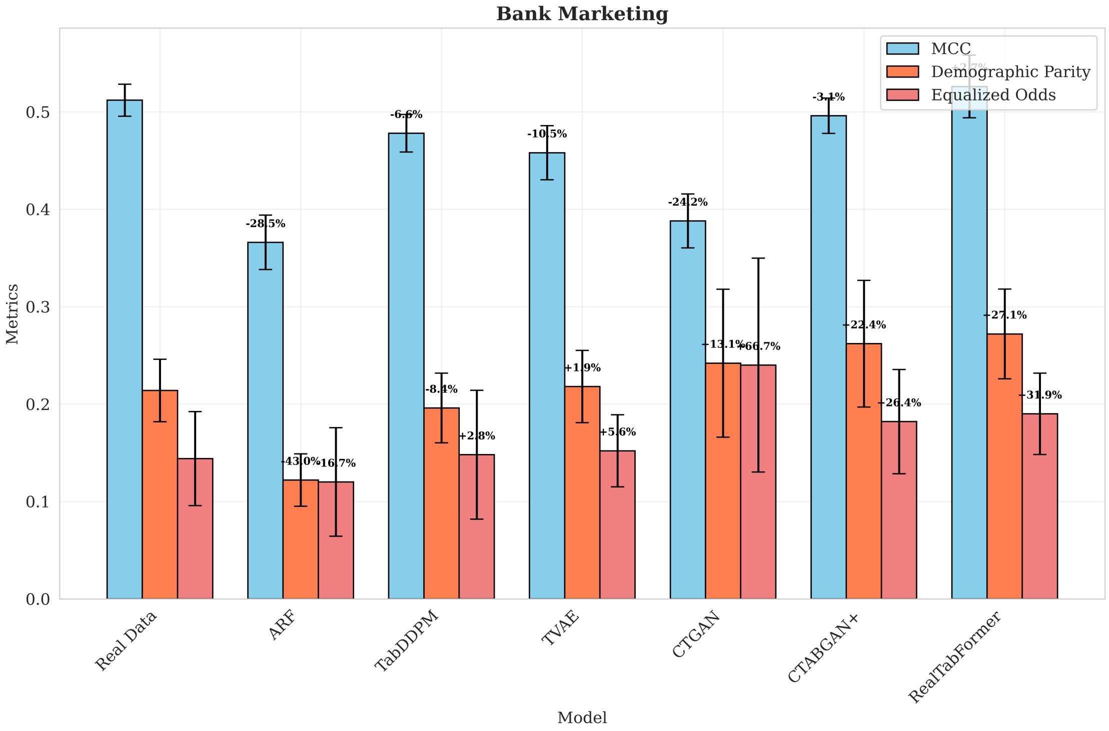


Figure 4.23: Results for MCC, Demographic Parity and Equalized Odds for the Bank Marketing dataset.

Table 4.8: Results of performance and fairness metrics for the original and synthetic versions of the Bank Marketing dataset with the LightGBM model.

Data Source	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	0.908 ± 0.004	0.95 ± 0.0	0.512 ± 0.016	0.214 ± 0.032	0.144 ± 0.048
TabDDPM	0.906 ± 0.005	0.95 ± 0.0	0.478 ± 0.019	0.196 ± 0.036	0.148 ± 0.066
RealTabFormer	0.902 ± 0.004	0.946 ± 0.005	0.526 ± 0.032	0.272 ± 0.046	0.19 ± 0.042
TVAE	0.9 ± 0.0	0.942 ± 0.004	0.458 ± 0.028	0.218 ± 0.037	0.152 ± 0.037
ARF	0.898 ± 0.004	0.944 ± 0.005	0.366 ± 0.028	0.122 ± 0.027	0.12 ± 0.056
CTGAN	0.892 ± 0.004	0.94 ± 0.0	0.388 ± 0.028	0.242 ± 0.076	0.24 ± 0.11
CTABGAN+	0.88 ± 0.01	0.93 ± 0.01	0.496 ± 0.018	0.262 ± 0.065	0.182 ± 0.054

Examining the fairness metrics for the Compas dataset in Figure 4.24, we observe similar patterns to those in the previous datasets. The Demographic Parity values show a slight increase for most synthetic datasets, with the highest increase of 4.9% for RealTabFormer and CTABGAN; the only exception is for CTGAN where it reduced by 23.2%. For Equalized Odds, most synthetic datasets also show increases compared to real data, with the highest values of 25.8% for CTABGAN and 13.5% for RealTabFormer, and the only reduction in this metric being for CTGAN with a decrease of 23.2%.

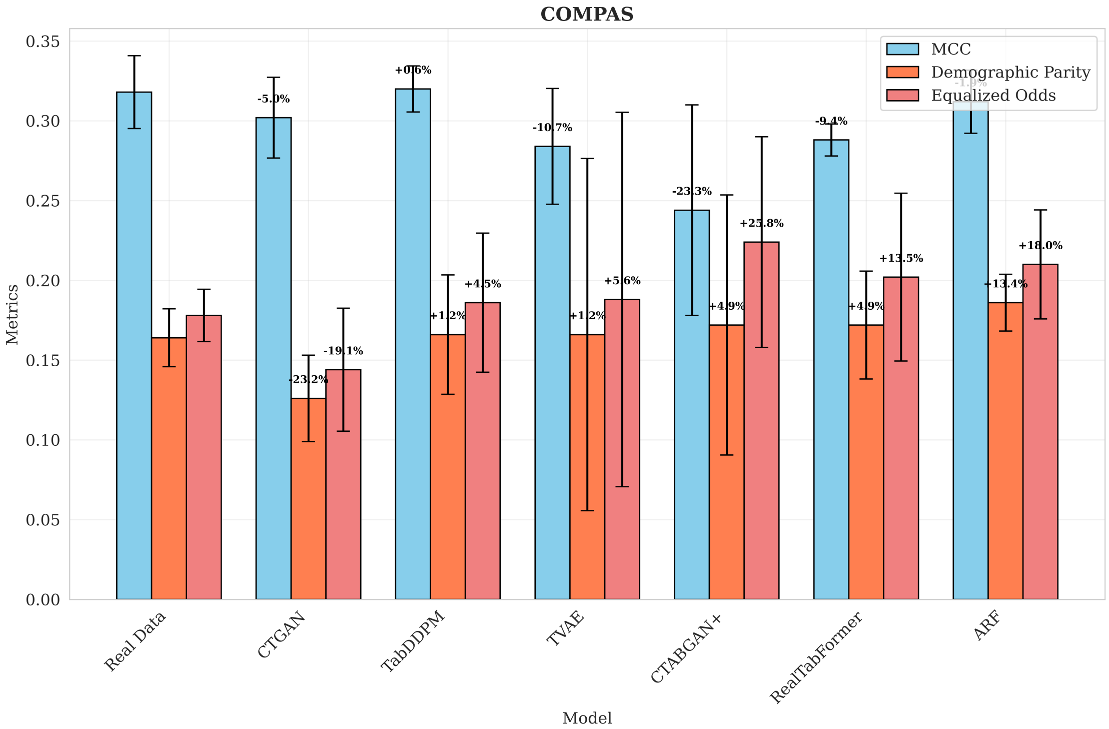


Figure 4.24: Results for MCC, Demographic Parity and Equalized Odds for the COMPAS dataset.

Table 4.9: Results of performance and fairness metrics for the original and synthetic versions of the compas dataset with the LightGBM model.

Data Source	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	0.666 ± 0.009	0.602 ± 0.015	0.318 ± 0.023	0.164 ± 0.018	0.178 ± 0.016
TabDDPM	0.668 ± 0.008	0.594 ± 0.009	0.32 ± 0.016	0.166 ± 0.041	0.186 ± 0.048
ARF	0.662 ± 0.013	0.59 ± 0.01	0.312 ± 0.022	0.186 ± 0.019	0.21 ± 0.037
CTGAN	0.658 ± 0.013	0.592 ± 0.038	0.302 ± 0.028	0.126 ± 0.03	0.144 ± 0.042
RealTabFormer	0.65 ± 0.007	0.558 ± 0.048	0.288 ± 0.011	0.172 ± 0.037	0.202 ± 0.058
TVAE	0.646 ± 0.021	0.602 ± 0.027	0.284 ± 0.04	0.166 ± 0.121	0.188 ± 0.129
CTABGAN+	0.63 ± 0.035	0.542 ± 0.046	0.244 ± 0.072	0.172 ± 0.089	0.224 ± 0.072

Finally, for the German dataset, despite the poor performance of all models, we see the same pattern for the fairness metrics as for the other datasets. For Demographic Parity, most models showed an increase in these metrics when compared to the real data, with the greatest being for CTABGAN with an increase of 117.6%. For Equalized Odds, CTGAN and RealTabFormer had slight decreases, TVAE maintained the same values, and Arf and CTABGAN showed high increases in this metric.

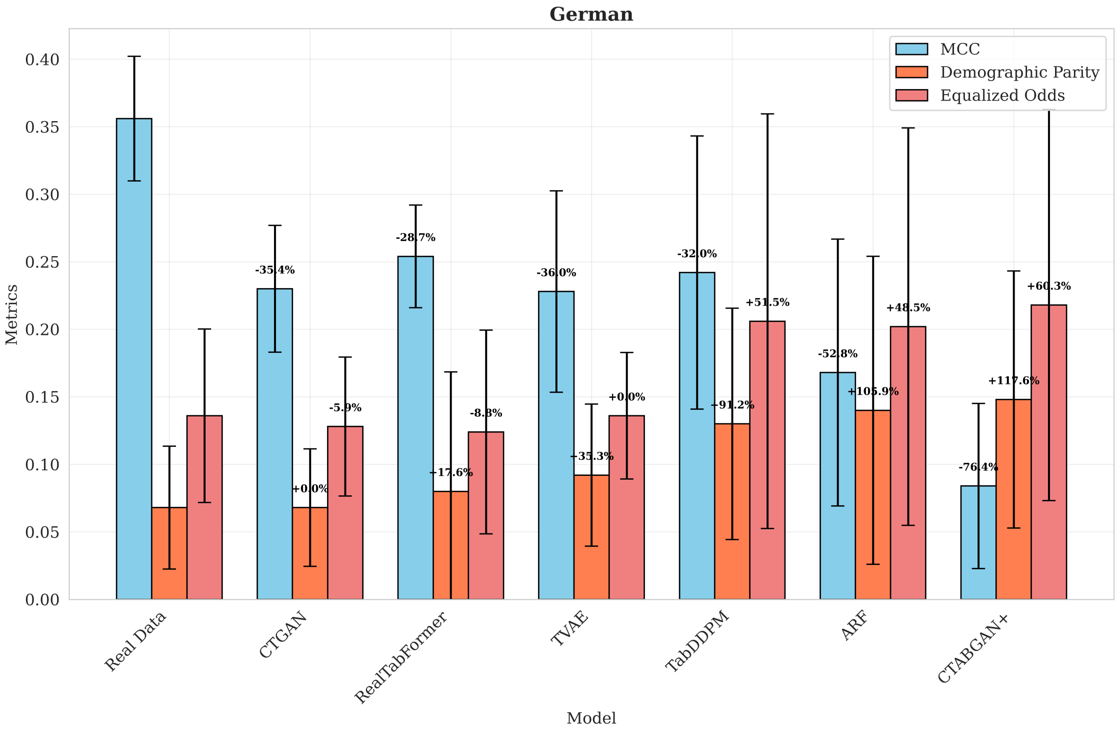


Figure 4.25: Results for MCC, Demographic Parity and Equalized Odds for the German dataset.

Table 4.10: Results of performance and fairness metrics for the original and synthetic versions of the German dataset with the LightGBM model.

Data Source	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	0.744 ± 0.025	0.522 ± 0.026	0.356 ± 0.046	0.068 ± 0.045	0.136 ± 0.064
TabDDPM	0.702 ± 0.046	0.426 ± 0.07	0.242 ± 0.111	0.13 ± 0.094	0.206 ± 0.168
RealTabFormer	0.702 ± 0.027	0.448 ± 0.027	0.254 ± 0.042	0.08 ± 0.097	0.124 ± 0.083
ARF	0.696 ± 0.03	0.324 ± 0.107	0.168 ± 0.108	0.14 ± 0.125	0.202 ± 0.161
TVAE	0.688 ± 0.037	0.434 ± 0.059	0.228 ± 0.082	0.092 ± 0.058	0.136 ± 0.051
CTGAN	0.678 ± 0.024	0.458 ± 0.043	0.23 ± 0.051	0.068 ± 0.048	0.128 ± 0.056
CTABGAN+	0.666 ± 0.031	0.248 ± 0.114	0.084 ± 0.067	0.148 ± 0.104	0.218 ± 0.159

4.4 Fairness Mitigation on Synthetic Datasets

Given the results obtained for the fairness metrics on the synthetic datasets, which indicate that synthetic datasets often lead to increases in Demographic Parity and Equalized Odds when compared to the original datasets, the importance of fairness mitigation becomes even more significant. In this section the results of the fairness mitigation experiments are shown and discussed. The Tables 4.11 to 4.18 and Figures 4.26 to 4.33 show the results of the experiments performed on the original and synthetic datasets. In these graphics, we can observe the outcomes of a baseline LightGBM classifier and a LightGBM classifier with two distinct fairness mitigation methods applied to it. Each method is run twice, with the classifier constrained by a different fairness metric: Demographic Parity or Equalized Odds.

The results demonstrate that fairness mitigation methods, such as Reductions and Threshold Optimization, are able to consistently reduce both Demographic Parity and Equalized Odds across various synthetic and real datasets. This is evident when comparing the baseline LightGBM classifier with its counterparts that employ these fairness mitigation methods, as illustrated in the tables and figures bellow, which showcase either Demographic Partity or Equalized Odds. Furthermore, in most cases, the effectiveness of fairness mitigation algorithms is quite comparable for models trained using synthetic or real datasets, even when the synthetic data resulted in greater unfairness compared to the original data.

Before delving into each experiment individually, it's important to note that while the experiments conducted with the Adult, COMPAS, and Bank Marketing datasets exhibit very similar patterns, this is not the case for the German dataset. The results from the German dataset demonstrate a clear different behavior. A plausible explanation for this discrepancy lies in the fact that the STDG Models were unable to generate reliable synthetic datasets for the German dataset. This could be attributed to the relatively small size of the dataset, as discussed in the preceding section.

4.4.1 Dataset: Adult

Now, proceeding to analyse each of the results of each experiment individually, it can be seen in Table 4.11 and in Figure 4.26 that in the Adult dataset, the use of fairness mitigation algorithms on the baseline classifiers trained on real data was effective in reducing the Demographic Parity without much loss in the performance. The Reductions and Threshold Optimizer methods were able to decrease the average Demographic parity across the five folds from 0.18 to 0.016 and 0.01, respectively, with the Accuracy going from 0.876 to 0.856 and 0.85 and a sharper but still small decrease in the MCC going from 0.64 to 0.574 and 0.564, respectively.

The results for the best synthetic data generator in this dataset, RealTabFormer, are very similar to those found in the real data. Even though the baseline classifier without fairness mitigation interventions has a greater Demographic Parity than its counterpart trained on the real data, the addition of the fairness algorithms to the prediction pipeline was able to decrease the Demographic Parity from 0.202 to 0.034 and 0.02 with the Reductions and Threshold Optimizer methods, respectively, while maintaining a MCC of 0.572 and 0.56 versus a MCC of 0.628 for the classifier without fairness interventions. This pattern repeats for the synthetic datasets generated by most of the STDG models, with small decreases in the performance metrics and significant reductions in the Demographic Parity, with the only exception being for the synthetic datasets generated by the CTABGAN+ model, in which there was a more significant decrease in the performance metrics when the fairness algorithms were added to the pipeline.

Table 4.11: Results of performance and fairness metrics for the fairness mitigation experiments of the Adult dataset with Demographic Parity as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.876 ± 0.005	0.714 ± 0.005	0.64 ± 0.01	0.18 ± 0.012	0.086 ± 0.017
	LightGBM with Reduction	0.856 ± 0.005	0.654 ± 0.005	0.574 ± 0.005	0.016 ± 0.011	0.292 ± 0.032
	LightGBM with Threshold Optimization	0.85 ± 0.0	0.648 ± 0.011	0.564 ± 0.005	0.01 ± 0.007	0.316 ± 0.036
RealTabFormer	LightGBM	0.868 ± 0.004	0.71 ± 0.016	0.628 ± 0.011	0.202 ± 0.022	0.094 ± 0.018
	LightGBM with Reduction	0.851 ± 0.01	0.659 ± 0.022	0.572 ± 0.022	0.034 ± 0.045	0.233 ± 0.064
	LightGBM with Threshold Optimization	0.846 ± 0.01	0.657 ± 0.021	0.56 ± 0.019	0.02 ± 0.034	0.262 ± 0.051
TabDDPM	LightGBM	0.856 ± 0.005	0.668 ± 0.008	0.582 ± 0.013	0.176 ± 0.011	0.108 ± 0.037
	LightGBM with Reduction	0.84 ± 0.0	0.605 ± 0.005	0.525 ± 0.009	0.021 ± 0.008	0.294 ± 0.036
	LightGBM with Threshold Optimization	0.836 ± 0.005	0.596 ± 0.007	0.511 ± 0.009	0.006 ± 0.013	0.337 ± 0.037
ARF	LightGBM	0.852 ± 0.004	0.654 ± 0.009	0.568 ± 0.008	0.176 ± 0.011	0.11 ± 0.042
	LightGBM with Reduction	0.837 ± 0.008	0.605 ± 0.014	0.521 ± 0.018	0.029 ± 0.016	0.251 ± 0.06
	LightGBM with Threshold Optimization	0.836 ± 0.005	0.609 ± 0.01	0.517 ± 0.01	0.012 ± 0.009	0.3 ± 0.038
TVAE	LightGBM	0.844 ± 0.009	0.666 ± 0.023	0.566 ± 0.013	0.23 ± 0.039	0.196 ± 0.029
	LightGBM with Reduction	0.84 ± 0.006	0.629 ± 0.033	0.537 ± 0.025	0.095 ± 0.044	0.13 ± 0.082
	LightGBM with Threshold Optimization	0.831 ± 0.012	0.631 ± 0.033	0.529 ± 0.026	0.083 ± 0.042	0.153 ± 0.09
CTGAN	LightGBM	0.83 ± 0.023	0.574 ± 0.165	0.496 ± 0.117	0.192 ± 0.082	0.224 ± 0.083
	LightGBM with Reduction	0.827 ± 0.015	0.556 ± 0.108	0.472 ± 0.074	0.039 ± 0.028	0.216 ± 0.064
	LightGBM with Threshold Optimization	0.822 ± 0.013	0.556 ± 0.103	0.464 ± 0.075	0.028 ± 0.028	0.266 ± 0.08
CTABGAN+	LightGBM	0.824 ± 0.031	0.618 ± 0.063	0.51 ± 0.083	0.234 ± 0.059	0.238 ± 0.138
	LightGBM with Reduction	0.774 ± 0.069	0.534 ± 0.099	0.391 ± 0.15	0.13 ± 0.146	0.278 ± 0.197
	LightGBM with Threshold Optimization	0.766 ± 0.057	0.512 ± 0.095	0.364 ± 0.137	0.095 ± 0.102	0.248 ± 0.113

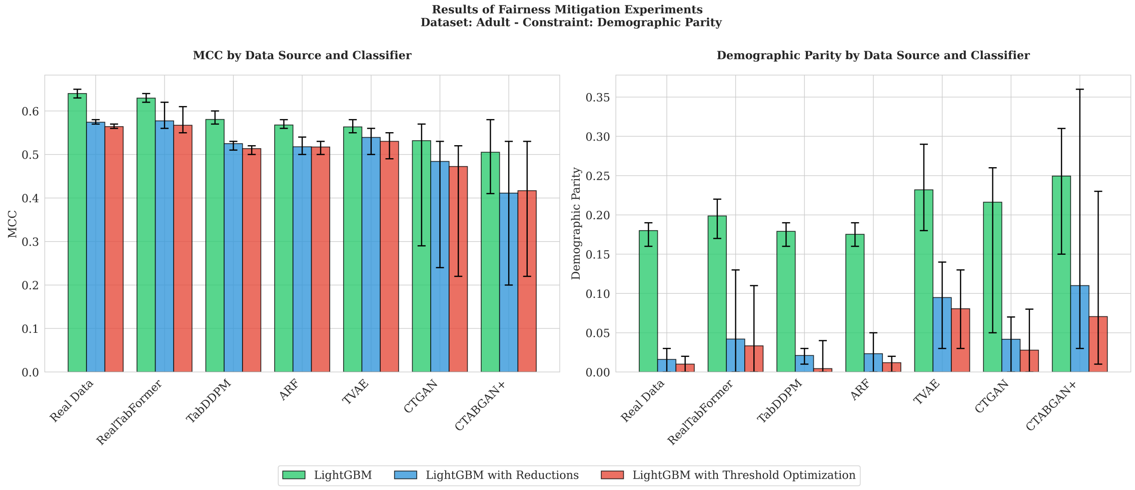


Figure 4.26: Results for MCC and Demographic Parity for the Adult dataset.

When the fairness mitigation algorithms were set to reduce Equalized Odds in the Adult dataset, the results were quite similar, as shown in Table 4.12 and Figure 4.27. It can be seen that, for the classifiers trained on real data, the addition of fairness mitigation algorithms decreased the Equalized Odds substantially with minimal decreases in the performance metrics. In this case, with RealTabFormer, the STDG model with the best performance metrics in this dataset, even though there were reductions in the Equalized Odds, it was not on the same level as it was with the classifiers trained on real data. Similar patterns were observed for the other STDG models, even in cases where there was a great increase in the Equalized Odds with the use of synthetic data by classifiers trained without fairness mitigation algorithms. This was the case with CTGAN and CTABGAN+ where they reported Equalized Odds values of 0.224 and 0.238, respectively, in opposition to 0.086 in their counterpart trained on real data. With the addition of fairness mitigation algorithms, these values reduced to 0.057 and 0.048 for CTGAN with the Reductions and the Thresholder Optimizer algorithms, respectively, and for 0.076 and 0.078 with CTABGAN+ with the same respective algorithms. For CTGAN, the use of the fairness algorithms led to increases in the performance metrics and for CTABGAN+ there were decreases.

Table 4.12: Results of performance and fairness metrics for the fairness mitigation experiments of the Adult dataset with Equalized Odds as constraint.

Data Source	**Classifier**	**Accuracy**	**F1-Score**	**MCC**	**Demographic Parity**	**Equalized Odds**
Original Data	LightGBM	0.876 ± 0.005	0.714 ± 0.005	0.64 ± 0.01	0.18 ± 0.012	0.086 ± 0.017
	LightGBM with Reduction	0.87 ± 0.0	0.694 ± 0.005	0.62 ± 0.01	0.126 ± 0.009	0.03 ± 0.007
	LightGBM with Threshold Optimization	0.86 ± 0.0	0.676 ± 0.009	0.598 ± 0.004	0.114 ± 0.009	0.03 ± 0.023
RealTabFormer	LightGBM	0.868 ± 0.004	0.71 ± 0.016	0.628 ± 0.011	0.202 ± 0.022	0.094 ± 0.018
	LightGBM with Reduction	0.865 ± 0.005	0.693 ± 0.021	0.614 ± 0.015	0.154 ± 0.018	0.053 ± 0.015
	LightGBM with Threshold Optimization	0.858 ± 0.004	0.678 ± 0.02	0.594 ± 0.017	0.142 ± 0.017	0.076 ± 0.043
TabDDPM	LightGBM	0.856 ± 0.005	0.668 ± 0.008	0.582 ± 0.013	0.176 ± 0.011	0.108 ± 0.037
	LightGBM with Reduction	0.852 ± 0.004	0.635 ± 0.005	0.558 ± 0.008	0.121 ± 0.01	0.043 ± 0.013
	LightGBM with Threshold Optimization	0.842 ± 0.004	0.623 ± 0.007	0.536 ± 0.007	0.107 ± 0.012	0.044 ± 0.026
ARF	LightGBM	0.852 ± 0.004	0.654 ± 0.009	0.568 ± 0.008	0.176 ± 0.011	0.11 ± 0.042
	LightGBM with Reduction	0.851 ± 0.005	0.643 ± 0.011	0.561 ± 0.015	0.115 ± 0.014	0.069 ± 0.031
	LightGBM with Threshold Optimization	0.842 ± 0.004	0.627 ± 0.005	0.535 ± 0.012	0.106 ± 0.01	0.045 ± 0.027
TVAE	LightGBM	0.844 ± 0.009	0.666 ± 0.023	0.566 ± 0.013	0.23 ± 0.039	0.196 ± 0.029
	LightGBM with Reduction	0.844 ± 0.007	0.642 ± 0.038	0.553 ± 0.022	0.163 ± 0.042	0.075 ± 0.04
	LightGBM with Threshold Optimization	0.831 ± 0.012	0.629 ± 0.027	0.528 ± 0.021	0.153 ± 0.043	0.114 ± 0.04
CTGAN	LightGBM	0.83 ± 0.023	0.574 ± 0.165	0.496 ± 0.117	0.192 ± 0.082	0.224 ± 0.083
	LightGBM with Reduction	0.836 ± 0.014	0.606 ± 0.098	0.521 ± 0.068	0.134 ± 0.038	0.057 ± 0.022
	LightGBM with Threshold Optimization	0.83 ± 0.015	0.602 ± 0.095	0.505 ± 0.071	0.119 ± 0.025	0.048 ± 0.058
CTABGAN+	LightGBM	0.824 ± 0.031	0.618 ± 0.063	0.51 ± 0.083	0.234 ± 0.059	0.238 ± 0.138
	LightGBM with Reduction	0.806 ± 0.037	0.578 ± 0.08	0.458 ± 0.107	0.131 ± 0.03	0.076 ± 0.057
	LightGBM with Threshold Optimization	0.802 ± 0.045	0.585 ± 0.076	0.461 ± 0.106	0.111 ± 0.059	0.078 ± 0.041

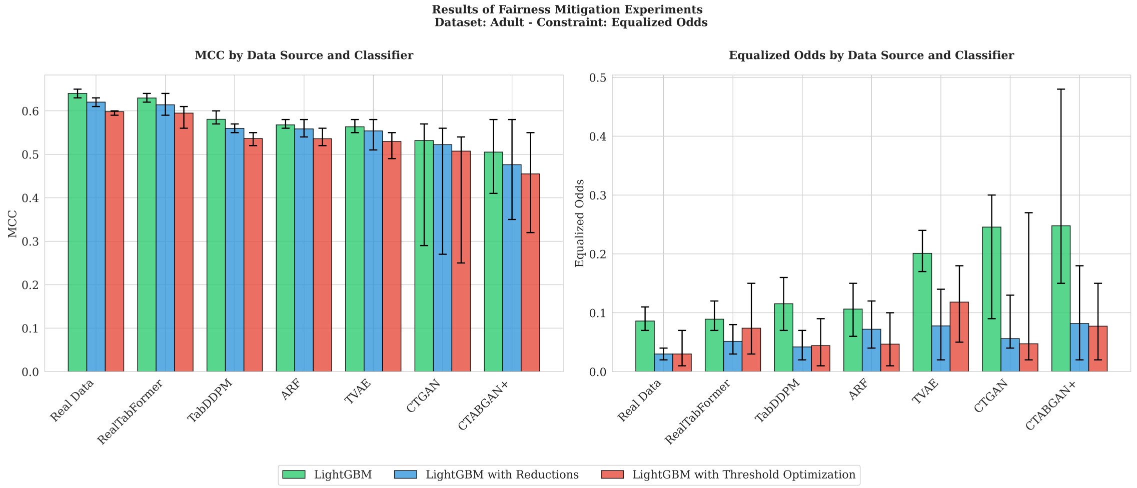


Figure 4.27: Results for MCC and Equalized Odds for the Adult dataset.

4.4.2 Dataset: Bank Marketing

The results for the fairness mitigation experiments with the Bank Marketing Dataset using Demographic Parity as constraint are shown in Table 4.28 and Figure 4.29. The Demographic Parity was significantly reduced with minor decreases in the performance metrics when fairness mitigation algorithms were added for both original and synthetic datasets. With the original dataset, the Reductions and the Threshold Optimizer method were able to reduce the Demographic Parity from 0.214 to 0.014 and 0.012, respectively, with a small decrease in the performance metrics with the Reductions algorithm and even a small increase in these metrics with the Threshold Optimization. A similar reduction in Demographic Parity can be noticed for all the synthetic datasets. Another aspect worth highlighting in this experiment is the evident trade-off between different fairness metrics. While it was possible to significantly reduce the Demographic Parity at a relatively low cost of performance, there was an increase in the Equalized Odds metric in all cases.

Table 4.13: Results of performance and fairness metrics for the fairness mitigation experiments of the Bank Marketing dataset with Demographic Parity as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.908 ± 0.004	0.556 ± 0.015	0.512 ± 0.016	0.214 ± 0.032	0.144 ± 0.048
	LightGBM with Reduction	0.906 ± 0.005	0.528 ± 0.011	0.484 ± 0.015	0.014 ± 0.013	0.258 ± 0.059
	LightGBM with Threshold Optimization	0.908 ± 0.004	0.564 ± 0.021	0.516 ± 0.018	0.012 ± 0.008	0.342 ± 0.053
TabDDPM	LightGBM	0.906 ± 0.005	0.51 ± 0.021	0.478 ± 0.019	0.196 ± 0.036	0.148 ± 0.066
	LightGBM with Reduction	0.904 ± 0.005	0.478 ± 0.036	0.446 ± 0.034	0.014 ± 0.005	0.226 ± 0.047
	LightGBM with Threshold Optimization	0.902 ± 0.004	0.496 ± 0.036	0.458 ± 0.03	0.012 ± 0.008	0.292 ± 0.054
RealTabFormer	LightGBM	0.902 ± 0.004	0.576 ± 0.04	0.526 ± 0.032	0.272 ± 0.046	0.19 ± 0.042
	LightGBM with Reduction	0.902 ± 0.004	0.546 ± 0.048	0.498 ± 0.034	0.01 ± 0.012	0.316 ± 0.099
	LightGBM with Threshold Optimization	0.9 ± 0.007	0.544 ± 0.043	0.496 ± 0.03	0.02 ± 0.014	0.364 ± 0.069
TVAE	LightGBM	0.9 ± 0.0	0.51 ± 0.035	0.458 ± 0.028	0.218 ± 0.037	0.152 ± 0.037
	LightGBM with Reduction	0.894 ± 0.005	0.478 ± 0.036	0.432 ± 0.029	0.018 ± 0.016	0.262 ± 0.081
	LightGBM with Threshold Optimization	0.894 ± 0.005	0.49 ± 0.023	0.436 ± 0.024	0.032 ± 0.015	0.352 ± 0.075
ARF	LightGBM	0.898 ± 0.004	0.362 ± 0.033	0.366 ± 0.028	0.122 ± 0.027	0.12 ± 0.056
	LightGBM with Reduction	0.896 ± 0.005	0.354 ± 0.038	0.358 ± 0.029	0.022 ± 0.004	0.112 ± 0.033
	LightGBM with Threshold Optimization	0.9 ± 0.0	0.424 ± 0.022	0.406 ± 0.011	0.018 ± 0.011	0.164 ± 0.038
CTGAN	LightGBM	0.892 ± 0.004	0.426 ± 0.04	0.388 ± 0.028	0.242 ± 0.076	0.24 ± 0.11
	LightGBM with Reduction	0.888 ± 0.004	0.404 ± 0.021	0.364 ± 0.018	0.02 ± 0.023	0.186 ± 0.061
	LightGBM with Threshold Optimization	0.882 ± 0.008	0.444 ± 0.029	0.386 ± 0.023	0.038 ± 0.019	0.302 ± 0.091
CTABGAN+	LightGBM	0.88 ± 0.01	0.556 ± 0.018	0.496 ± 0.018	0.262 ± 0.065	0.182 ± 0.054
	LightGBM with Reduction	0.878 ± 0.008	0.534 ± 0.021	0.47 ± 0.025	0.05 ± 0.041	0.346 ± 0.072
	LightGBM with Threshold Optimization	0.87 ± 0.01	0.53 ± 0.02	0.466 ± 0.017	0.048 ± 0.029	0.374 ± 0.062

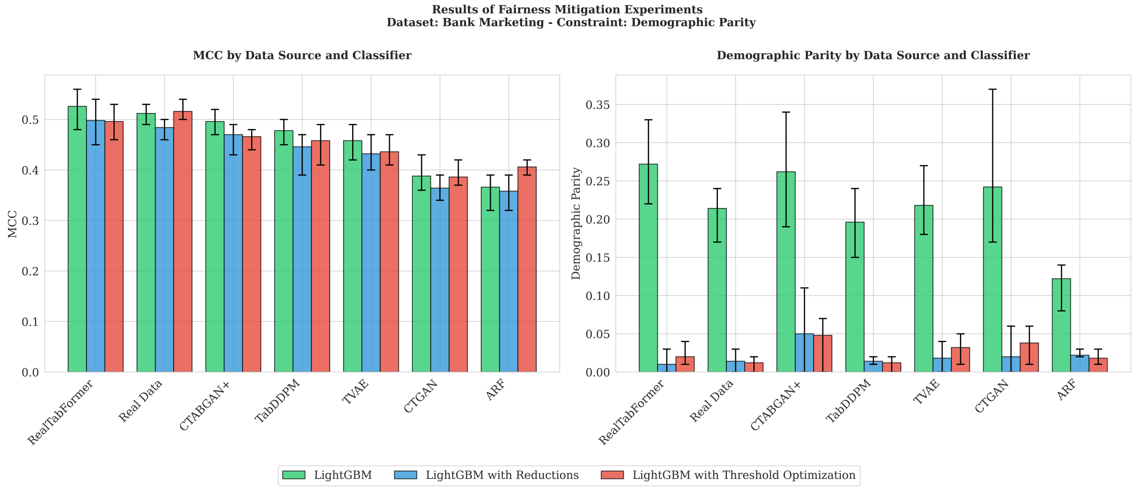


Figure 4.28: Results for MCC and Demographic Parity for the Bank Marketing dataset.

Similar results are found when Equalized Odds is used as constraint in this dataset, as shown in the Table 4.14 and Figure 4.29. There was a significant reduction in Equalized Odds when fairness mitigation algorithm were applied to the baseline classifier, for all datasets, including the original and the synthetic ones. The reduction in these datasets was not as large for Demographic Parity when Demographic Parity was set as constraint, but they were still quite relevant and also with only minor decreases in the performance metric. In contrast to the trade-off observed when Demographic Parity is set as a constraint during optimization for Equalized Odds, there was no increase in Demographic Parity but a reasonable de-

Table 4.14: Results of performance and fairness metrics for the fairness mitigation experiments of the Bank Marketing dataset with Equalized Odds as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.908 ± 0.004	0.556 ± 0.015	0.512 ± 0.016	0.214 ± 0.032	0.144 ± 0.048
	LightGBM with Reduction	0.91 ± 0.0	0.544 ± 0.011	0.506 ± 0.011	0.128 ± 0.019	0.068 ± 0.026
	LightGBM with Threshold Optimization	0.902 ± 0.004	0.498 ± 0.027	0.458 ± 0.016	0.098 ± 0.016	0.06 ± 0.041
TabDDPM	LightGBM	0.906 ± 0.005	0.51 ± 0.021	0.478 ± 0.019	0.196 ± 0.036	0.148 ± 0.066
	LightGBM with Reduction	0.902 ± 0.004	0.488 ± 0.019	0.456 ± 0.018	0.12 ± 0.02	0.066 ± 0.018
	LightGBM with Threshold Optimization	0.9 ± 0.0	0.488 ± 0.031	0.454 ± 0.021	0.102 ± 0.011	0.07 ± 0.022
RealTabFormer	LightGBM	0.902 ± 0.004	0.576 ± 0.04	0.526 ± 0.032	0.272 ± 0.046	0.19 ± 0.042
	LightGBM with Reduction	0.904 ± 0.005	0.568 ± 0.047	0.522 ± 0.04	0.17 ± 0.034	0.106 ± 0.017
	LightGBM with Threshold Optimization	0.898 ± 0.008	0.498 ± 0.018	0.452 ± 0.023	0.114 ± 0.032	0.074 ± 0.021
TVAE	LightGBM	0.9 ± 0.0	0.51 ± 0.035	0.458 ± 0.028	0.218 ± 0.037	0.152 ± 0.037
	LightGBM with Reduction	0.9 ± 0.0	0.496 ± 0.036	0.452 ± 0.026	0.124 ± 0.036	0.09 ± 0.04
	LightGBM with Threshold Optimization	0.888 ± 0.008	0.458 ± 0.064	0.408 ± 0.049	0.088 ± 0.044	0.096 ± 0.055
ARF	LightGBM	0.898 ± 0.004	0.362 ± 0.033	0.366 ± 0.028	0.122 ± 0.027	0.12 ± 0.056
	LightGBM with Reduction	0.896 ± 0.005	0.358 ± 0.039	0.358 ± 0.033	0.076 ± 0.017	0.038 ± 0.013
	LightGBM with Threshold Optimization	0.896 ± 0.005	0.394 ± 0.069	0.376 ± 0.044	0.078 ± 0.022	0.048 ± 0.02
CTGAN	LightGBM	0.892 ± 0.004	0.426 ± 0.04	0.388 ± 0.028	0.242 ± 0.076	0.24 ± 0.11
	LightGBM with Reduction	0.89 ± 0.0	0.404 ± 0.047	0.37 ± 0.037	0.086 ± 0.048	0.082 ± 0.035
	LightGBM with Threshold Optimization	0.886 ± 0.009	0.466 ± 0.03	0.408 ± 0.022	0.09 ± 0.048	0.12 ± 0.071
CTABGAN+	LightGBM	0.88 ± 0.01	0.556 ± 0.018	0.496 ± 0.018	0.262 ± 0.065	0.182 ± 0.054
	LightGBM with Reduction	0.88 ± 0.01	0.548 ± 0.019	0.486 ± 0.023	0.128 ± 0.032	0.134 ± 0.034
	LightGBM with Threshold Optimization	0.88 ± 0.012	0.536 ± 0.023	0.47 ± 0.028	0.092 ± 0.042	0.166 ± 0.077

crease. This highlights the significance of evaluating multiple fairness metrics and carefully selecting the appropriate one in each scenario.

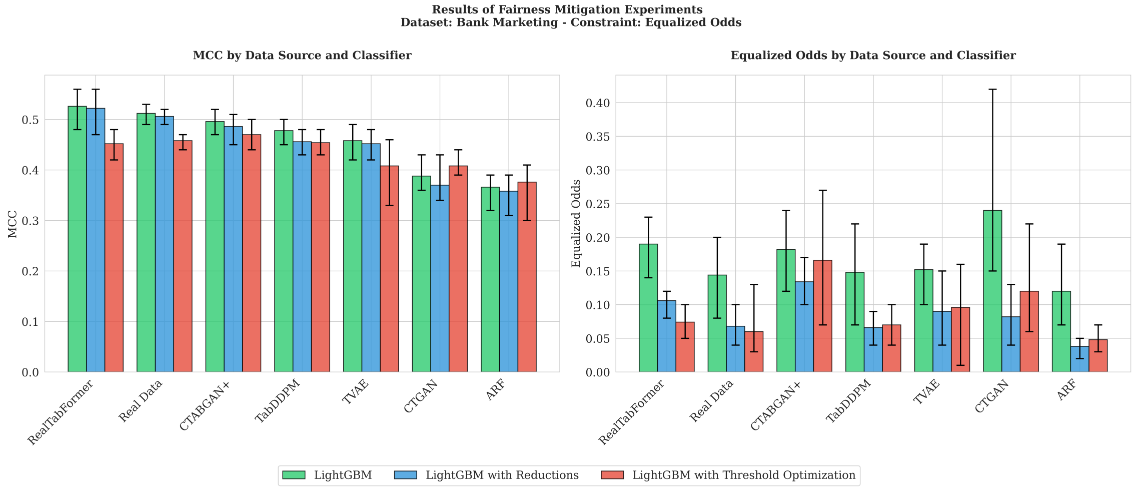


Figure 4.29: Results for MCC and Equalized Odds for the Bank Marketing dataset.

4.4.3 Dataset: COMPAS

Examining the COMPAS dataset results, presented in Table 4.15 and Figure 4.30, the fairness mitigation algorithms demonstrated consistent effectiveness in reducing Demographic Parity across all models. For classifiers trained on real data, the Reductions and Threshold Optimizer methods successfully reduced Demographic Parity from 0.164 to 0.018 and 0.016, respectively, with relatively modest decreases in performance metrics - Accuracy declined from 0.666 to 0.660 and 0.656, while MCC values decreased from 0.318 to 0.310 and 0.302, respectively.

Similar patterns were found for the classifiers trained on synthetic datasets. In

Table 4.15: Results of performance and fairness metrics for the fairness mitigation experiments of the Compas dataset with Demographic Parity as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.666 ± 0.009	0.602 ± 0.015	0.318 ± 0.023	0.164 ± 0.018	0.178 ± 0.016
	LightGBM with Reduction	0.66 ± 0.007	0.596 ± 0.018	0.31 ± 0.019	0.018 ± 0.018	0.054 ± 0.027
	LightGBM with Threshold Optimization	0.656 ± 0.005	0.598 ± 0.022	0.302 ± 0.015	0.016 ± 0.013	0.066 ± 0.035
TabDDPM	LightGBM	0.668 ± 0.008	0.594 ± 0.009	0.32 ± 0.016	0.166 ± 0.041	0.186 ± 0.048
	LightGBM with Reduction	0.661 ± 0.007	0.589 ± 0.012	0.309 ± 0.012	0.026 ± 0.024	0.053 ± 0.014
	LightGBM with Threshold Optimization	0.662 ± 0.007	0.589 ± 0.013	0.31 ± 0.016	0.019 ± 0.015	0.062 ± 0.029
ARF	LightGBM	0.662 ± 0.013	0.59 ± 0.01	0.312 ± 0.022	0.186 ± 0.019	0.21 ± 0.037
	LightGBM with Reduction	0.659 ± 0.012	0.585 ± 0.016	0.306 ± 0.022	0.032 ± 0.012	0.065 ± 0.015
	LightGBM with Threshold Optimization	0.661 ± 0.014	0.582 ± 0.012	0.312 ± 0.027	0.018 ± 0.017	0.051 ± 0.019
CTGAN	LightGBM	0.658 ± 0.013	0.592 ± 0.038	0.302 ± 0.028	0.126 ± 0.03	0.144 ± 0.042
	LightGBM with Reduction	0.648 ± 0.009	0.581 ± 0.028	0.284 ± 0.021	0.02 ± 0.02	0.054 ± 0.025
	LightGBM with Threshold Optimization	0.649 ± 0.011	0.577 ± 0.032	0.285 ± 0.023	0.018 ± 0.018	0.06 ± 0.027
RealTabFormer	LightGBM	0.65 ± 0.007	0.558 ± 0.048	0.288 ± 0.011	0.172 ± 0.037	0.202 ± 0.058
	LightGBM with Reduction	0.649 ± 0.007	0.556 ± 0.04	0.286 ± 0.012	0.034 ± 0.023	0.056 ± 0.031
	LightGBM with Threshold Optimization	0.645 ± 0.005	0.549 ± 0.034	0.28 ± 0.008	0.03 ± 0.021	0.059 ± 0.019
TVAE	LightGBM	0.646 ± 0.021	0.602 ± 0.027	0.284 ± 0.04	0.166 ± 0.121	0.188 ± 0.129
	LightGBM with Reduction	0.642 ± 0.017	0.599 ± 0.023	0.279 ± 0.034	0.072 ± 0.047	0.095 ± 0.06
	LightGBM with Threshold Optimization	0.643 ± 0.022	0.6 ± 0.028	0.279 ± 0.042	0.056 ± 0.033	0.082 ± 0.035
CTABGAN+	LightGBM	0.63 ± 0.035	0.542 ± 0.046	0.244 ± 0.072	0.172 ± 0.089	0.224 ± 0.072
	LightGBM with Reduction	0.61 ± 0.032	0.525 ± 0.035	0.208 ± 0.06	0.054 ± 0.043	0.117 ± 0.057
	LightGBM with Threshold Optimization	0.614 ± 0.025	0.516 ± 0.016	0.213 ± 0.055	0.054 ± 0.047	0.097 ± 0.049

this dataset, TabDDPM was the best performer, achieving performance metrics even slightly better than it‘s counterpart in real data for the baseline classifier without fairness interventions, with a MCC of 0.320 versus 0.318 on the real data. The addition of the fairness algorithms reduced Demographic Parity from 0.166 to 0.026 and 0.019, with minimal performance degradation. ARF demonstrated similar patterns, showing effective fairness mitigation with Demographic Parity reductions from 0.186 to 0.032 and 0.018, while maintaining MCC scores of 0.306 and 0.312. RealTabFormer, despite being the top performer in previous datasets, showed more modest results in COMPAS, with baseline MCC of 0.288. Still, the use of fairness algorithms led to improvements in the Demographic Parity from 0.172 to 0.034 and 0.030. CTABGAN+ demonstrated the most significant performance challenges, with the lowest baseline MCC of 0.244 and substantial decreases when fairness algorithms were applied, though still achieving meaningful fairness improvements from 0.172 to 0.054 in both mitigation approaches. It can be noted that overall the Threshold Optimization method had an advantage over the Reductions method in this dataset with Demographic Parity set as constraints, as it achieved equal or smaller values of Demographic Parity at a similar performance drop.

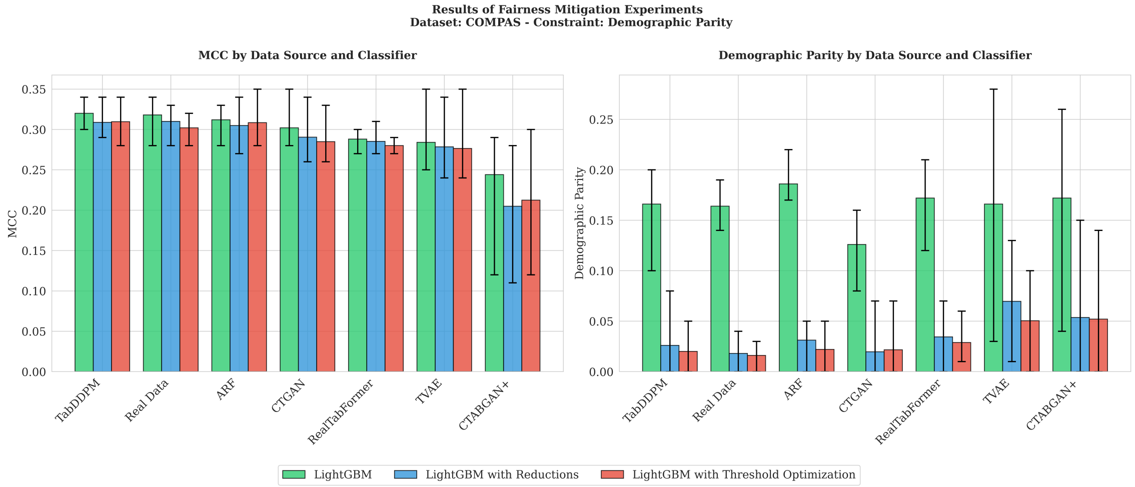


Figure 4.30: Results for MCC and Demographic Parity for the Compas dataset.

When configured to minimize Equalized Odds, as shown in Table 4.16 and Figure 4.31, the results revealed similar patterns. Real data classifiers effectively reduced Equalized Odds from 0.178 to 0.048 and 0.056 with the Reductions and Threshold Optimizer methods, while maintaining reasonable performance with MCC values of 0.302 and 0.298. Among the synthetic data generators, the response was similar in comparison to the real data. TabDDPM maintained strong performance while reducing Equalized Odds from 0.186 to 0.061 and 0.051, though with a more noticeable performance drop in the Threshold Optimizer approach with it‘s MCC going from 0.320 to 0.270. ARF, CTGAN demonstrated similar patterns overall. RealTabFormer exhibited mixed results, with a slightly smaller decrease in the Reductions method, which reduced Equalized Odds to 0.086. In contrast, it experienced a greater decrease with the Thresholds Optimization methods, albeit at a higher performance penalty. TVAE and CTABGAN+ demonstrated the worst results after fairness mitigation, both in terms of fairness reduction and performance drop. Despite performing worse than the other STDG models, fairness mitigation was able to reduce Equalized Odds to values around 50% of the original in these cases.

Table 4.16: Results of performance and fairness metrics for the fairness mitigation experiments of the Compas dataset with Equalized Odds as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
TabDDPM	LightGBM	0.668 ± 0.008	0.594 ± 0.009	0.32 ± 0.016	0.166 ± 0.041	0.186 ± 0.048
	LightGBM with Reduction	0.661 ± 0.007	0.59 ± 0.011	0.31 ± 0.009	0.033 ± 0.027	0.061 ± 0.02
	LightGBM with Threshold Optimization	0.639 ± 0.009	0.586 ± 0.022	0.27 ± 0.015	0.052 ± 0.026	0.051 ± 0.018
Original Data	LightGBM	0.666 ± 0.009	0.602 ± 0.015	0.318 ± 0.023	0.164 ± 0.018	0.178 ± 0.016
	LightGBM with Reduction	0.658 ± 0.004	0.592 ± 0.015	0.302 ± 0.013	0.032 ± 0.022	0.048 ± 0.03
	LightGBM with Threshold Optimization	0.654 ± 0.011	0.594 ± 0.032	0.298 ± 0.023	0.03 ± 0.027	0.056 ± 0.029
ARF	LightGBM	0.662 ± 0.013	0.59 ± 0.01	0.312 ± 0.022	0.186 ± 0.019	0.21 ± 0.037
	LightGBM with Reduction	0.661 ± 0.013	0.585 ± 0.021	0.31 ± 0.023	0.043 ± 0.022	0.059 ± 0.025
	LightGBM with Threshold Optimization	0.648 ± 0.02	0.587 ± 0.02	0.289 ± 0.031	0.066 ± 0.02	0.068 ± 0.036
CTGAN	LightGBM	0.658 ± 0.013	0.592 ± 0.038	0.302 ± 0.028	0.126 ± 0.03	0.144 ± 0.042
	LightGBM with Reduction	0.648 ± 0.012	0.585 ± 0.035	0.287 ± 0.025	0.034 ± 0.01	0.055 ± 0.018
	LightGBM with Threshold Optimization	0.646 ± 0.013	0.585 ± 0.021	0.28 ± 0.025	0.029 ± 0.009	0.045 ± 0.035
RealTabFormer	LightGBM	0.65 ± 0.007	0.558 ± 0.048	0.288 ± 0.011	0.172 ± 0.037	0.202 ± 0.058
	LightGBM with Reduction	0.65 ± 0.004	0.556 ± 0.04	0.289 ± 0.008	0.063 ± 0.02	0.086 ± 0.043
	LightGBM with Threshold Optimization	0.645 ± 0.009	0.54 ± 0.029	0.276 ± 0.02	0.042 ± 0.013	0.05 ± 0.023
TVAE	LightGBM	0.646 ± 0.021	0.602 ± 0.027	0.284 ± 0.04	0.166 ± 0.121	0.188 ± 0.129
	LightGBM with Reduction	0.644 ± 0.02	0.598 ± 0.026	0.278 ± 0.039	0.08 ± 0.061	0.102 ± 0.061
	LightGBM with Threshold Optimization	0.634 ± 0.025	0.599 ± 0.034	0.265 ± 0.043	0.081 ± 0.048	0.102 ± 0.048
CTABGAN+	LightGBM	0.63 ± 0.035	0.542 ± 0.046	0.244 ± 0.072	0.172 ± 0.089	0.224 ± 0.072
	LightGBM with Reduction	0.619 ± 0.031	0.533 ± 0.039	0.222 ± 0.063	0.039 ± 0.024	0.088 ± 0.047
	LightGBM with Threshold Optimization	0.616 ± 0.035	0.522 ± 0.015	0.218 ± 0.068	0.042 ± 0.031	0.084 ± 0.053

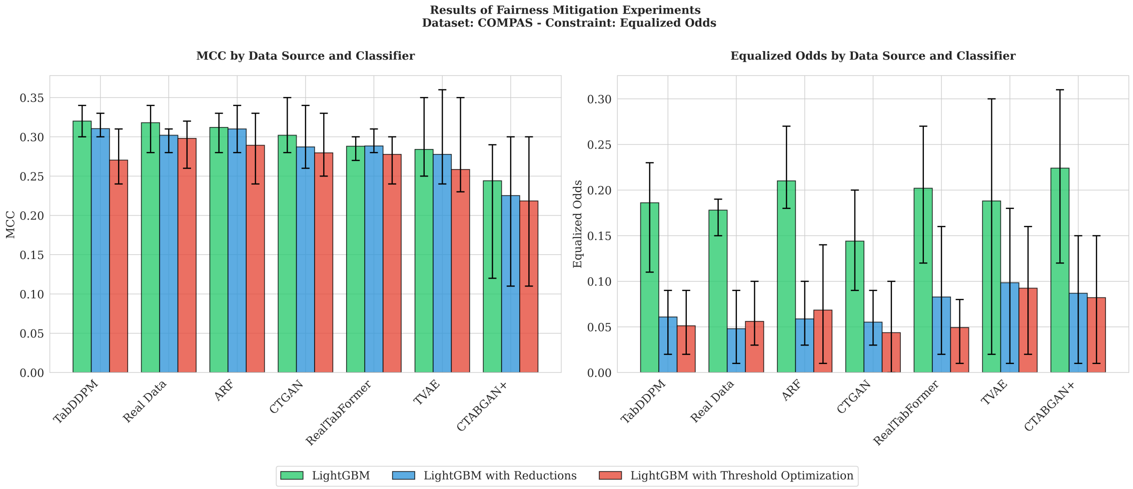


Figure 4.31: Results for MCC and Equalized Odds for the Compas dataset.

4.4.4 Dataset: German

Analyzing the German dataset results with Demographic Parity as constraint, as shown in Table 4.17 and Figure 4.32, it can be seen very distinct patterns as those observed in the previous datasets. For either real data or the synthetic ones, the fairness mitigation algorithms were not able to reduce the unfairness effectively. For real data with the Reductions algorithm there was as decrease in Demographic Parity from 0.068 to 0.042 with a small decrease in the MCC, going from 0.356 to 0.332. With the Threshold Optimization algorithm, there was an increase on Demographic Parity with the usage of the method. This behavior is counterintuitive and may be due to poor optimization related to the size of dataset, or to distribution shifts in the test sets compared to the training sets.

Table 4.17: Results of performance and fairness metrics for the fairness mitigation experiments of the German dataset with Demographic Parity as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.744 ± 0.025	0.522 ± 0.026	0.356 ± 0.046	0.068 ± 0.045	0.136 ± 0.064
	LightGBM with Reduction	0.736 ± 0.017	0.502 ± 0.032	0.332 ± 0.031	0.042 ± 0.019	0.092 ± 0.079
	LightGBM with Threshold Optimization	0.738 ± 0.02	0.476 ± 0.04	0.32 ± 0.057	0.088 ± 0.037	0.158 ± 0.091
RealTabFormer	LightGBM	0.702 ± 0.027	0.448 ± 0.027	0.254 ± 0.042	0.08 ± 0.097	0.124 ± 0.083
	LightGBM with Reduction	0.692 ± 0.025	0.44 ± 0.022	0.234 ± 0.039	0.066 ± 0.042	0.116 ± 0.049
	LightGBM with Threshold Optimization	0.695 ± 0.039	0.424 ± 0.056	0.234 ± 0.074	0.089 ± 0.05	0.166 ± 0.077
TabDDPM	LightGBM	0.702 ± 0.046	0.426 ± 0.07	0.242 ± 0.111	0.13 ± 0.094	0.206 ± 0.168
	LightGBM with Reduction	0.682 ± 0.058	0.376 ± 0.07	0.184 ± 0.129	0.212 ± 0.221	0.301 ± 0.237
	LightGBM with Threshold Optimization	0.681 ± 0.041	0.385 ± 0.083	0.186 ± 0.105	0.197 ± 0.205	0.329 ± 0.269
ARF	LightGBM	0.696 ± 0.03	0.324 ± 0.107	0.168 ± 0.108	0.14 ± 0.125	0.202 ± 0.161
	LightGBM with Reduction	0.705 ± 0.03	0.36 ± 0.106	0.204 ± 0.107	0.121 ± 0.084	0.186 ± 0.101
	LightGBM with Threshold Optimization	0.693 ± 0.014	0.292 ± 0.049	0.147 ± 0.053	0.098 ± 0.042	0.143 ± 0.055
TVAE	LightGBM	0.688 ± 0.037	0.434 ± 0.059	0.228 ± 0.082	0.092 ± 0.058	0.136 ± 0.051
	LightGBM with Reduction	0.675 ± 0.028	0.412 ± 0.044	0.195 ± 0.061	0.046 ± 0.03	0.088 ± 0.043
	LightGBM with Threshold Optimization	0.67 ± 0.025	0.391 ± 0.048	0.173 ± 0.054	0.141 ± 0.063	0.204 ± 0.088
CTGAN	LightGBM	0.678 ± 0.024	0.458 ± 0.043	0.23 ± 0.051	0.068 ± 0.048	0.128 ± 0.056
	LightGBM with Reduction	0.68 ± 0.019	0.451 ± 0.039	0.225 ± 0.047	0.056 ± 0.029	0.104 ± 0.038
	LightGBM with Threshold Optimization	0.68 ± 0.027	0.435 ± 0.05	0.214 ± 0.062	0.111 ± 0.086	0.174 ± 0.113
CTABGAN+	LightGBM	0.666 ± 0.031	0.248 ± 0.114	0.084 ± 0.067	0.148 ± 0.104	0.218 ± 0.159
	LightGBM with Reduction	0.66 ± 0.037	0.22 ± 0.097	0.06 ± 0.044	0.074 ± 0.05	0.134 ± 0.089
	LightGBM with Threshold Optimization	0.666 ± 0.041	0.272 ± 0.13	0.118 ± 0.075	0.102 ± 0.087	0.145 ± 0.089

Similar results were found for the synthetic datasets. In 4 of the 6 datasets, the use of the Threshold Optimization method led to increases in Demographic Parity. With the Reductions method, there was an increase for one dataset and decreases for the others, but not at the same level of effectiveness as seen with the Adult, Bank Marketing, and COMPAS datasets. The largest reduction in Demographic Parity occurred with the Reductions method on the dataset generated with TVAE, decreasing from 0.092 with the regular classifier to 0.046. The worst case was with the dataset generated using TabDDPM, where the Threshold Optimization method increased Demographic Parity from 0.13 to 0.197, and the Reductions method led to an increase, reaching 0.212.

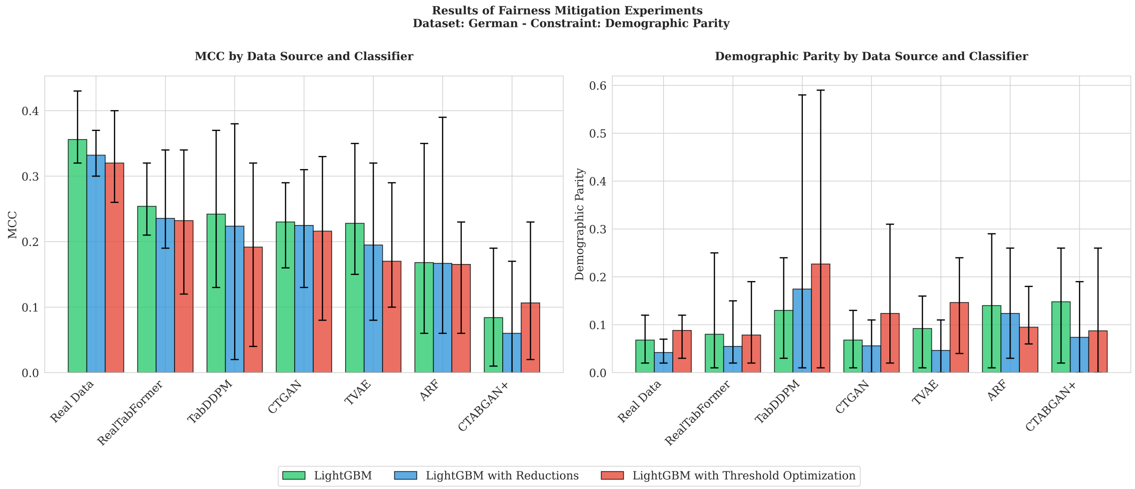


Figure 4.32: Results for MCC and Demographic Parity for the German dataset.

When configured to minimize Equalized Odds, as shown in Table 4.18 and Figure 4.33, the results were even more concerning. There was virtually no reduction

Table 4.18: Results of performance and fairness metrics for the fairness mitigation experiments of the German dataset with Equalized Odds as constraint.

Data Source	Classifier	Accuracy	F1-Score	MCC	Demographic Parity	Equalized Odds
Original Data	LightGBM	0.744 ± 0.025	0.522 ± 0.026	0.356 ± 0.046	0.068 ± 0.045	0.136 ± 0.064
	LightGBM with Reduction	0.744 ± 0.025	0.522 ± 0.026	0.356 ± 0.046	0.068 ± 0.045	0.136 ± 0.064
	LightGBM with Threshold Optimization	0.742 ± 0.023	0.51 ± 0.037	0.348 ± 0.051	0.076 ± 0.042	0.126 ± 0.08
RealTabFormer	LightGBM	0.702 ± 0.027	0.448 ± 0.027	0.254 ± 0.042	0.08 ± 0.097	0.124 ± 0.083
	LightGBM with Reduction	0.698 ± 0.029	0.448 ± 0.027	0.252 ± 0.043	0.074 ± 0.099	0.12 ± 0.086
	LightGBM with Threshold Optimization	0.698 ± 0.025	0.454 ± 0.03	0.256 ± 0.033	0.068 ± 0.077	0.134 ± 0.068
TabDDPM	LightGBM	0.702 ± 0.046	0.426 ± 0.07	0.242 ± 0.111	0.13 ± 0.094	0.206 ± 0.168
	LightGBM with Reduction	0.702 ± 0.046	0.426 ± 0.07	0.242 ± 0.111	0.13 ± 0.094	0.206 ± 0.168
	LightGBM with Threshold Optimization	0.689 ± 0.032	0.418 ± 0.059	0.216 ± 0.08	0.164 ± 0.088	0.262 ± 0.156
ARF	LightGBM	0.696 ± 0.03	0.324 ± 0.107	0.168 ± 0.108	0.14 ± 0.125	0.202 ± 0.161
	LightGBM with Reduction	0.696 ± 0.03	0.32 ± 0.111	0.166 ± 0.109	0.14 ± 0.125	0.206 ± 0.156
	LightGBM with Threshold Optimization	0.692 ± 0.03	0.362 ± 0.105	0.188 ± 0.103	0.164 ± 0.108	0.236 ± 0.147
TVAE	LightGBM	0.688 ± 0.037	0.434 ± 0.059	0.228 ± 0.082	0.092 ± 0.058	0.136 ± 0.051
	LightGBM with Reduction	0.68 ± 0.038	0.426 ± 0.066	0.214 ± 0.089	0.086 ± 0.062	0.134 ± 0.055
	LightGBM with Threshold Optimization	0.67 ± 0.035	0.432 ± 0.049	0.198 ± 0.068	0.054 ± 0.056	0.108 ± 0.056
CTGAN	LightGBM	0.678 ± 0.024	0.458 ± 0.043	0.23 ± 0.051	0.068 ± 0.048	0.128 ± 0.056
	LightGBM with Reduction	0.678 ± 0.024	0.458 ± 0.043	0.23 ± 0.051	0.068 ± 0.048	0.128 ± 0.056
	LightGBM with Threshold Optimization	0.68 ± 0.023	0.454 ± 0.038	0.224 ± 0.048	0.078 ± 0.068	0.126 ± 0.05
CTABGAN+	LightGBM	0.666 ± 0.031	0.248 ± 0.114	0.084 ± 0.067	0.148 ± 0.104	0.218 ± 0.159
	LightGBM with Reduction	0.666 ± 0.031	0.248 ± 0.114	0.084 ± 0.067	0.148 ± 0.104	0.218 ± 0.159
	LightGBM with Threshold Optimization	0.672 ± 0.037	0.276 ± 0.118	0.11 ± 0.073	0.142 ± 0.101	0.212 ± 0.148

in Equalized Odds for the real data or the synthetic datasets using either of the two fairness mitigation methods. In some cases, such as with RealTabFormer, TabDDPM, and ARF, the Thresholding Optimization methods even led to increases in Equalized Odds, with values of 0.134, 0.262, and 0.236, compared to the regular classifier baselines, which showed values of 0.124, 0.206, and 0.202, respectively. The only decreases occurred for RealTabFormer with the Reductions method, which was an insignificant decrease of just 0.004 in Equalized Odds, and for TVAE with Threshold Optimization, where there was a small decrease, dropping from 0.136 in the regular classifier baseline to 0.108.

This pattern suggests that the German dataset presents unique challenges for fairness mitigation algorithms, possibly due to its smaller size, different feature distributions, or inherent data characteristics that make the fairness-performance trade-off more difficult to optimize effectively.

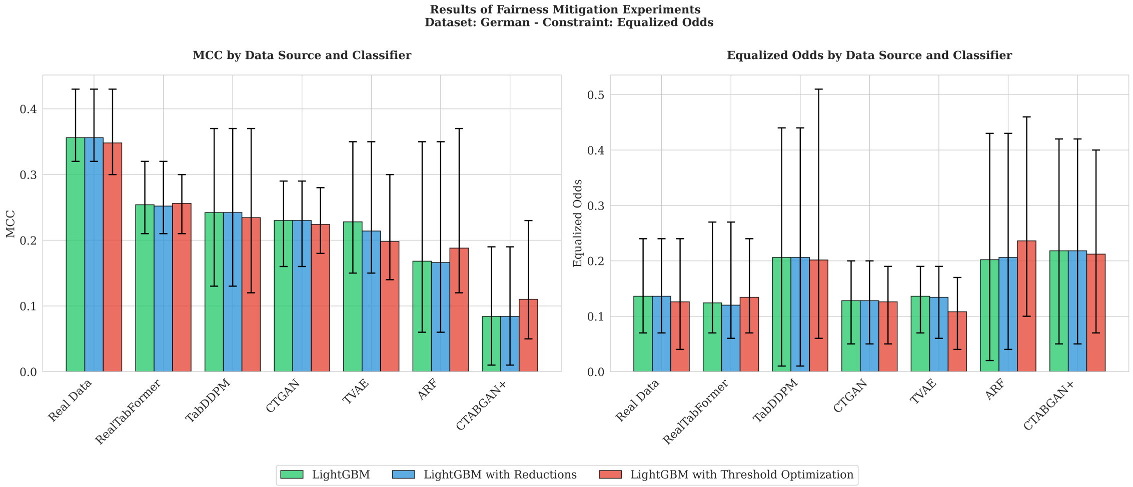


Figure 4.33: Results for MCC and Equalized Odds for the German dataset.

Chapter 5

Conclusion

In recent years, synthetic tabular data generation has experienced significant growth in both research and industry applications. However, the presence of biases in synthetic datasets remains largely unexplored, raising critical questions about their implications for fairness-sensitive applications. This dissertation addressed three fundamental research questions: **RQ1** - Do synthetic datasets preserve the unfairness present in the original data? **RQ2** - Do they potentially exacerbate this unfairness? **RQ3** - Can fairness mitigation algorithms effectively reduce unfairness when applied to classifiers trained with synthetic data?

A comprehensive empirical evaluation was conducted using six synthetic tabular data generation models across four widely used datasets in fairness literature (Adult, Bank Marketing, COMPAS, and German). A rigorous experimental framework was established, incorporating hyperparameter tuning of STDG models and cross-validation throughout the entire workflow.

The hyperparameter optimization revealed distinct patterns among models. TVAE, CTGAN, and CTABGAN+ exhibited high sensitivity to hyperparameters with greater variability in performance, while RealTabFormer and ARF demonstrated the lowest variability. RealTabFormer, ARF, and TabDDPM achieved the highest overall scores in the evaluation.

With respect to **RQ1** and **RQ2**, the fairness evaluation demonstrated that synthetic datasets frequently worsen classifier fairness compared to original data. For the Adult dataset, RealTabFormer showed increases in Demographic Parity and Equalized Odds of 10.3% and 3.5%, respectively, despite being the best-performing model with only a 1.6% decrease in MCC. Even worse, CTABGAN+ exhibited increases of 38.6% in Demographic Parity and 188.1% in Equalized Odds. Similar patterns emerged across the other datasets, with the German dataset showing particularly poor results across all models, likely due to its smaller size (MCC decreases ranging from 28.7% to 76.4%).

Regarding **RQ3**, fairness mitigation algorithms proved effective across synthetic datasets. In the Adult dataset, applying fairness mitigation to RealTabFormer-generated data reduced Demographic Parity from 0.202 to 0.034 (Reductions) and 0.02 (Threshold Optimization), closely matching real data performance, where reductions went from 0.18 to 0.016 and 0.01, respectively. This pattern held consistently across the Adult, Bank Marketing, and COMPAS datasets, where fairness mitigation achieved substantial reductions in both Demographic Parity and Equalized Odds while maintaining reasonable performance levels.

The effectiveness of fairness mitigation on synthetic data closely followed the effect of these algorithms applied to real data, particularly for higher-utility synthetic datasets. Models achieving superior machine learning utility also demonstrated the most effective fairness improvements after mitigation. However, the German dataset proved an exception, with fairness mitigation algorithms showing limited or counterproductive effects across all models and methods.

These findings suggest a viable strategy for fairness-sensitive applications: selecting high-utility synthetic data generation models that closely approximate the original distribution, followed by direct application of fairness mitigation algorithms. The results indicate that while synthetic datasets may initially amplify unfairness, this can be effectively addressed through post-processing fairness interventions, provided the underlying synthetic data quality is sufficient.

Future research could expand this evaluation to include hyperparameter tuning of classifiers and fairness methods, investigate root causes of increased unfairness in synthetic data, and explore fairness-aware and privacy-focused STDG models. Furthermore, extending this analysis to include other fairness definitions such as intersectional fairness could be an interesting avenue of research. Additionally, investigating synthetic data for fairness-oriented data augmentation and oversampling presents promising research directions.

References

FEUERRIEGEL, S., HARTMANN, J., JANIESCH, C., et al. "Generative ai", *Business & Information Systems Engineering*, v. 66, n. 1, pp. 111–126, 2024.

ASSEFA, S. A., DERVOVIC, D., MAHFOUZ, M., et al. "Generating synthetic data in finance: opportunities, challenges and pitfalls". In: *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

HERNANDEZ, M., EPELDE, G., ALBERDI, A., et al. "Synthetic data generation for tabular health records: A systematic review", *Neurocomputing*, v. 493, pp. 28–45, 2022.

FONSECA, J., BACAO, F. "Tabular and latent space synthetic data generation: a literature review", *Journal of Big Data*, v. 10, n. 1, pp. 115, 2023.

MOSTLY AI. "MOSTLY AI Raises $25M to Bring Synthetic Data to Every Enterprise". Jan 2022. Disponível em: <https://mostly.ai/news/mostly-ai-raises-25m-to-bring-synthetic-data-to-every-enterprise>. Press release announcing Series B funding round led by Molten Ventures.

SILICONANGLE. "Nvidia reportedly acquires Gretel for $320M+ to strengthen AI training tools", *SiliconANGLE*, Mar 2025. Disponível em: <https://siliconangle.com/2025/03/19/nvidia-reportedly-acquires-gretel-320m-strengthen-ai-training-tools/>. Report on Nvidia's acquisition of synthetic data platform startup Gretel Labs Inc.

DUNNE, D., DUNNE, K., BARANEK, D. "JP Morgan and the Case of Synthetic Data", *SSRN Electronic Journal*, Apr 2024. doi: 10.2139/ssrn.5336310. Disponível em: <https://ssrn.com/abstract=5336310>. Available at SSRN.

XU, L., SKOULARIDOU, M., CUESTA-INFANTE, A., et al. "Modeling Tabular data using Conditional GAN". In: Wallach, H., Larochelle, H., Beygelzimer, A., et al. (Eds.), *Advances in Neural Information Processing Systems*, v. 32. Curran Associates, Inc., 2019. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf>.

ZHAO, Z., KUNAR, A., BIRKE, R., et al. "Ctab-gan: Effective table data synthesizing". In: *Asian Conference on Machine Learning*, pp. 97–112. PMLR, 2021.

XU, D., YUAN, S., ZHANG, L., et al. "Fairgan: Fairness-aware generative adversarial networks". In: *2018 IEEE international conference on big data (big data)*, pp. 570–575. IEEE, 2018.

RAJABI, A., GARIBAY, O. O. "Tabfairgan: Fair tabular data generation with generative adversarial networks", *Machine Learning and Knowledge Extraction*, v. 4, n. 2, pp. 488–501, 2022.

VAN BREUGEL, B., KYONO, T., BERREVOETS, J., et al. "Decaf: Generating fair synthetic data using causally-aware generative networks", *Advances in Neural Information Processing Systems*, v. 34, pp. 22221–22233, 2021.

EITAN, Y., CAVAGLIONE, N., ARBEL, M., et al. "Fair synthetic data does not necessarily lead to fair models". In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

BHANOT, K., QI, M., ERICKSON, J. S., et al. "The problem of fairness in synthetic healthcare data", *Entropy*, v. 23, n. 9, pp. 1165, 2021.

GANEV, G., OPRISANU, B., DE CRISTOFARO, E. "Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data". In: *International Conference on Machine Learning*, pp. 6944–6959. PMLR, 2022.

LIU, Q., DEHO, O., VADIEE, F., et al. "Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms". In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 591–600, 2025.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., et al. "Generative Adversarial Nets". In: Ghahramani, Z., Welling, M., Cortes, C., et al. (Eds.), *Advances in Neural Information Processing Systems*, v. 27. Curran Associates, Inc., 2014. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.

ZHAO, Z., KUNAR, A., BIRKE, R., et al. "Ctab-gan+: Enhancing tabular data synthesis", *arXiv preprint arXiv:2204.00401*, 2022.

KOTELNIKOV, A., BARANCHUK, D., RUBACHEV, I., et al. "Tabddpm: Modelling tabular data with diffusion models". In: *International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

BORISOV, V., SESSLER, K., LEEMANN, T., et al. "Language Models are Realistic Tabular Data Generators". In: *ICLR*, 2023.

SOLATORIO, A., DUPRIEZ, O. "Generating synthetic data using REaLTabFormer, and assessing the probabilistic measure of statistical disclosure risk". In: *UNECE Expert Meeting on Statistical Data Confidentiality*, pp. 26–28, 2023.

WATSON, D. S., BLESCH, K., KAPAR, J., et al. "Adversarial random forests for density estimation and generative modeling". In: *International Conference on Artificial Intelligence and Statistics*, pp. 5357–5375. PMLR, 2023.

HERNADEZ, M., EPELDE, G., ALBERDI, A., et al. "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions", *Methods of information in medicine*, v. 62, n. S 01, pp. e19–e38, 2023.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., et al. "CatBoost: unbiased boosting with categorical features", *Advances in neural information processing systems*, v. 31, 2018.

MEHRABI, N., MORSTATTER, F., SAXENA, N., et al. "A survey on bias and fairness in machine learning", *ACM computing surveys (CSUR)*, v. 54, n. 6, pp. 1–35, 2021.

KAMIRAN, F., CALDERS, T. "Classifying without discriminating". In: *2009 2nd international conference on computer, control and communication*, pp. 1–6. IEEE, 2009.

DWORK, C., HARDT, M., PITASSI, T., et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

KAMISHIMA, T., AKAHO, S., SAKUMA, J. "Fairness-aware learning through regularization approach". In: *2011 IEEE 11th international conference on data mining workshops*, pp. 643–650. IEEE, 2011.

ZEMEL, R., WU, Y., SWERSKY, K., et al. "Learning fair representations". In: *International conference on machine learning*, pp. 325–333. PMLR, 2013.

LOUIZOS, C., SWERSKY, K., LI, Y., et al. "The variational fair autoencoder". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, May 2016.

PADALA, M., GUJAR, S. "Fnnc: Achieving fairness through neural networks". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI. https://www. ijcai. org/proceedings/2020/0315. pdf Go to original source*, 2020.

MADRAS, D., CREAGER, E., PITASSI, T., et al. "Learning adversarially fair and transferable representations". In: *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

HARDT, M., PRICE, E., SREBRO, N. "Equality of opportunity in supervised learning", *Advances in neural information processing systems*, v. 29, 2016.

KAMIRAN, F., KARIM, A., ZHANG, X. "Decision theory for discrimination-aware classification". In: *2012 IEEE 12th international conference on data mining*, pp. 924–929. IEEE, 2012.

BORISOV, V., LEEMANN, T., SESSLER, K., et al. "Deep neural networks and tabular data: A survey", *IEEE transactions on neural networks and learning systems*, 2022.

AGARWAL, A., BEYGELZIMER, A., DUDÍK, M., et al. "A reductions approach to fair classification". In: *International conference on machine learning*, pp. 60–69. PMLR, 2018.

WEERTS, H., DUDÍK, M., EDGAR, R., et al. "Fairlearn: Assessing and Improving Fairness of AI Systems". 2023. Disponível em: <`http://jmlr.org/papers/v24/23-0389.html`>.

CHENG, V., SURIYAKUMAR, V. M., DULLERUD, N., et al. "Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 149–160, 2021.

YALE, A., DASH, S., BHANOT, K., et al. "Synthesizing quality open data assets from private health research studies". In: *International Conference on Business Information Systems*, pp. 324–335. Springer, 2020.

PEREIRA, M., KSHIRSAGAR, M., MUKHERJEE, S., et al. "Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data", *Plos one*, v. 19, n. 2, pp. e0297271, 2024.

LE QUY, T., ROY, A., IOSIFIDIS, V., et al. "A survey on datasets for fairness-aware machine learning", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 12, n. 3, pp. e1452, 2022.

BECKER, B., KOHAVI, R. "Adult". UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

ANGWIN, J., LARSON, J., MATTU, S., et al. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*, 2016. Disponível em: <`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`>.

AKIBA, T., SANO, S., YANASE, T., et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.